Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

JORGE LUIZ NUNES DOS SANTOS JUNIOR

INTERFACES ENTRE LEXICOGRAFIA E DIALETOLOGIA: POR UM PROTÓTIPO DE VOCABULÁRIO DIALETAL ELETRÔNICO DA REGIÃO NORTE DO BRASIL

Três Lagoas/MS
2023

JORGE LUIZ NUNES DOS SANTOS JUNIOR

INTERFACES ENTRE LEXICOGRAFIA E DIALETOLOGIA: POR UM PROTÓTIPO DE VOCABULÁRIO DIALETAL ELETRÔNICO DA REGIÃO NORTE DO BRASIL

Tese apresentada ao Programa de Pós-graduação em Letras da Universidade Federal de Mato Grosso do Sul, *Campus* de Três Lagoas, área de concentração Estudos Linguísticos, como requisito para a obtenção do título de Doutor em Letras.

Linha de pesquisa: Análise, descrição e documentação de línguas

Orientadora: Profa. Dra. Aparecida Negri Isquerdo (UFMS/CNPq).

Coorientador: Prof. Dr. Fabrice Charles Bernard Issac (Université Paris Nord).

Três Lagoas/MS
2023

“Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)”.

Dedico este trabalho à minha esposa, companheira de todas as horas, e à minha filha que veio ao mundo meses antes deste projeto nascer.

AGRADECIMENTOS

A Deus por ter sustentado minha família e a mim nas adversidades que enfrentamos no decorrer desta pesquisa.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) pelo apoio financeiro.

Ao Programa de Pós-Graduação em Letras, coordenação de curso e secretaria pelo suporte aos discentes.

À minha orientadora, professora Dra. Aparecida Negri Isquerdo, por ter me apresentado a Dialetologia e por todas as suas aulas, orientações, avaliações e, sobretudo, pela paciência e dedicação.

Ao meu coorientador, professor Dr. Fabrice Charles Bernard Issac, por ter me apresentado o *XML* e me mostrar como usá-lo para inovar a metodologia da pesquisa científica no ramo da Linguística.

À professora Dra. Elizabete Aparecida Marques pelas aulas e sugestões que ajudaram este trabalho a tomar forma, sobretudo, quando me disse no Exame de Qualificação que a pesquisa precisava focalizar na resolução de um problema.

Ao professor Dr. Renato Rodrigues Pereira pelas aulas, conversas e minuciosas sugestões que enriqueceram o trabalho e me ajudaram a crescer como pesquisador.

Ao meu colega de turma Roosevelt Vicente Ferreira pelas inúmeras vezes em que estudamos e discutimos sobre temas relacionados à Linguística.

Aos demais colegas e professores com quem tive contato durante esses quatro anos de estudos.

SANTOS JUNIOR, Jorge Luiz Nunes dos. *Interfaces entre Lexicografia e Dialetologia: por um protótipo de vocabulário dialetal eletrônico da região Norte do Brasil*, 2023. 209 f. Tese (Doutorado em Letras) – Universidade Federal de Mato Grosso do Sul, Três Lagoas, MS, 2023.

RESUMO

Esta Tese situa-se na intersecção da Lexicografia com a Dialetologia e a Linguística Computacional e surge da necessidade de contribuir para o preenchimento da lacuna relacionada ao tratamento lexicográfico e eletrônico dos dados dialetais do Projeto Atlas Linguístico do Brasil (ALiB), no caso, os documentados nas 18 localidades que compõem a rede de pontos do Projeto ALiB, no interior da região Norte do Brasil e, por extensão, colaborar com o Dicionário Dialetal Brasileiro (MACHADO FILHO, 2010) que tem como propósito produzir um dicionário dialetal a partir dos dados do ALiB. O estudo tem como objetivo principal produzir e apresentar um protótipo lexicográfico digital desenvolvido a partir do *corpus* dialetal do Projeto ALiB documentado no interior da região Norte do país. Assim, o produto final desta Tese se caracteriza por um conjunto de ferramentas eletrônicas que formam o protótipo do *Vocabulário Dialetal da região Norte do Brasil* (*VoDiNorte*), disponível no link <http://vodinorte.bombadil.fr/>. Essas ferramentas, por sua vez, são capazes de recuperar informações dialetais estruturadas em um banco de dados em *XML* (*Extensible Markup Language*) e apresentá-las ao usuário em uma organização lexicográfica. Destacamos, ainda, que essas ferramentas estão classificadas em dois grupos, a saber: i) ferramentas prototípicas – destinadas ao usuário comum e intermediário e que se assemelham aos produtos lexicográficos eletrônicos convencionais e, ii) um motor de busca avançada no *corpus* de pesquisa, voltado para atender ao usuário avançado, que se caracteriza como um produto inovador do ponto de vista da Lexicografia Eletrônica. Além disso, a Tese apresenta a metodologia empregada com o objetivo de encorajar pesquisadores do ramo da Linguística a repensarem a terceirização de serviços relacionados à programação de softwares de baixa complexidade, abrindo espaço para a exploração de um novo horizonte em que o estudioso pode investir em conhecimentos pontuais para desenvolver suas próprias ferramentas computacionais, ou seja, a criação de soluções informatizadas que atendam a objetivos específicos de uma determinada pesquisa científica. Em síntese, esta Tese demonstra não apenas as funcionalidades do protótipo do *VoDiNorte*, como também estende um convite aos pesquisadores das áreas relacionadas à Linguística, para que possam considerar a possibilidade de criarem seus *corpora* de pesquisa em *XML* experimentando, assim, as maneiras de manipulação de dados e de recuperação de informação de forma automatizada e personalizada a exemplo do protótipo do *VoDiNorte* que pode ser visto, sob um viés técnico, como uma aplicação web que funciona a partir de uma base de dados em *XML*. Além disso, o banco de dados desta Tese pode servir de base para outros produtos e/ou integrar outros projetos como, por exemplo, o Dicionário Dialetal Brasileiro (MACHADO FILHO, 2010) tendo em vista que o *XML* é uma linguagem compatível e de fácil utilização junto a outros sistemas informatizados. Destacamos, ainda, que esta pesquisa está fundamentada na Lexicografia (TARP, 2019; 2018; 2011; 2008); (FUERTES-OLIVERA; BERGENHOLTZ, 2015); (FUERTES-OLIVERA; TARP, 2014); (LEROYER, 2011); (PORTO DAPENA, 2004), na Dialetologia (CARDOSO, 2010); (RADTKE; THUN, 1996); (CHAMBERS; TRUDGILL, 1994); Linguística de *Corpus*

(McENERY; HARDIE, 2012); (O'KEEFFE; McCARTHY, 2010); (BERBER SARDINHA, 2004); (SINCLAIR, 1991) e na Linguística Computacional (KEDIA; RASU, 2020); (SRINIVASA-DESIKAN, 2018); (HARERT, 2004; 2005). Por fim, a pesquisa demonstrou que a criação de bases de dados em *XML* para o processamento de dados lexicais fornece ao pesquisador subsídios para o desenvolvimento de ferramentas computacionais que podem ser úteis aos estudos linguísticos, além de possibilitarem a criação de produtos digitais a partir de dados de uma pesquisa científica.

Palavras-chave: Projeto ALiB; Lexicografia; Dialetologia; banco de dados em *XML* ferramentas computacionais.

ABSTRACT

This Thesis is located at the intersection of Lexicography with Dialectology and Computational Linguistics and arises from the need to contribute to filling the gap related to the lexicographical and electronic treatment of the dialectal data from the Atlas Linguístico do Brasil (AliB) Project, documented in 18 locations which compose the network of the ALiB Project points, at the interior from Brazil Northern region and, by extension, collaborate with the Brazilian Dialectal Dictionary (MACHADO FILHO, 2010) which aims to produce a dialect dictionary from ALiB data. The main study objective is to produce and to present a digital lexicographical prototype developed from ALiB Project dialectal corpus. Thus, the final product of this Thesis is characterized by a set of electronic tools which compose the prototype of the Dialectal Vocabulary from the Brazil Northern region (VoDiNorte), available at the link <http://vodinorte.bombadil.fr/>. These tools are capable of retrieving structured dialectal information from an XML (Extensible Markup Language) database and present it to the user in a lexicographic organization. These tools are classified into two groups: i) prototypical tools - designed for the common and intermediate user and resembling conventional electronic lexicographic products, and ii) an advanced search engine in the research corpus, aimed at the advanced user, which is innovative from the perspective of Electronic Lexicography. Furthermore, the Thesis presents the methodology employed to encourage linguistics researchers to rethink the outsourcing of services related to low-complexity software programming, opening up space to the exploration of a new horizon in which the scholar can invest in specific knowledge to develop their own computational tools, that is, the creation of computerized solutions that meet specific objectives of a particular scientific research. In summary, this Thesis demonstrates not only the functionalities of the VoDiNorte prototype as a product but also extends an invitation to researchers in Linguistics related areas to consider the possibility of creating their own research corpora in XML, experimenting with ways of manipulating data and retrieving information in an automated and personalized manner, similar to the VoDiNorte prototype, which can be seen, from a technical standpoint, as a web application that functions from an XML database. Further the database of this Thesis can serve as a basis to another products and/or integrate another projects such as the Dicionário Dialetal Brasileiro (MACHADO FILHO, 2010), given that XML is a compatible and easy to use language alongside other computerized systems. We highlight that this research is based on Lexicography (TARP, 2019; 2018; 2011; 2008); (FUERTES-OLIVERA; BERGENHOLTZ, 2015); (FUERTES-OLIVERA; TARP, 2014); (LEROYER, 2011); (PORTO DAPENA, 2004), on Dialectology (CARDOSO, 2010); (RADTKE; THUN, 1996); (CHAMBERS; TRUDGILL, 1994); on Corpus Linguistics (McENERY; HARDIE, 2012); (O'KEEFFE; MCCARTHY, 2010); (BERBER SARDINHA, 2004); (SINCLAIR, 1991) and on Computational Linguistics (KEDIA; RASU, 2020); (SRINIVASA-DESIKAN, 2018); (HARERT, 2004; 2005). Finally, the research demonstrated that creating XML databases for lexical data provides researchers with subsidies to the development of computational tools that can be useful to linguistic studies, in addition to enabling the creation of digital products from data of a scientific research.

Keywords: ALiB Project; Lexicography; Dialectology; XML database; computational tools.

ATA DE DEFESA DE TESE PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS DOUTORADO

Aos trinta dias do mês de março do ano de dois mil e vinte e três, às treze horas, na Plataforma Google Meet, da Fundação Universidade Federal de Mato Grosso do Sul, reuniu-se a Banca Examinadora composta pelos membros: Aparecida Negri Isquerdo (UFMS), Conceição de Maria de Araújo Ramos (UFMA), Daniela Barreiro Claro (UFBA), Elizabete Aparecida Marques (UFMS) e Renato Rodrigues Pereira (UFMS), sob a presidência do primeiro, para julgar o trabalho do aluno: **JORGE LUIZ NUNES DOS SANTOS JUNIOR**, CPF [redacted], do Programa de Pós-Graduação em Letras, Curso de Doutorado, da Fundação Universidade Federal de Mato Grosso do Sul, apresentado sob o título **"Vocabulário dialetal: protótipo de um produto lexicográfico eletrônico a partir de dados geolinguísticos da região Norte do Brasil"** e orientação de Aparecida Negri Isquerdo. A presidente da Banca Examinadora declarou abertos os trabalhos e agradeceu a presença de todos os Membros. A seguir, concedeu a palavra ao aluno que expôs sua Tese. Terminada a exposição, os senhores membros da Banca Examinadora iniciaram as arguições. Terminadas as arguições, a presidente da Banca Examinadora fez suas considerações. A seguir, a Banca Examinadora reuniu-se para avaliação, e após, emitiu parecer expresso conforme segue:

EXAMINADORES

Dra. Aparecida Negri Isquerdo (Interno)

Dr. Renato Rodrigues Pereira (Interno)

Dra. Elizabete Aparecida Marques (Interno)

Dra. Daniela Barreiro Claro (Externo)

Dra. Conceição de Maria de Araújo Ramos (Externo)

Dra. Beatriz Aparecida Alencar (Interno) (Suplente)

Dr. Bruno Oliveira Maroneze (Externo) (Suplente)

RESULTADO FINAL:

Aprovado

OBSERVAÇÕES: A banca destaca o caráter inovador do trabalho, no que se refere à interface entre Lexicografia eletrônica, Dialetologia e Ciência da Computação - inovação metodológica no desenvolvimento da ferramenta. Além disso, sugere ajuste no título da Tese como segue: **Interfaces entre Lexicografia e Dialetologia: por um protótipo de vocabulário dialetal eletrônico da região Norte do Brasil**.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

ALiB – Atlas Linguístico do Brasil

ALPB – Atlas Linguístico da Paraíba

ALPR – Atlas Linguístico do Paraná

ALAM – Atlas Linguístico do Amazonas

ALSAM – Atlas Linguístico do Sul Amazonense

ALS – Atlas Linguístico de Sergipe

APFB – Atlas prévio dos falares baianos

CSS – *Cascading Style Sheets*

DDB – Dicionário Dialetal Brasileiro

DTD – *Document Type Definition*

EALMG – Esboço de um Atlas Linguístico de Minhas Gerais

FLEX – *FieldWorks Language Explorer*

GPS – *Global Positioning System*

HTML – *HyperText Marukp Language*

IBGE – *Instituto Brasileiro de Geografia e Estatística*

ID - Identificação

PLN – Processamento de Linguagem Natural

QSL – Questionário Semântico-lexical

TN – Tradução nossa

UL – Unidade lexical

OCR – *Optical Character Recognition*

VoDiNorte – Vocabulário Dialetal da região Norte do Brasil

XML – *Extensive Markup Language*

SUMÁRIO

INTRODUÇÃO

As obras lexicográficas apresentam uma tipologia[1] bastante variada, tendo em vista que são concebidas para diversas finalidades. Com isso em mente é importante esclarecer, de imediato, a diferença entre dicionário e vocabulário para a compreensão do escopo desta Tese. Assim, um dicionário geral como, por exemplo, os *dicionários de língua* ou *tesauros* em que o número de verbetes pode ultrapassar 500.000 entradas têm a pretensão de descrever todo o léxico de um povo, ou seja, busca na medida do possível repertoriar o maior montante possível do universo lexical de uma língua. O vocabulário, por sua vez, se concentra numa porção lexical de menor proporção, geralmente vinculada a uma área de especialidade ou a uma região geográfica. A quantidade de entradas, de verbetes nesse tipo de obra é significativamente menor. Em alguns casos, vocabulários podem contribuir como fonte de dados para o desenvolvimento de dicionários de maior porte.

Relação dessa natureza pode ser observada no Projeto do *Dicionário Dialetal Brasileiro (DDB)* (MACHADO FILHO, 2010), que tem como objetivo mais amplo descrever o vocabulário dialetal de habitantes das cinco regiões brasileiras com base nos dados geolinguísticos documentados pelo *Projeto Atlas Linguístico do Brasil (ALiB)*[2]. O Projeto do DDB, em andamento na Universidade Federal da Bahia (UFBA) sob a coordenação do Prof. Dr. Américo Venâncio Lopes Machado Filho, constitui-se em um projeto de ampla envergadura que reúne especialistas das áreas da Dialetologia, da Lexicografia e das Ciências da Informação. Associados a esse projeto, trabalhos de pós-graduação têm contribuído para o tratamento lexicográfico dos dados do Projeto ALiB a partir de porções de *corpora* que reúnem localidades da rede de pontos do projeto, representativas das diferentes regiões do Brasil. Esses estudos, vinculado ao Projeto do DDB, configuram-se como vocabulários dialetais em termos lexicográficos:

i) *Vocabulário dialetal baiano*, Tese de Doutorado de autoria de Isamar Neiva de Santana, defendida em 2017 na UFBA, trabalho orientado pelo Prof. Dr. Américo

[1] Esse assunto será detalhado no item *1.3. Tipologia lexicográfica*.

[2] Projeto interinstitucional com sede na Universidade Federal da Bahia (UFBA) que tem a meta mais ampla de produzir um atlas geral da língua portuguesa do Brasil. Maiores informações sobre o Projeto ALiB podem ser obtidas por meio de consulta ao site do projeto: <https://alib.ufba.br/>. Acesso em: 15

[2] Projeto interinstitucional com sede na Universidade Federal da Bahia (UFBA) que tem a meta mais ampla de produzir um atlas geral da língua portuguesa do Brasil. Maiores informações sobre o Projeto ALiB podem ser obtidas por meio de consulta ao site do projeto: <https://alib.ufba.br/>. Acesso em: 15 nov. 2022.

Venâncio Lopes Machado Filho que analisou os dados relativos às localidades da rede de pontos do Projeto ALiB pertencentes ao estado da Bahia;

ii) *Vocabulário dialetal maranhense: a contribuição do Maranhão para o Dicionário Dialetal Brasileiro*, Dissertação de Mestrado produzida por Camila Maramaldo Ferreira Robson, com base nos dados do Projeto ALiB relativos a duas localidades do interior (Tuntum e Bacabal) e da capital, do estado do Maranhão (São Luís). O trabalho de dissertação foi defendido em 2017, na Universidade Federal do Maranhão (UFMA), sob orientação do Prof. Dr. José de Ribamar Mendes Bezerra;

iii) *Vocabulário dialetal do Centro-Oeste: interfaces entre a Lexicografia e a Dialetologia*, Tese produzida por Daniela de Souza Silva Costa, com base em dados da rede de pontos do Projeto ALiB pertencente à região Centro-Oeste. A Tese foi defendida em 2018, na Universidade Estadual de Londrina (UEL), sob a orientação da Profa. Dra. Aparecida Negri Isquerdo.[3]

Além desses trabalhos já concluídos, há duas outras pesquisas em andamento, em nível de doutorado, que estão desenvolvendo produtos lexicográficos a partir dos dados dialetais do Projeto ALiB, sob a orientação do Prof. Dr. Américo Venâncio Lopes Machado Filho, a saber: i) *Vocabulário dialetal do Sul e do Sudeste do Brasil com base nos dados do Projeto ALiB*, de autoria de Cemary Correia de Sousa, iniciada em 2020 e, ii) *Vocabulário dialetal da região Nordeste*, de Maria José Ferreira da Silva, iniciada em 2021.

No que se refere, exclusivamente, à região Norte do Brasil, até onde foi possível apurar, há um único estudo de cunho lexicográfico realizado a partir dos dados dialetais do ALiB: o *Vocabulário dialetal da região Norte do Brasil: um estudo das capitais com base nos dados do Projeto ALiB*, produzido por Cemary Correia de Souza, como dissertação de Mestrado defendida na Universidade Federal da Bahia (UFBA), em 2019, sob a orientação do Prof. Dr. Américo Venâncio Lopes Machado Filho.

Destacamos que os trabalhos mencionados são essencialmente lexicográficos e, no tocante à região Norte do Brasil, podemos elencar outros estudos lexicais voltados

[3] Registramos também a Tese de Doutorado *Léxico brasileiro em dicionários monolíngues e bilíngues: estudo metalexicográfico da variação em perspectiva dialetal e histórica*, de autoria de Aniele Souza de Oliveira, defendida em 2017, sob a orientação do Prof. Dr. Américo Venâncio Lopes Machado Filho, na UFBA. Essa tese não analisou *corpus* inédito do Projeto ALiB, mas sim cotejou dados lexicográficos com dados lexicais registrados nas cartas lexicais que compõem o vol. 2 do ALiB (CARDOSO et al, 2014). Um dos produtos dessa tese foi a elaboração de um "pequeno glossário dialetal bilíngue em que são reunidas as unidades lexicais do ALiB, privilegiando-se a variação diatópica na elaboração das definições" (OLIVEIRA, 2017, p. vi).

para a produção de atlas linguísticos, especificados na sequência: i) *Atlas Linguístico do Amazonas (ALAM),* de autoria de Maria Luiza de Carvalho Cruz (2004); ii) *Atlas Linguístico Sonoro do Pará (ALISPA),* de autoria de Abdelhak Razky (2004); iii) *Atlas dos falares do baixo Amazonas (AFBAM),* de Roseanny Melo de Brito (2011); iv) *Pelos caminhos da cartografia linguística paraense: um estudo semântico-lexical do Distrito Mosqueiro numa perspectiva socioeducacional,* trabalho produzido por Talita Rodrigues de Sá (2013); v) *Atlas Linguístico do Amapá (ALAP),* de autoria de Abdelhak Razky, Celeste Maria da Rocha Ribeiro, Romário Duarte Sanches (2017); vi) *Atlas Linguístico dos Falares de Manaus (ALFAMA),* produzido por Leticia Pinto Cardoso (2018); vii) *Atlas Linguístico do Sul Amazonense (ALSAM),* de autoria de Edson Galvão Maia (2018); viii) *Atlas Linguístico Topodinâmico e Topoestático do Estado do Tocantins (ALITTETO),* de autoria de Greize Alves da Silva (2018); ix) *Atlas Etnolinguístico do Acre (ALAC): fronteiras léxicas*, produzida por Luísa Galvão Lessa Kalberg (2018); x) *Atlas Linguístico do Acre (ALiAC)*, de Lindinalva Messias, em desenvolvimento na Universidade Federal do Acre e o *Atlas Linguístico de Rondônia (ALiRO)*, de Iara Telles, também em desenvolvimento.

Ainda na esteira de registro de trabalhos dialetais e lexicográficos, Sá (2021) realiza uma análise de dados de dois atlas linguísticos amazonenses, a saber: o *Atlas Linguístico do Amazonas (ALAM),* (CRUZ, 2004) e o Atlas *Linguístico do Sul Amazonense (ALSAM)* (MAIA, 2018), comparando-os com os registros lexicográficos de Houaiss (2009), de Ferreira (2010) e de Michaelis (2015). Esse estudo teve como objetivo comparar definições relativas ao item lexical *igarapé*, registradas nesses três dicionários com os dados mapeados nos dois atlas linguísticos selecionados. O autor concluiu que há divergências entre as acepções lexicográficas de *igarapé* em relação aos dados registrados nos atlas linguísticos amazonenses, o que revela a necessidade de atualização dos dicionários quanto ao registro de marcas diatópicas (SÁ, 2021, p. 224).

Com base nesse panorama de trabalhos dialetais e, sobretudo, lexicográficos, nota-se uma lacuna no que diz respeito a estudos lexicográficos com dados do Projeto ALiB documentados no interior da região Norte do Brasil, pois a pesquisa de Correia de Souza (2019) analisou apenas dados das capitais.

Nesse sentido, uma comparação dos dados geolinguísticos do ALiB documentados nas capitais com os oriundos do interior poderá revelar particularidades significativas, tendo em vista que nas capitais os informantes entrevistados possuem um perfil de escolaridade que contempla indivíduos com o Ensino Fundamental incompleto

e com o Ensino Superior, ao passo que no interior os informantes deviam possuir o Ensino Fundamental incompleto de escolaridade.

Diante do exposto, a seguinte pergunta norteou os passos metodológicos desta Tese, a saber: como dar tratamento informatizado aos dados orais do Projeto ALiB de modo a recuperar informações de maneira automática e desenvolver um produto lexicográfico on-line?

A partir desse questionamento, a seguinte hipótese foi levantada: investir na aquisição de conhecimentos pontuais relacionados à Ciência da Computação permite ao linguista desenvolver ferramentas computacionais de baixa complexidade dispensando, dessa forma, a contratação de programadores para realizar a automação de determinadas tarefas, além de abrir horizontes para uma metodologia de trabalho interdisciplinar.

Esta Tese tem como objetivo geral elaborar um protótipo do *Vocabulário dialetal da região Norte do Brasil* (VoDiNorte), disponível no link <http://vodinorte.bombadil.fr/>, com base em dados orais que integram o *corpus* do Projeto ALiB – Atlas Linguístico do Brasil, documentados em localidades do interior da região Norte do Brasil, ou seja, 18 localidades que integram a rede de pontos do ALiB. Esses dados são oriundos das respostas fornecidas por 72 informantes naturais dessas localidades para as 202 perguntas do Questionário Semântico-Lexical do Projeto ALiB (QSL-ALiB), gravadas em áudio pela equipe de pesquisa do projeto.

Os objetivos específicos, por sua vez, foram assim estabelecidos:

i) dar tratamento lexicográfico e eletrônico aos dados lexicais do Projeto ALiB documentados nas 18 localidades da rede de pontos do interior da região Norte do Brasil;

ii) elaborar ferramentas computacionais para a recuperação de dados de forma automática;

iii) criar uma aplicação web responsável por acessar o banco de dados e exibir informações do protótipo do *VoDiNorte* de maneira on-line;

iv) contribuir com o Projeto *Dicionário Dialetal Brasileiro* em relação aos dados do ALiB armazenados em *XML*.

Entende-se que os impactos desta Tese podem ser vistos em três frentes que se interligam naturalmente, a saber: i) o uso gratuito e on-line do protótipo do *VoDiNorte* e a contribuição das ferramentas construídas para o desenvolvimento de produtos lexicográficos genuinamente eletrônicos; ii) as possibilidades de estudos futuros; iii) o incentivo a pesquisadores que desejem investir na aquisição de conhecimentos

específicos relacionados à Ciência da Computação para desenvolver suas próprias soluções computacionais podendo, dessa forma, utilizar recursos de programação para desenvolver pesquisas no campo da Linguística.

Por fim, a Tese está organizada da seguinte maneira:

Capítulo 1 – Léxico e Lexicografia – no qual discutimos o tipo de relação existente entre o léxico e o ato de nomear, além de relacionar essa prática aos dicionários que tradicionalmente podem ser vistos como repositórios das línguas no decorrer do tempo. Além disso, aspectos teóricos da Lexicografia Impressa são abordados nesse capítulo a fim de oferecer um contraponto reflexivo em relação aos pressupostos apresentados no capítulo seguinte.

No *Capítulo 2 – Caminhos para uma Lexicografia Eletrônica* – focalizamos a evolução da Lexicografia ao longo de sua história de modo que, atualmente, é possível identificar uma mudança teórica-metodológica, no que diz respeito ao uso das ferramentas computacionais utilizadas na construção de dicionários eletrônicos inovadores. Assim, são discutidos critérios de classificação, planejamento e produção de obras lexicográficas eletrônicas, bem como fazemos menção à Teoria Funcional da Lexicografia (TARP, 2008) que orienta o labor lexicográfico a partir do uso das modernas tecnologias computacionais.

O *Capítulo 3 – Interfaces da Lexicografia* – discorre sobre o caráter interdisciplinar da Lexicografia que pode ser visto em suas diversas interrelações com outras disciplinas, com foco nos laços atados entre a Dialetologia e a Lexicografia Dialetal perspectiva de destaque nesta Tese, uma vez que os dados trabalhados são de cunho dialetal. Além disso, há de se pontuar o entremeado metodológico que foi tecido a partir da Linguística de *Corpus* e, principalmente, da Linguística Computacional.

Por sua vez, o *Capítulo 4 – Percurso metodológico* – esboça o trabalho desenvolvido com o *corpus* oral do Projeto ALiB, no tocante aos dados circunscritos aos municípios do interior da região Norte do Brasil, a fim de tratá-los lexicográfica e eletronicamente. Assim, ferramentas computacionais foram desenvolvidas para o processamento desses dados, além de uma aplicação web, destinada a oferecer aos usuários um conjunto de ferramentas que, juntas, apresentam diferentes modos de acesso aos dados do protótipo do *VoDiNorte*.

O *Capítulo 5 – Apresentação do protótipo* – reúne explicações sobre o funcionamento do protótipo do *VoDiNorte*, bem como esclarece sobre as ferramentas disponíveis nos três perfis de usuário que são acessados na página inicial da aplicação

web, a saber: i) usuário comum – leigo e estudantes do Ensino Fundamental; ii) usuário intermediário – leigo, estudantes do Ensino Médio e estudantes do Ensino Superior; iii) usuário avançado – pesquisadores em geral e professores universitários.

Finalmente, as *Conclusões* retomam a pergunta de pesquisa e a hipótese apresentada nesta Introdução a partir da experiência vivenciada no decorrer da construção da Tese. Além do mais, avaliamos os resultados alcançados, projetamos as possibilidades de estudos futuros e convidamos a todos os estudiosos interessados em aprender a desenvolver suas próprias soluções computacionais no âmbito de suas pesquisas, a investir na apropriação de conhecimentos específicos relacionados à Ciência da Computação. Após as conclusões elencamos as referências e os anexos.

CAPÍTULO 1 – LÉXICO E LEXICOGRAFIA

Este capítulo tem como foco a discussão acerca da relação entre o léxico e o processo de catalogação da realidade que circunda o falante, focalizando sua capacidade de transformação, tomando como pressuposto que o léxico pode ser comparado a um organismo vivo que nasce e se desenvolve com o passar dos anos. A importância de se registrar o acervo lexical das línguas naturais em dicionários e os principais pressupostos teórico-metodológicos da Lexicografia também são temas discutidos neste Capítulo.

1.1. O ato de nomear

O ato de nomear se realiza por meio da palavra, cujo conceito, para Biderman (2001, p. 109-123), é muito relativo. A complexidade que envolve o tema é discutida detalhadamente pela autora que chega ao consenso de que, no âmbito dos estudos morfológicos, sintáticos e semânticos, o termo *unidade lexical (UL)* (BIDERMAN, 2001, p. 155) representa a melhor maneira de se referir ao conceito de *palavra*. Dessa forma, nesta Tese fez-se uso desse termo em detrimento do uso de *palavra*, tido como vago do ponto de vista científico.

Nomear objetos, pessoas, lugares e tudo aquilo que rodeia o ser humano é uma prática antiga que pode ser observada desde os registros contidos nas Escrituras Sagradas, segundo a qual Adão, primeiro homem criado por Deus, recebe a tarefa de dar nomes aos animais e às plantas ao seu redor.

Mito ou realidade, o fato é que o homem sempre nomeou e sempre continuará a exercitar essa prática que pode ser constatada também no âmbito da Terminologia[4], no qual novas UL surgem em função do desenvolvimento tecnológico de uma área específica como as ligadas ao universo da Informática que acabam, frequentemente, sendo incorporadas ao léxico comum como é o caso de *backup, banner, bluetooth, bug, cookie, download, PC, print, hacker, hashtag, log in, log out, podcast, scanner, spam*, dentre outras.

[4] Campo de estudos que investiga as linguagens de especialidade que surgem da comunicação estabelecida entre especialistas em uma determinada área do conhecimento.

Todavia, destacamos que esse processo de nomeação não é exclusivo das áreas envolvidas na produção do conhecimento, tendo em vista que o uso da linguagem, em situações específicas, por um grupo de pessoas pode ocorrer em várias circunstâncias como, por exemplo, entre trabalhadores das mais variadas profissões.

Um exemplo típico é o uso das denominações *tarisca*[5] e *tipiti*[6] nas casas de farinha existentes na região Norte do Brasil. Trata-se de terminologias utilizadas no processo de fabricação da farinha de mandioca e que dificilmente são utilizadas em situações comunicativas fora do ambiente de trabalho, ou seja, são termos que ainda não extrapolaram o contexto da linguagem de especialidade em que foram concebidos[7].

Ainda em relação ao ato de nomear, é preciso esclarecer que esse feito se realiza numa língua natural por intermédio de um sistema linguístico em que o léxico é gerado a partir de dois ingredientes essenciais, isto é, um *referente* situado no universo cognoscível que rodeia o homem, como apresentado por Biderman (2006, p. 35) e a escolha lexical que esse homem realiza a partir do *sistema de possibilidades*, proposto por Coseriu (1980, p. 122).

Desse modo, o léxico criado pelo homem se desenvolve e ganha novas formas, conotações e usos distintos. Essa riqueza pode ser estudada a partir de uma perspectiva diatópica como, por exemplo, nos estudos dialetológicos e geolinguísticos. Vale destacar que a norma lexical[8] de uma dada região é resultante da combinação de dois fatores complementares, a saber: i) o conjunto das convenções herdadas pela gramática da norma padrão, que oferece ao falante o sistema de possibilidades da língua. ii) a carga extralinguística proveniente das influências do ambiente físico e social (SAPIR, 1969, p. 44).

Tendo em vista que o processo de nomeação é uma atividade que sempre acompanhará o ser humano no curso de sua história e que o léxico de uma língua se transforma com o passar do tempo, é preciso registrar os itens lexicais e seus sentidos dentro de um espaço de tempo específico, para que essa riqueza lexical não se perca. Sendo assim, os dicionários de língua são instrumentos que se configuram como

[5] Lâmina utilizada nos moinhos que trituram a mandioca na fabricação de farinha.
[6] Ferramenta de origem indígena, feita de cipó trançado, utilizada na fabricação da farinha de mandioca que tem a finalidade de extrair a água da pasta de mandioca ralada.
[7] Essa temática é explorada com mais profundidade em Santos Junior; Isquerdo (2022, no prelo).
[8] Falar típico de uma comunidade de falantes que se forma ao longo dos anos e que pode, em algumas situações de uso, fugir à norma gramatical da língua. Segundo Oliveira (2001, p. 110), a norma lexical é "entendida como costume, a tradição continuada que se verifica nos hábitos linguísticos de uma comunidade."

repositórios das línguas naturais.

1.2. O registro lexical por meio dos dicionários

O exercício do ato de nomear ao longo dos anos está associado a uma carga cultural que é inerente aos indivíduos que vivem em comunidade. Assim, a relação desenvolvida entre o léxico e a cultura resulta num patrimônio vocabular da língua (SEABRA, 2015, p. 66) que não pode ser ignorado.

Dessa maneira, partindo de uma perspectiva lexicográfica, a ação de relacionar um *referente* a um *conceito* constitui-se num mecanismo de registro e preservação lexical de uma nação. Isso porque o dinamismo das línguas naturais, que pode ser comparado a um organismo vivo, se caracteriza pela capacidade de transformação muito sensível ao tempo e aos fatores extralinguísticos. Desse modo, a dicionarização é uma prática importante para a identificação e compreensão das UL que entrarão em desuso no futuro e quais estarão mais vivas na boca do povo, tendo em vista que o léxico é vivo e está em constante mudança (BIDERMAN, 2001, p. 197).

Os dicionários gerais guardam uma imensidão de retratos de tempos pretéritos que remetem aos mais variados usos e contextos de uma língua. Essa particularidade ilustra a relevância da Lexicografia que age como um guardião do léxico, prestando um serviço de singular importância às línguas naturais. No entanto, esta não é a única tarefa de um dicionário, pois esses tipos de obras, normalmente, são concebidas para funcionarem como uma ferramenta de consulta lexicográfica que pode explorar temáticas variadas a partir das necessidades de seus usuários, o que resulta em diferentes tipos de dicionários. No próximo tópico são discutidos critérios estabelecidos pelos estudos lexicográficos para avaliar e classificar um dicionário.

1.3. Tipologia lexicográfica

Os dicionários apresentam uma tipologia muito vasta e essa multiplicidade se justifica pelo princípio de que as necessidades de consulta dos usuários são, igualmente, diversificadas. Assim, um consulente pode abrir um dicionário em busca de informações semânticas, gramaticais, etimológicas, de tradução, de produção textual, enfim, as demandas de consulta dos usuários são variadas o que exige a consulta de distintos tipos de dicionário, ou seja, para cada tipo de demanda há um tipo de

dicionário.

Evidentemente, a maioria dos dicionários são planejados para atender a diferentes demandas de pesquisas em seus verbetes. Tradicionalmente, os dicionários escolares, por exemplo, oferecem dados lexicográficos relacionados à ortografia, à separação silábica, à classificação gramatical, à definição e ao exemplo de uso do verbete em questão. Desse modo, esse tipo de obra atende a, pelo menos, cinco demandas de pesquisa de um estudante. Porém, esse dicionário não apresentará verbetes relacionados a uma linguagem de especialidade, pois não foi projetado para atender a esse público.

Desse modo, várias obras lexicográficas são publicadas com a finalidade de atender a essa diversidade de público, afinal, não pode ser desconsiderado o fato de serem os usuários que movimentam o mercado editorial de obras lexicográficas. Logo, compreender a tipologia das obras lexicográficas reveste-se de importância, pois se constitui num mecanismo de classificação e avaliação. Haensch (1997, p. 49), por exemplo, pondera que, na tarefa de classificação de um dicionário, o mais sensato é procurar responder a seguinte pergunta: Quais são as características e as funções a que se presta uma obra lexicográfica?

Baseado nessa questão, o autor estabelece uma classificação tipológica composta por treze critérios, a saber: 1) a quantidade de entradas que demonstra a dimensão do dicionário; 2) se o dicionário é enciclopédico ou linguístico; 3) o número de línguas contemplado pela obra, ou seja, monolíngue, bilíngue ou multilíngue; 4) o grupo a que a obra se destina, ou seja, os tipos de usuários previstos; 5) suporte impresso ou eletrônico; 6) normativo ou descritivo; 7) o sistema linguístico em que o dicionário se baseou, ou seja, o tipo de *corpus* que foi utilizado para compor a obra; 8) o tipo de ordenação das entradas, ou seja, semasiológico, onomasiológico, inverso, família de palavras, situações de comunicação; 9) a natureza do léxico registrado que pode ser geral ou parcial; 10) se o dicionário é integral (tesouro), representativo (caudal léxico extenso) ou seletivo (menor seleção de entradas); 11) se os dicionários são de recepção (passivos) ou de produção (ativos); 12) repertórios lexicográficos não autônomos e materiais léxicos "escondidos" representados, na maioria das vezes, por glossários terminológicos produzidos por especialistas que não são divulgados fora do âmbito de sua especialidade e; 13), o uso de ilustrações (HAENSCH, 1997, p. 49-59).

Porto Dapena (2002, p. 42- 76), por seu turno, propõe uma organização das obras lexicográficas a partir de dois grupos que, por sua vez, apresentam subdivisões, de

acordo com o objetivo do dicionário. O primeiro grupo abarca os *dicionários não linguísticos*, compreendendo as enciclopédias, os dicionários enciclopédicos e os dicionários terminológicos. O segundo grupo, bem mais vasto, abrange os *dicionários linguísticos* que são divididos em sete categorias, a saber: 1) perspectiva temporal em que os dicionários podem ser sincrônicos ou diacrônicos (históricos, etimológicos); 2) extensão e volume das entradas que se divide segundo o número de línguas (monolíngue, bilíngue ou plurilíngue) e a amplitude do conjunto léxico, que se secciona em dois subgrupos: os dicionários gerais (tesouro, dicionário manual, dicionário de bolso, dicionário abreviado) e os dicionários particulares ourestritos/especiais que, por sua vez, se subdividem em dois outros grupos, os de restrição externa (dicionários dialetais, gíricos, profissionais, terminológicos) e os de restrição interna, compreendendo os dicionários gramaticais (pronúncia, ortográficos, morfológicos, sintáticos) e os dicionários textuais (refrães, locuções, fraseologias, citações, frases célebres); 3) nível ou plano linguístico que abrangem dicionários de língua, dicionários da norma (dicionário normativo e dicionário de uso) e os dicionários do discurso (vocabulários e glossários); 4) microestrutura, compreendendo os dicionários descritivos (dicionários definitórios como o Dicionário da Real Academia Espanhola) e os dicionários não descritivos (índices de palavras, concordâncias); 5) ordem, podendo ser alfabéticos (diretos, inversos como os dicionários de rimas), estatísticos ou de frequência, de famílias etimológicas e ideológicos ou analógicos que, por sua vez, podem se apresentar como temáticos, ideológicos ou de sinônimos e antônimos (distintivos e acumulativos); 6) finalidade que compreende os dicionários pedagógicos (dicionários infantis e escolares e os dicionários para estrangeiros) e outro subgrupo destinado aos dicionários da descodificação e codificação (semasiológicos e onomasiológicos) e, por fim, 7) suporte podendo ser apresentados no formato de dicionários em papel ou eletrônicos (PORTO DAPENA, 2002, p. 42-76).

Por sua vez, Rodríguez Barcia (2016, p. 98-99) apresenta seis critérios para a classificação de dicionários que são detalhados, a seguir:

1) *Critérios qualitativos*: consiste em uma classificação que leva em consideração as qualidades de um dicionário, isto é, as características básicas da obra. Os critérios qualitativos apresentam seis subcritérios: i) *enciclopedismo* no qual se enquadra o dicionário enciclopédico e o dicionário de língua; ii) *perspectiva temporal* em que o dicionário pode ser diacrônico ou sincrônico; iii) *número de línguas* que podem ser dos seguintes tipos: dicionário monolíngue, dicionário bilíngue e dicionário

multilíngue; iv) *presença de ilustrações* como, por exemplo, o dicionário ilustrado ou o dicionário visual; vi) *idade das pessoas destinatárias*, ou seja, o dicionário pode ser infantil ou para adultos; vii) *línguas das pessoas destinatárias*, ou seja, o dicionário é elaborado para pessoas nativas ou para estrangeiros.

2) *Critérios quantitativos:* representados pela avaliação da extensão do dicionário e a seleção do léxico registrado e são divididos em dois subcritérios, a saber: i) *extensão* em que é possível encontrar o dicionário abreviado, o dicionário manual, o dicionário de bolso, ou seja, obras que variam em relação à quantidade de entradas presentes em sua nomenclatura; ii) *seleção do léxico,* ou seja, o dicionário exaustivo ou integral, também conhecido como dicionário geral, além do dicionário seletivo ou representativo como é o caso, por exemplo, do dicionário dialetal, dicionário de dúvidas, dicionário de socioletos, dicionário especializado.

3) *Critérios estruturais:* estabelecidos em função do tipo de ordenação das entradas, bem como a consideração das vozes em seu contexto. Três subcritérios detalham essa classificação: i) *ordenação*, ou seja, o dicionário pode ser semasiológico ou onomasiológico; ii) *eixo sintagmático* no qual se enquadram o dicionário de valência, o dicionário de colocações, os dicionários de construção e regime e os dicionários fraseológicos; iii) eixo paradigmático em que os dicionários podem ser de sinônimos, de ideias e afins, de homônimos, de antônimos, inversos, de rima, ideológicos e tesauros.

4) *Critérios funcionais:* classificação dos dicionários a partir das funções gerais a que se concentram e das necessidades específicas de consulta que buscam satisfazer e são divididos em três subcritérios, a saber: i) *Dicionários de consulta* que podem ser de codificação (produção) ou de descodificação; ii) *Dicionários de aprendizagem* enquadrados em dicionários escolares, dicionários de segundas línguas e dicionários de dúvidas; iii) *Dicionários para profissionais* em que os dicionários podem ser especializados, terminológicos, etimológicos, temáticos e voltados para a tradução.

5) *Critérios puristas*: Relacionados ao caráter normativo dos repertórios lexicográficos e apresentam os seguintes subcritérios: i) *Dicionários normativos* (prescritivos) que são representados pelos dicionários acadêmicos e pelos dicionários abalizados por instituições educativas e governamentais; ii) *Dicionários descritivos* formado por dicionários de uso.

6) *Critérios formais:* rotulam os dicionários por meio do suporte. Apresentam dois subcritérios, a saber: i) *Dicionários impressos*; ii) *Dicionários digitais* que se

caracterizam pelos dicionários em suporte eletrônicos, dicionários digitais, dicionários on-line e aplicações para celular.

A partir desses critérios, o protótipo do *VoDiNorte* pode ser classificado da seguinte maneira:

Quadro 1: Características tipológicas do protótipo do *VoDiNorte*

Critério	Subcritério	Tipo
1) Qualitativo	i) Enciclopedismo	Dicionário de língua
	ii) Perspectiva temporal	Dicionário sincrônico
	iii) Número de línguas	Dicionário monolíngue
	iv) Presença de ilustrações	Dicionário ilustrado
	v) Idade dos destinatários	Dicionário para adultos
	vi) Língua dos destinatários	Dicionário para nativos
2) Quantitativo	i) Extensão	Dicionário abreviado
	ii) Seleção do léxico	Dicionário seletivo ou representativo (dialetal)
3) Estruturais	i) Ordenação	Dicionário semasiológico
	ii) Eixo sintagmático	
	iii) Eixo paradigmático	
4) Funcionais	i) Dicionários de consulta	Dicionário de decodificação
	ii) Dicionários de aprendizagem	
	iii) Dicionários para profissionais	
5) Purista	i) Dicionários normativos	
	ii) Dicionários descritivos	Dicionário de uso
6) Formal	i) Dicionários impressos	
	ii) Dicionários digitais	Dicionário on-line

Fonte: Elaboração do autor com base em Rodríguez Barcia (2016, p. 98-99).

Destacamos, ainda, que o conjunto de critérios elaborados pela Lexicografia com vistas a realizar uma classificação tipológica dos dicionários se norteia por um tripé que, na realidade, além de subsidiar a avaliação dessas obras configura-se na essência do labor lexicográfico:

> Un análisis general permite comprobar cómo los criterios se distribuyen en aspectos internos y externos en relación a la elaboración, producción y distribución del diccionario. La finalidad, personas destinatarias y soporte de publicación condicionan notablemente los elementos definitorios de los diferentes diccionarios[9] (RODRÍGUEZ BARCIA, 2016, p. 95).

Desse modo, a qualidade de um dicionário gira em torno da tríade *finalidade*,

[9] "Uma análise geral permite verificar como os critérios estão distribuídos em aspectos internos e externos em relação à elaboração, produção e distribuição do dicionário. A finalidade, os destinatários e o suporte de publicação condicionam notavelmente os elementos definidores dos diferentes dicionários" (T.N.).

pessoas destinatárias e *suporte de publicação*. Outros critérios são estabelecidos a partir dessa base como anteriormente mencionado, ou seja, os objetivos da produção de uma obra lexicográfica devem ser estabelecidos conforme o perfil do usuário, levando em consideração o suporte que o dicionário assumirá.

Tendo em vista que a elaboração de um dicionário é norteada pelo tipo de usuário a que se destina, faz-se necessário estabelecer os destinatários que se beneficiarão da obra. Isso implica delimitar os objetivos de um dicionário a partir de seus usuários. No entanto, uma abordagem focada apenas nos objetivos é insuficiente para a avaliação de um dicionário. Nesse sentido, Tarp (2018, p. 244) argumenta que, para se definir ou classificar um dicionário de maneira genuinamente científica, sem correr o risco de deixar algum tipo de obra de fora, deve-se partir de três critérios básicos, ou seja: o *propósito*, o *conteúdo* e a *forma*.

Quanto ao *propósito*, uma obra lexicográfica é concebida como uma ferramenta nas mãos do usuário a fim de cumprir uma função específica, como é o caso de uma chave de fenda, de um martelo ou um alicate, ou seja, cada ferramenta é apropriada para executar um tipo específico de trabalho/função. Por sua vez, o *conteúdo* diz respeito ao dado organizado em formato lexicográfico que determinada ferramenta possibilita. Esse dado lexicográfico, que foi extraído de dados brutos e trabalhado para ganhar o formato lexicográfico, é interpretado pelo usuário que tem a responsabilidade de transformá-lo na informação que está procurando e isso implica que o consulente compreenda o funcionamento do verbete lexicográfico que está sendo consultado, isto é, o usuário deve saber manejar a ferramenta que está utilizando. Em relação à *forma*, o dicionário é visto como uma ferramenta que pode assumir o formato impresso ou eletrônico que, inclusive, pode ser associado a outros produtos/ferramentas digitais como é o caso dos corretores ortográficos (TARP, 2018, p. 244-245).

O conjunto dos critérios apresentados até aqui funcionam como instrumentos de avaliação lexicográfica, à medida que possibilitam o agrupamento dos dicionários a partir de suas semelhanças e diferenças, o que permite identificar pontos positivos e negativos de cada obra. Vale destacar que um critério não anula o outro. A contribuição dos lexicógrafos, nesse sentido, configura-se como uma prestação de serviços à Lexicografia e que podem ocorrer enfoques distintos, mas não excludentes, como é o caso dos critérios formulados por Tarp (2018, p. 247) para classificar dicionários a partir do *propósito*, que são divididos em quatro categorias principais, a saber:

> 1 Dictionaries with communicative functions, that is, dictionaries designed to assist their users in solving problems related to written and oral communication (text reception, text production, translation and text revision).
> 2 Dictionaries with cognitive functions, that is, dictionaries designed to transmit knowledge to their users.
> 3 Dictionaries with operative functions, that is, dictionaries designed to assist their users
> in performing specific types of action.
> 4 Dictionaries with interpretive functions, that is, dictionaries designed to assist their users in interpreting non-linguistic signs[10] (TARP, 2018, p. 247).

O enfoque desse autor na perspectiva do *propósito* se justifica pelo fato de uma obra lexicográfica ser concebida para atender a demandas específicas de consulta de uma classe específica de usuários. Com base nessas necessidades, a Teoria Funcional da Lexicografia[11] foi desenvolvida para dar suporte teórico-metodológico ao planejamento e produção de dicionários que funcionam como *ferramentas eletrônicas de consulta lexicográfica*.

A seguir, focalizamos as partes constituintes de um dicionário, bem como a terminologia criada para se referir a cada uma delas, de acordo com o arcabouço teórico da Lexicografia.

1.4. Megaestrutura, macroestrutura e microestrutura

Conhecer as partes constituintes de um dicionário não é saber destinado apenas aos lexicógrafos. Os usuários também devem compreender, ao menos em nível elementar, como um dicionário é arquitetado a fim de tirar o máximo de proveito da obra. Desse modo, o consulente terá uma experiência positiva ao compreender como manusear o dicionário, ou seja, como ler os conteúdos registrados em cada uma das distintas partes que formam essas obras.

---

[10] "1 Dicionários com funções comunicativas, ou seja, dicionários projetados para auxiliar seus usuários na resolução de problemas relacionados à comunicação escrita e oral (recepção de texto, produção de texto, tradução e revisão de texto).
2 Dicionários com funções cognitivas, ou seja, dicionários projetados para transmitir conhecimento aos seus usuários.
3 Dicionários com funções operativas, ou seja, dicionários destinados a auxiliar seus usuários na execução de tipos específicos de ação.
4 Dicionários com funções interpretativas, ou seja, dicionários destinados a auxiliar seus usuários na interpretação de signos não linguísticos" (T.N.).

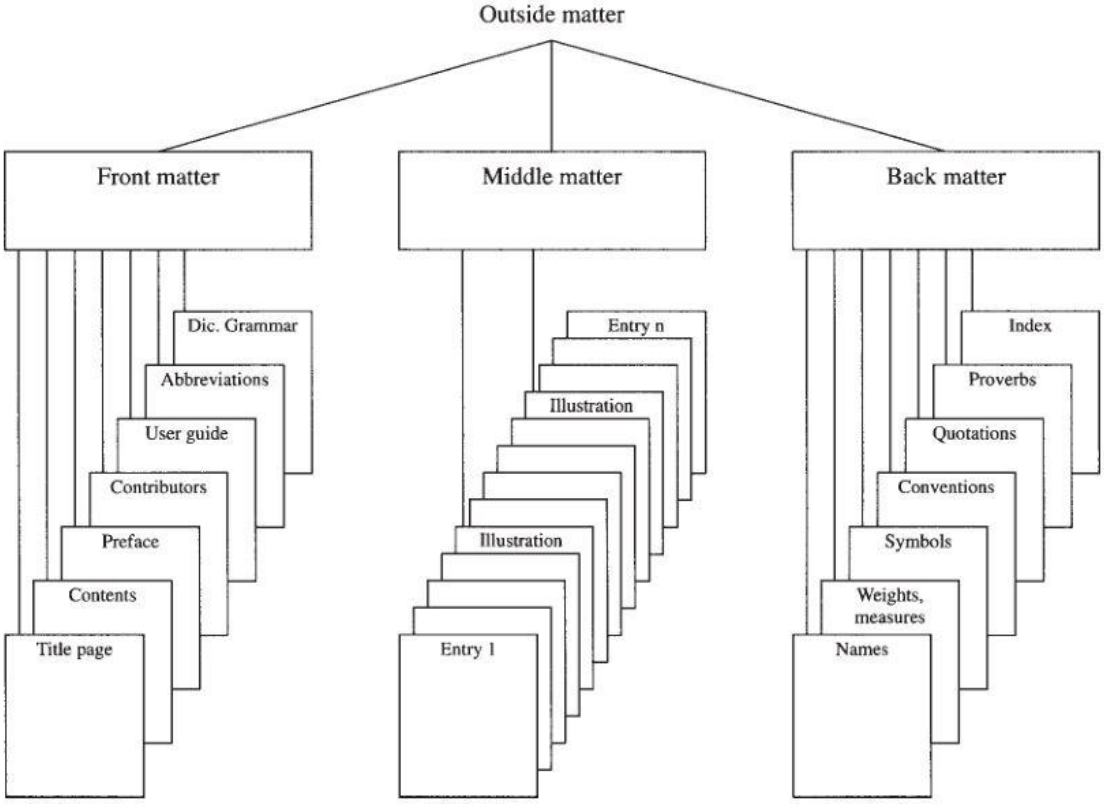
[11] Maiores detalhes são apresentados no item *2.8. Teoria Funcional da Lexicografia*.

Rey-Debove (1971, p. 21), por exemplo, concebe a macroestrutura como o conjunto de informações que podem ser lidas na vertical, enquanto a microestrutura é compreendida pelo conteúdo que pode ser lido na horizontal. Nessa perspectiva, as informações lidas na vertical são aquelas que figuram na macroestrutura como, por exemplo, o prefácio, as instruções ao leitor, a nomenclatura, ao passo que a leitura das informações na horizontal representa o conjunto de dados organizados após o lema.

Haensch (1997), por sua vez, organiza as informações dispostas nos dicionários a partir de duas classificações básicas, a saber: macroestrutura e microestrutura. Dessa forma, para esse autor, a macroestrutura corresponde à organização dos materiais que formam o corpo do dicionário como, por exemplo, o conjunto das entradas e demais informações como o prólogo, o prefácio, as instruções para o usuário, os anexos, as listas de abreviaturas etc. A microestrutura, por sua vez, é composta por todos os dados que compõem o verbete, podendo variar a depender da tipologia do dicionário como, por exemplo, o lema, a indicação de pronúncia, informações morfológicas, gramaticais entre outras (HAENSCH, 1997, p. 39-41).

Na mesma direção de organização das informações, Hartman; James (1998, p. 91; 94) e Hartmann (2001, p. 58) postulam que a microestrutura de um dicionário é formada pelas informações lexicográficas organizadas após a cabeça do verbete, ao passo que a macroestrutura é composta pela nomenclatura do dicionário. Esses autores ainda pontuam que a macroestrutura pode ser complementada pela *outside matter*, que é formada pela *front matter*, como é o caso do prefácio e das instruções de uso; pela *middle matter*, a exemplo das ilustrações; e da *back matter* no qual são registradas a bibliografia, os anexos etc. O conjunto da macroestrutura e da *outside matter* denomina-se megaestrutura, como é possível observar a seguir:

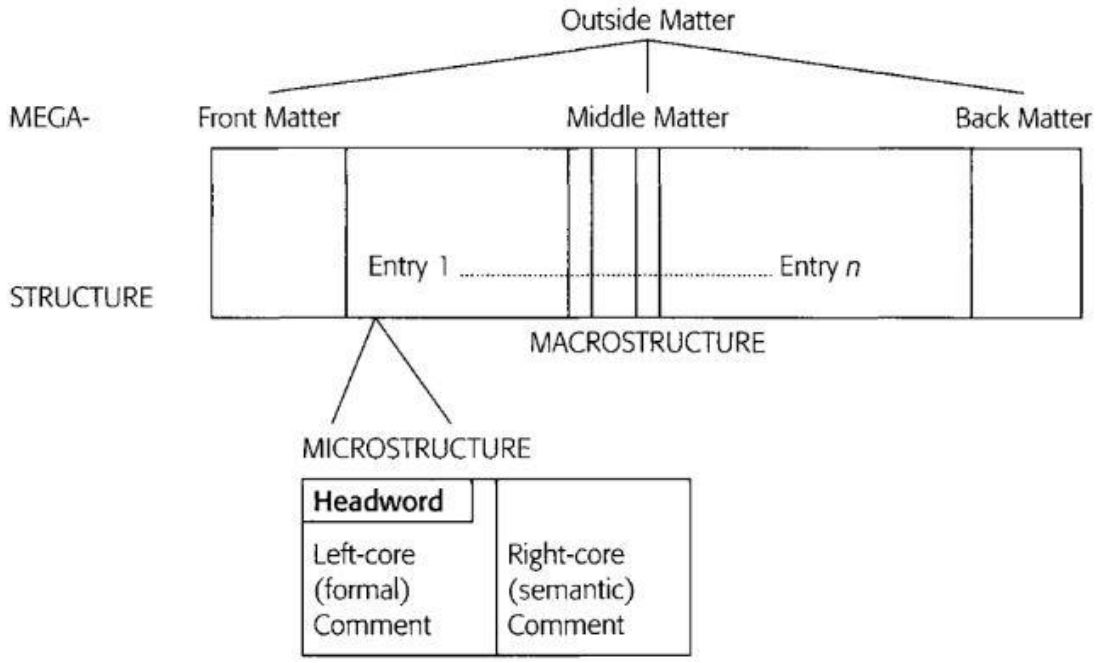
Figura 1: A megaestrutura lexicográfica.



Fonte: Hartman; James (1998, p. 92).

Ao se focalizar um verbete da macroestrutura de um dicionário sob o ângulo de uma lupa, algumas características próprias podem ser observadas como as mostradas na figura a seguir:

Figura 2: Mega-, macro- e microestrutura dos dicionários.



Fonte: Hartmann (2001, p. 59).

A Figura 2 mostra a diferença entre macroestrutura e microestrutura, além de especificar que o verbete é constituído por três elementos, a saber: o lema (*headword*); dados sobre a forma como, por exemplo, a ortografia, a morfologia, a pronúncia (*left-core*); informações semânticas como a definição, os exemplos (*right-core*). Essa organização dá indícios de uma hierarquia na organização dos elementos que compõem a microestrutura dos dicionários.

Porto Dapena (2002, p. 183), por sua vez, defende que a microestrutura de um dicionário é compreendida de duas partes, a saber: i) *Parte enunciativa*: composta do lema e as informações lexicográficas que antecedem à definição; ii) *Parte informativa*: destinado à definição e demais dados complementares como exemplos e remissivas.

Já Seco (2003, p. 25-29) explica que o verbete lexicográfico é formado por dois enunciados e que seus elementos ocupam um lugar fixo em relação à ordem em que aparecem na microestrutura dessas obras. Desse modo, o *primeiro enunciado* é formado pelo lema e seguido imediatamente por dados relacionados à etimologia, à classe gramatical, às informações cronológicas, aos dados do espaço geográfico de ocorrência e à atividade ou área social. Em seguida, o *segundo enunciado* é formado pela definição e, quando houver, pelos dados complementares (SECO, 2003, p. 25-26).

A proposta de Seco (2003) foi sintetizada por Pereira (2018, p. 43) em um quadro que permite uma visualização didática dos componentes que fazem parte do *primeiro enunciado* e do *segundo enunciado,* como é possível visualizar a seguir:

Quadro 2: Lugar fixo das informações do primeiro enunciado da microestrutura.

Primeiro enunciado	Palavra-entrada	1°	2°	3°	4°	5°	Segundo enunciado	6°
		Etimologia	Categoria gramatical	Vigência cronológica	Ambito geográfico	Ambito da atividade ou nível social	Definição	Informação adicional se necessário

Fonte: Pereira (2018, p. 43).

Destacamos, ainda, que nem todo dicionário apresenta todas as cinco informações lexicográficas em seu *primeiro enunciado*, dada a diversidade tipológica que essas obras podem assumir. Porém, quando são apresentadas ao usuário tendem seguir essa ordem.

O tópico a seguir aborda a questão da definição lexicográfica que se constitui em um elemento de grande importância nos dicionários.

1.5. Definição lexicográfica

Antes de adentrar ao tema das definições lexicográficas, é preciso considerar o fato de ocorrer, com frequência, uma incongruência quanto ao emprego dos termos *significado, sentido, acepção* e *definição*[12], como observa Medina Guerra (2003):

> En realidad, todo eso no es más que una consecuencia de la imprecisión terminológica que domina a la lexicografía; comprensible, por otro lado, si se tiene en cuenta que tan solo en una época relativamente reciente se ha comenzado a considerar a la lexicografía como parte da lingüística, como parte de una ciencia[13] (MEDINA GUERRA, 2003, p. 129).

Segundo a autora o *significado* é um valor que possui uma representação na língua por meio de uma unidade léxica enquanto o *sentido* é a variante de um significado. A acepção, por sua vez, é entendida como um sentido consolidado pelo uso

[12] A definição, em um sentido amplo que abrange todas as áreas do conhecimento, é discutida por Sager (2000) a partir de uma análise filosófica.

[13] "Na realidade, tudo isso não é mais que uma consequência da imprecisão terminológica que domina a Lexicografia; compreensível, por outro lado, si se tem em conta que somente em uma época relativamenterecente que se começou a considerar a Lexicografia como parte da Linguística, como parte de uma ciência" (T.N.).

e aceito por uma comunidade de falantes. Finalmente, a definição equivale ao modo de expressar, de descrever determinado sentido. Na Lexicografia, a definição representa o procedimento de catalogação de cada uma das acepções de um verbete (MEDINA GUERRA, 2003, p. 131).

A partir dessas considerações, notamos que o termo *significado* é utilizado de forma equivocada, quando se diz respeito à consulta de um verbete no dicionário. Trata-se de um uso que já está cristalizado na língua e pertence ao senso comum muito bem difundido.

Desse modo, no bojo das discussões lexicográficas, é preciso observar que o *significado* de um signo não representa um conceito, pois, sendo o *significado* um valor, pertence ao campo mental, intuitivo, motivado por inferências e conhecimento de mundo de cada falante. Nesse sentido, Ramón Trujillo (1988, p. 9 *apud* MEDINA GUERRA 2003, p. 131) pondera que:"Sólo se pueden describirse o definirse cada uno de los usos de un signo. Por ello diremos que el diccionario no puede registrar más que acepciones, variantes o usos, pero no significados[14]."

Realizadas essas ponderações, passamos a tratar da definição lexicográfica[15] que, para Porto Dapena (2002, p. 269), é "[...] todo tipo de equivalencia establecida entre la entrada y cualquier expresión explicativa de la misma en un diccionario monolingüe.[16]" Complementa ainda o mesmo autor que esse tipo de relação equivalente é construída por meio de dois elementos: "[...]el **definido** o **definiendum**, representado por la entrada del artículo lexicográfico, y el **definidor** o **definiens**, que es la expresión explicativa y que en el lenguaje corriente lamamos más especificadamente también **definición**[17]" (PORTO DAPENA, 2002, p. 269).

No labor lexicográfico, o texto definitório exige uma atenção especial, pois se trata de uma informação lexicográfica importante. Desse modo, entender a complexidade que envolve o tema é o primeiro passo para compreender os equívocos que, frequentemente, são identificados em definições lexicográficas.

---

[14] "Só se pode descrever ou definir cada um dos usos de um signo. Por isso diremos que o dicionário só pode registrar acepções, variantes ou usos, e não significados" (T.N.).

[15] A definição lexicográfica também é discutida por outros autores como, por exemplo, em Bosque (1982, p. 105-123); Bajo Pérez (2000, p. 35-52); Medina Guerra (2003, p. 129-146); Seco (2003 p. 25-58).

[16] "[...] todo tipo de equivalência estabelecida entre a entrada e qualquer expressão explicativa da mesma em um dicionário monolíngue" (T.N.).

[17] "[...] o **definido** ou **definiendum,** representado pela entrada (verbete), e o **definidor** ou **definiens**, que é a expressão explicativa e que na linguagem corrente chamamos mais especificadamente também **definição**" (T.N.).

Bajo Pérez (2000, p. 35), por exemplo, chama a atenção para o princípio aristotélico, segundo o qual uma definição deve partir de um gênero próximo e de uma diferença específica, ou seja:

> [...] para definir, se recurre primero al hiperónimo inmediatamente superior después se señalan las características adicionales: por ejemplo, <<rosa>> se define así en el DRAE: flor (hiperónimo) del rosal, notable por su belleza, la suavidad de su fragancia su color, generalmente encarnado poco subido… (características adicionales)[18] (BAJO PÉREZ, 2000, p. 35).

Na verdade, construir uma relação entre definido e definidor não é tarefa simples e são muitos os problemas identificados pela metalexicografia[19]. No âmbito da definição lexicográfica, por exemplo, é possível afirmar que uma definição bem escrita deve ser:

> [...] **completa** – no puede faltar ningún rasgo característico, pero tampoco debe ser demasiado amplia –, **no circular** – lo definido no debe entrar en la definición –, **no negativa** – no debe señalarse lo que lo definido no es, sino lo que es –, **no metafórica ni figurada** – el lenguaje figurado o metafórico resultaría impreciso y equívoco, por lo que no permitiría identificar lo definido[20] (BAJO PÉREZ, 2000, p. 35).

Somamos a essas condições os apontamentos de Medina Guerra (2003, p. 132-133) que, ao discutir a complexidade do labor da escrita da definição lexicográfica, lista os seguintes princípios:

> 1. La unidad léxica definida no debe figurar en la definición.
> 2. La definición no debe translucir ninguna ideología.
> 3. La definición debe participar de las características de la lengua de su época y las palabras con que se codifique han de ser sencillas a la vez que claras y precisas[21] (MEDINA GUERRA, 2003, p. 133).

Entre tantos parâmetros estabelecidos para aperfeiçoar a qualidade das

[18] "[...] para definir, se recorre primeiro ao hiperônimo imediatamente superior, depois são indicadas as características adicionais: por exemplo, <<rosa>> é definido assim no DRAE: flor (hiperônimo) da roseira, notável por sua beleza, suavidade de sua fragrância, sua cor geralmente vermelho escuro (características adicionais)" (T.N.).

[19] Compreendida como a parte da lexicografia teórica que estuda os dicionários em um amplo contexto, a fim de identificar problemas metodológicos e apresentar soluções.

[20] "[...] **completa** – não pode faltar nenhum traço característico, mas também não deve ser muito ampla –, **não circular** – o definido não deve entrar na definição –, **não negativa** – não se deve apontar o que o definido não é, mas o que é –, **não metafórica nem figurada** – a linguagem figurada ou metafórica seria imprecisa e equivocada, por tanto não permitiria a identificar o que é o definido" (T.N.).

[21] "1. A unidade léxica definida não deve figurar na definição.
2. A definição não deve transmitir nenhuma ideologia.
3. A definição deve utilizar as características da língua de seu tempo e as palavras utilizadas devem ser simples, claras e precisas" (T.N.).

definições lexicográficas, há um teste simples, chamado de teste de substituição ou comutação, que pode ser aplicado para avaliar um texto definitório. Conforme Seco (2003, p. 32), "Si el enunciado definidor puede sustituir al término definido, en un enunciado de habla, sin que el sentido objetivo de este se altere, el enunciado definidor es válido[22]". Vale acrescentar que, nessa prova, os itens lexicais a serem substituídos devem pertencer à mesma classe gramatical, ou seja, um substantivo substituirá um substantivo e assim por diante.

Porto Dapena (2002, p. 310), por seu turno, já ponderara que a prova da comutação é utilizada para verificar em que medida o texto definitório é equivalente semanticamente ao verbete em questão, pois na definição ocorrem dois enunciados: o enunciado parafrástico e o contorno que são elementos acrescentados para complementar o contexto da definição. O autor exemplifica essa relação com o seguinte verbete: "**Confluir**. *Intr*. Juntarse dos o más ríos u otras corrientes de agua en un mismo lugar." Assim, é possível observar que *juntarse* constitui o enunciado parafrástico, enquanto as demais informações representam o contorno definicional. Ou seja, ao aplicar o teste de comutação é possível verificar que o verdadeiro argumento dessa definição é *juntarse*, pois pode ser substituído perfeitamente num enunciado qualquer como, por exemplo: "Aquí confluyen el Miño y el Sil = Aquí se juntan el Miño y el Sil" (PORTO DAPENA, 2002, p. 310).

Outro exemplo da aplicação do teste de comutação é encontrado em Sepulveda (2006, p. 117) a partir do verbo *guardar*, em espanhol: "**guardar** conservar o retener [una cosa]". Nesse exemplo, o verdadeiro conteúdo expresso pelo verbo *guardar* é o enunciado *conservar* ou *retener*, porque é a única porção do texto definitório que pode ser substituída pelo definido em um texto real de fala. Isto é, ao aplicar o teste de comutação a partir de um texto real de fala, temos: "La viuda **guarda** o **conserva** todos los manuscritos". No entanto, se no teste de comutação o definido for substituído por todo o texto definitório, incluindo o contorno, sinalizado pelos colchetes, a sentença reproduziria uma agramaticalidade: "La viuda **conserva** o **guarda** [uma cosa] todos los manuscritos" (SEPULVEDA, 2006, p. 117).

Vale destacar que, mesmo sendo a prova da substituição ou comutação um teste não aplicável a todos os tipos de definições, esse método se constitui numa maneira de

[22] "Se o enunciado definidor puder substituir o termo definido, em um enunciado de fala, sem que o sentido objetivo deste se altere, o enunciado definidor é válido" (T.N.).

verificar a qualidade do texto definitório. Somamos a isso o fato de que é preciso ter coerência, sistematicidade e rigor metodológico na produção lexicográfica (MEDINA GUERRA, 2003, p. 138).

Na sequência, aprofundamos o tema das definições lexicográficas, a partir da posição de Seco (2003) e do detalhamento da tipologia proposta por Porto Dapena (2002), bem como apresentamos o tipo de definição adotada para integrar a microestrutura do protótipo do *VoDiNorte*.

1.6. Tipologia das definições lexicográficas

Segundo Seco (2003, p. 34), há dois tipos básicos de definição lexicográfica, a saber: i) *definições em metalinguagem do conteúdo* que buscam descrever um referente e são formadas por todos os substantivos e a maioria dos adjetivos, verbos e advérbios; ii) *definições em metalinguagem do signo*, isto é, uma classe de definição considerada imprópria, pois sua finalidade não é definir um referente, mas oferecer explicações sobre a função das interjeições e de vocábulos gramaticais como, por exemplo, preposições, conjunções, pronomes e artigos.

Ainda, de acordo com esse autor, uma diferença distintiva entre esses dois tipos de definições lexicográficas é que apenas as definições em metalinguagem do conteúdo aceitam o teste de substituição (SECO, 2003, p. 34).

O conjunto das *definições em metalinguagem do conteúdo* e das *definições em metalinguagem do signo* foram, sistematicamente, descritas e exemplificadas por Porto Dapena (2002, p. 266-290, cuja proposta foi posteriormente sintetizada por Pereira (2018, p. 46- 50). O quadro a seguir sintetiza essas informações.

Quadro 3: Tipologia das definições lexicográficas.

<table>
<tr><td rowspan="13">Metalinguagem do conteúdo</td><td rowspan="3">Enciclopédica</td><td colspan="4">Descritiva</td></tr>
<tr><td colspan="4">Teleológica</td></tr>
<tr><td colspan="4">Genética</td></tr>
<tr><td>Ostensiva</td><td colspan="4"></td></tr>
<tr><td rowspan="9">Conceitual</td><td rowspan="7">Perifrástica</td><td rowspan="5">Substancial</td><td colspan="2">Includente positiva</td></tr>
<tr><td colspan="2">Includente negativa</td></tr>
<tr><td colspan="2">Excludente ou antonímica</td></tr>
<tr><td colspan="2">Participativa ou metonímica</td></tr>
<tr><td colspan="2">Aproximativa ou analógica</td></tr>
<tr><td colspan="2">Aditiva</td></tr>
<tr><td>Relacional</td><td colspan="2">Morfossemântica</td></tr>
<tr><td rowspan="3">Sinonímica</td><td>Sinonímica propriamente dita</td><td>Complexa</td><td>Mista</td></tr>
<tr><td colspan="3">Parassinonímica</td></tr>
<tr><td colspan="3">Pseudoperifrástica</td></tr>
<tr><td rowspan="3">Metalinguagem do signo</td><td rowspan="3">Funcional</td><td colspan="4">Morfossintática</td></tr>
<tr><td colspan="4">Contextual</td></tr>
<tr><td colspan="4">Pragmática</td></tr>
<tr><td></td><td>Hibrida</td><td colspan="4"></td></tr>
</table>

Fonte: Porto Dapena (2002) e Pereira (2018).

Com base no quadro 3, segue uma breve descrição de cada tipologia mencionada. Na medida do possível, substituímos os exemplos originalmente apresentados por Porto Dapena (2002) e por Pereira (2018), em língua espanhola, por exemplos em língua portuguesa, retirados do Dicionário Houaiss (2009)[23].

A definição *enciclopédica,* também chamada de definição das coisas, busca descrever o definido de forma pormenorizada. Essa definição pode ser do tipo *descritiva*, *teleológica* ou *genética*. Desse modo, a definição *descritiva* não deve responder *que é o definido,* mas, sim, c*omo é o definido.* (figura 3):

[23] Obra de referência no tocante à língua portuguesa do Brasil.

Figura 3: Verbete *cavalo* no Dicionário Houaiss (2009).

Fonte: Houaiss (2009)

Na definição de *cavalo* apresentada no Houaiss (2009), observamos que *mamífero perissodátilo da fam. dos equídeos* tem caráter de definição linguística, pois é bastante pontual e objetiva. Porém, as demais características que descrevem o referente revelam que se trata de uma definição predominantemente *enciclopédica descritiva*.

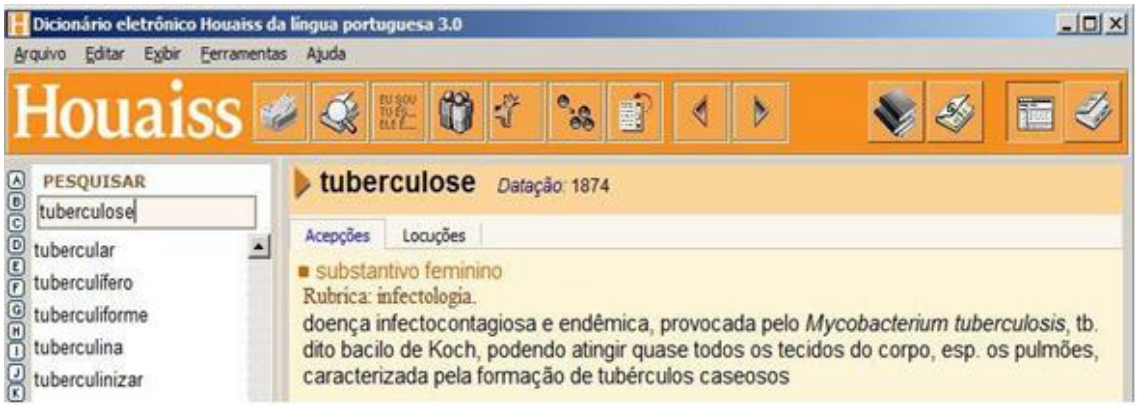
Já a definição *teleológica* caracteriza o objeto por sua finalidade ou seu destino como, por exemplo, no verbete *barômetro* que descreve a utilidade do instrumento (figura 4):

Figura 4: Verbete *barômetro* no Dicionário Houaiss (2009).

Fonte: Houaiss (2009).

Percebemos, na figura 5, que a definição g*enética* caracteriza o objeto por sua origem ou causa, como se pode observar no verbete *tuberculose*, ilustrado a seguir, em que há menção do agente biológico causador da doença:

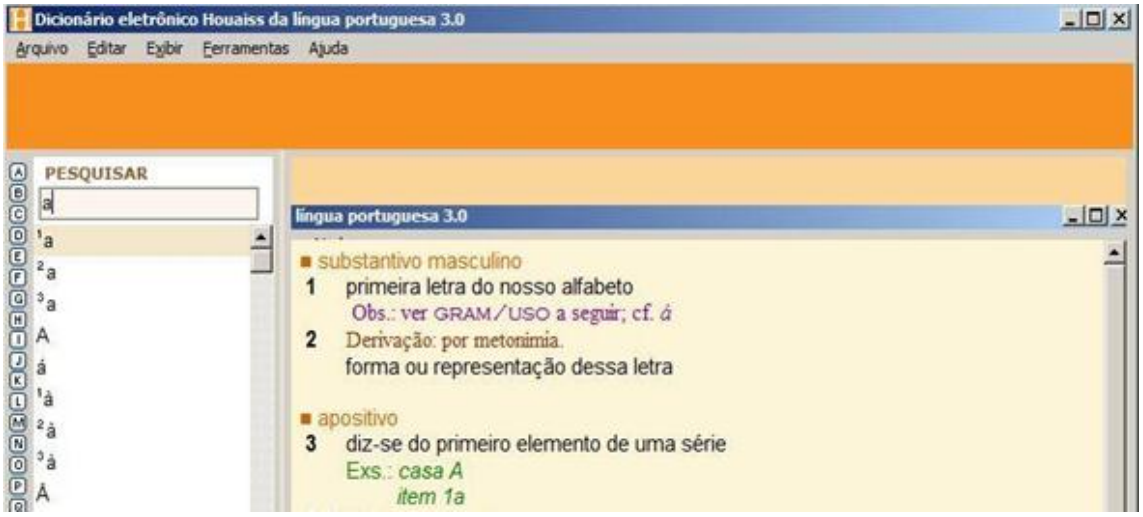
Figura 5: Verbete *tuberculose* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009)

A *definição ostensiva*, por sua vez, não é propriamente uma definição, já que se caracteriza por transpor integralmente ou parcialmente o referente no lugar do definidor, como pode ser observado na terceira acepção do verbete apresentado na figura 6.

Figura 6: Verbete *a* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

Há ainda casos de *definições ostensivas* do tipo icônicas que recorrem a recursos imagéticos como, por exemplo, ilustrações e fotografias muito utilizadas nas enciclopédias. Nesse sentido, devemos observar as construções produzidas mediante imagens verbais que remetem a um conhecimento do mundo externo e geral do consulente, a fim de definir o referente, como podemos observar no verbete apresentado

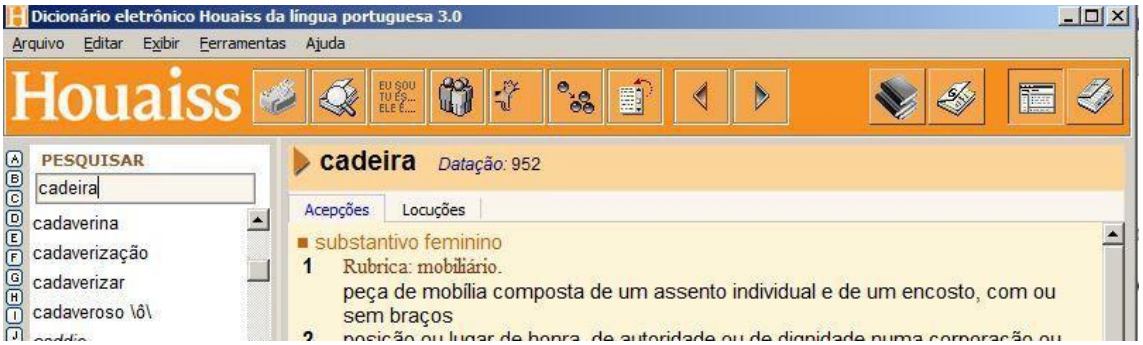
na figura a seguir:

Figura 7: Verbete *verde* no Dicionário Houaiss (2009).

Fonte: Houaiss (2009)

Constatamos, a partir da figura 7, que a definição do item léxico *verde* conta com o conhecimento de mundo do consulente ao associar essa cor com a relva. A *definição conceitual*, por sua vez, é formada pela metalinguagem do conteúdo,ou seja, é utilizada uma equivalência para expressar um conceito, como se verifica no verbete *cadeira*, ilustrado a seguir (figura 8).

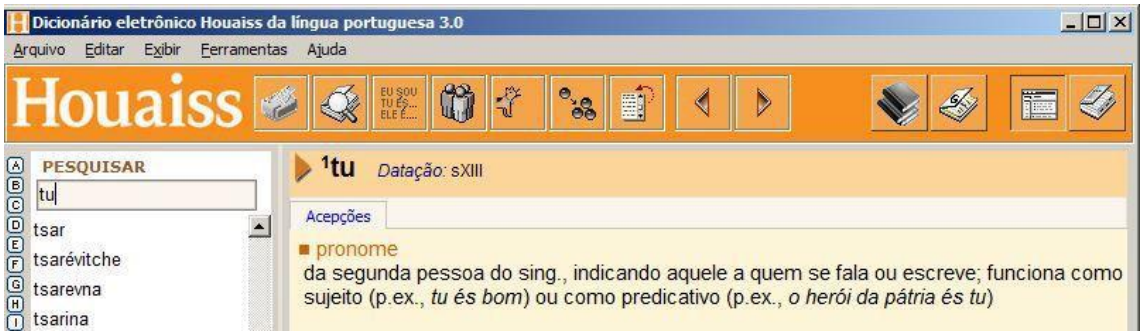
Figura 8: Verbete *cadeira* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

Ao seu turno, a *definição funcional* – chamada de *imprópria* em entradas de conteúdo gramatical, por falta de significado léxico – é escrita em metalinguagem do signo, pois sua finalidade é explicar a função do definido como podemos observar a seguir, por meio do verbete *tu* (figura 9).

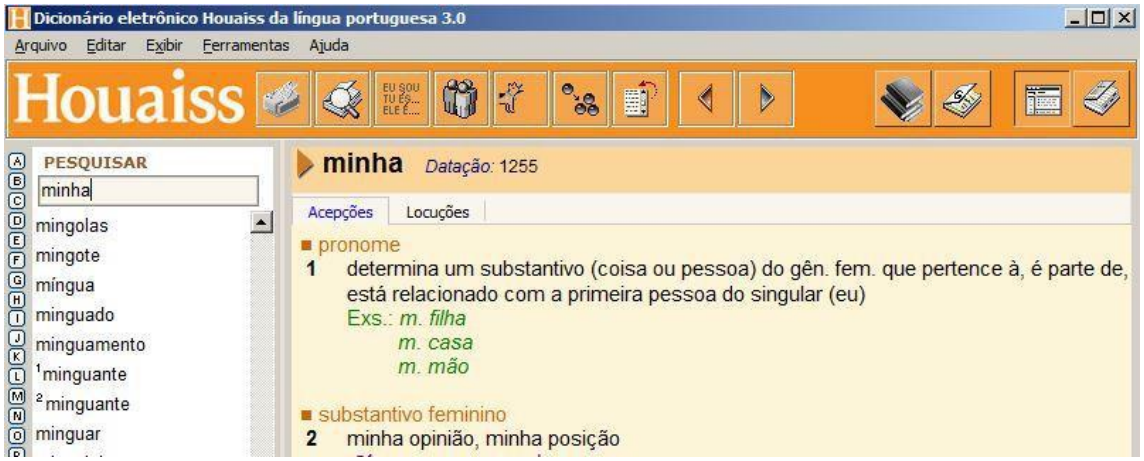
Figura 9: Verbete *tu* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

Destacamos, ainda, que a *definição funcional*, por apresentar informações sobre o funcionamento gramatical, contextual e pragmático de um definido, se distribui em três subtipos, a saber: morfossintático, contextual e pragmático. A seguir, temos um exemplo de *definição funcional morfossintática* (figura 10).

Figura 10: Verbete *minha* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

O verbete *minha*, apresentado na figura 10, ilustra uma acepção com uma descrição morfossintática e exemplos. A seguir, a *definição funcional contextual*, é exemplificada com o verbete apresentado na figura 11.

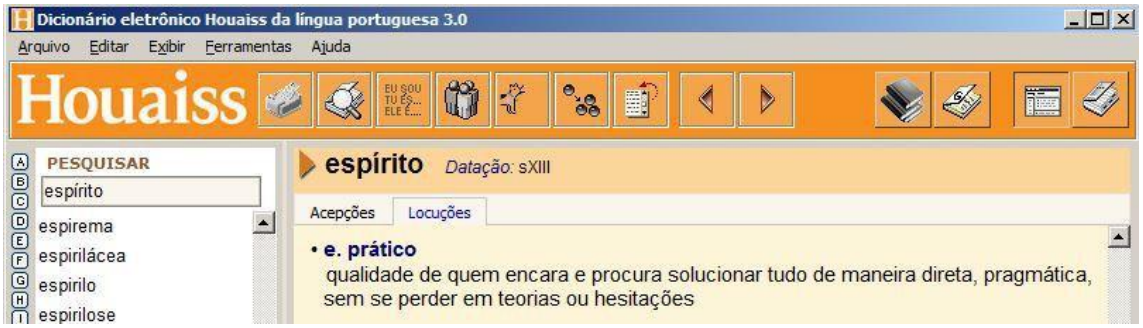
Figura 11: Verbete *batido* no Dicionário Houaiss (2009).

Fonte: Houaiss (2009).

As acepções constantes na figura 11 reúnem diferentes contextos de uso de *batido*. Desse modo, há uma acepção para descrever o respectivo contexto, seguido de um exemplo.

Já a *definição funcional pragmática*, que evidencia inferências construídas por determinado contexto de fala, pode ser observada a partir da locução apresentada na figura 12.

Figura 12: Verbete *espírito prático* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

Há, ainda, definições do tipo *híbridas* que combinam enunciados de *definições conceituais* e *funcionais*. Essa mescla pode ser encontrada na *definição funcional* do

tipo *contextual*, que geralmente são iniciadas com enunciados do tipo *nome de*, *diz-se de* ou *aplica-se* etc. Exemplo:

> **Ene.** Nombre de la letra n.
> **Gitano.** Dícese de los individuos de un pueblo originario de la India, extendido por gran parte de la Europa, que mantienen en gran parte un nomadismo y han conservado rasgos físicos y culturales propios.
> **Diablo.** Nombre general de los ángeles arrojados al abismo, y de cada uno de ellos (PORTO DAPENA, 2002, p. 283).

O autor explica que essas definições ficariam melhores se as expressões introdutórias fossem retiradas. Desse modo, sugere a seguinte reescrita: "**Ene.** Letra n. **Gitano.** Individuo de un pueblo originario de la India… **Diablo.** Ángel arrojado al abismo" (PORTO DAPENA, 2002, p. 283).

Nesse impasse, a preferência do lexicógrafo deve recair na definição conceitual ao invés de na definição funcional. Todavia, nem sempre isso é possível e, incluso, pois há situações em que não há remédio a não ser recorrer à definição híbrida. Exemplo: "**Aguileño.** Dícese del rostro largo delgado. **Atenorado.** Dícese de la voz parecida a la de tenor de los instrumentos cuyo sonido tiene timbre parecido" (PORTO DAPENA, 2002, p. 284).

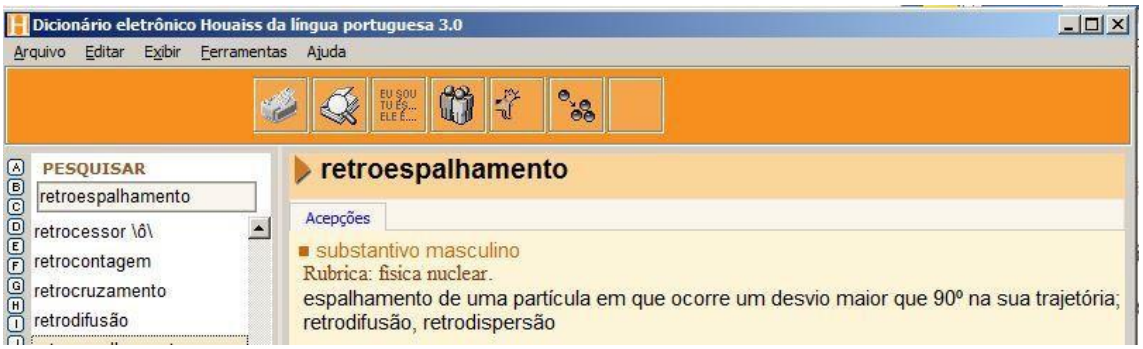
Nos exemplos anteriores, explica o autor, se os enunciados introdutórios (*dícese del rostro* e *dícese de la voz*) fossem retirados, a definição conceitual de cada acepção não seria suficiente para esclarecer ao consulente o contexto exato de cada definidor. Porém, nos referidos exemplos, é possível separar a parte funcional da parte conceitual, que é considerada a verdadeira definição, resultando em verbetes mais claros. Por exemplo: "**Aguileño.** Largo y delgado. Se disse del rostro. **Atenorado.** Parecido a la voz del tenor. Se dice de la voz de instrumentos musicales" (PORTO DAPENA, 2002, p. 284).

A *definição conceitual* é subdividida em *perifrástica* (composta por um enunciado frasal) e *sinonímica* (que busca aproximar o definidor e o definido por meio de um sinônimo). Por sua vez, a *definição conceitual perifrástica* se divide em *substancial* e *relacional*. A *substancial* apresenta seis classificações, a saber: *includente positiva*, *includente negativa*, *excludente* ou *antonímica*, *participativa* ou *metonímica*, *aproximativa* ou *analógica* e *aditiva*. Por sua vez, a *relacional* contém uma classificação *morfossemântica*.

A *definição perifrástica sinonímica*, por sua vez, é subdividida em *sinonímica propriamente dita*, *parassinonímica* e *pseudoperifrástica*.

A *definição sinonímica* propriamente dita é aquela que apresenta uma equivalência sinonímica válida. Quando é composta por um único sinônimo é chamada *simples* e quando é formada por mais de um sinônimo recebe o nome de *complexa* ou *cumulativa*. Há, ainda, a definição que reúne uma *definição sinonímica* e uma *definição parafrástica* que, neste caso, é denominada *mista* como podemos observar no exemplo a seguir (figura 13).

Figura 13: Verbete *retroespalhamento* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

A *definição parassinonímica* é aquela que não atende ao princípio da equivalência semântica e, portanto, não é aceitável. Assim, há casos em que o definidor não representa uma definição sinonímica, mas um hiperônimo, como no exemplo apresentado por Porto Dapena (2002, p. 287): "**Maroma.** Cuerda". Nesse caso, o vocábulo *maroma* significa *cuerda gruesa*. Ou seja, *cuerda* é hiperônimo de *maroma*. Essa relação também pode ocorrer quando o texto definitório representa um hipônimo da entrada como, por exemplo: "**Calzado.** Zapato o alpargata". Nessa definição *zapato* e *alpargata* funcionam como hipônimos de *calzado* (PORTO DAPENA, 2002, p. 287).

A *definição pseudoperifrástica*, por sua vez, é aquela que reúne um contexto, chamado de *contorno definicional*, que complementa o sentido expresso pelo termo sinônimo, como podemos observar a seguir (figura 14).

Figura 14: Verbete *forjar* no Dicionário Houaiss (2009).

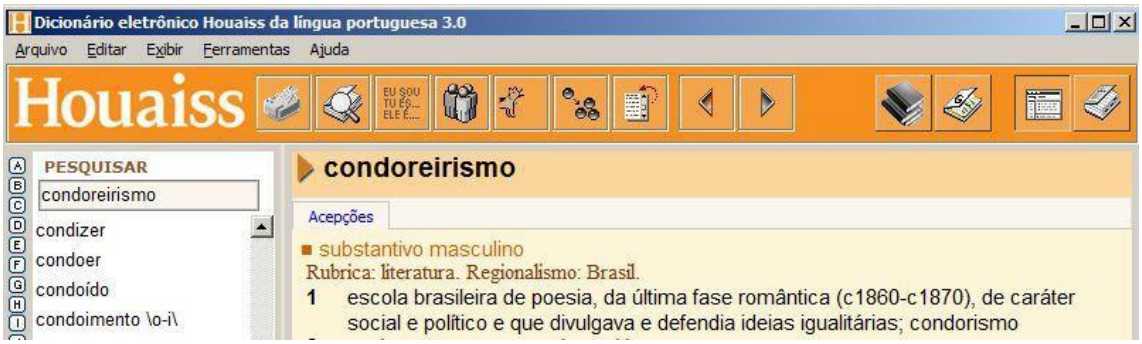
Fonte: Houaiss (2009).

Observamos na figura 14, mais especificadamente na segunda acepção do verbete *forjar*, que os sinônimos *modelar* e *fabricar* constituem o eixo central do sentido do item lexical em questão. Desse modo, os demais enunciados são classificados como contextuais ou elementos do *contorno definicional*.

A *definição perifrástica relacional* busca estabelecer uma relação entre definido e definidor por meio de outros elementos relacionados gramaticalmente. Ou seja, o definidor inicia-se com um núcleo da mesma categoria gramatical do definido e é acrescido de complementos ou, ainda, há casos em que não há um núcleo, isto é, há somente um elemento que transpõe – por meio de um pronome relativo ou preposição – o sentido do verbete para um sintagma transposto – oração ou sintagma nominal. Essa relação pode ser observada em "**Imparcial.** Que juzga o procede con imparcialidad. **Honestamente.** Con honestidade", Vale acrescentar que a correspondência total ou parcial entre a entrada e o texto definitório recebe a classificação de *definição relacional morfossemântica*, quando se trata de definido formado por unidades lexicais compostas ou derivadas. Por exemplo: **"Ilegítimo.** No legítimo. **Barbiluengo.** Que tiene la barba larga" (PORTO DAPENA, 2002, p. 291-292).

A *definição substancial includente positiva* ou *hiperonímica* é aquela construída nos moldes aristotélicos, em que o definidor é escrito por meio de um gênero próximo (hiperônimo) e de uma diferença específica (elementos que concretizam o significado do definido). Por exemplo:

Figura 15: Verbete *condoreirismo* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

Observamos, por meio da Figura 15, que o enunciado *escola brasileira de poesia* representa o gênero próximo e as demais informações representam a diferença específica. Vale observar que esse é um tipo considerado ideal de definição lexicográfica.

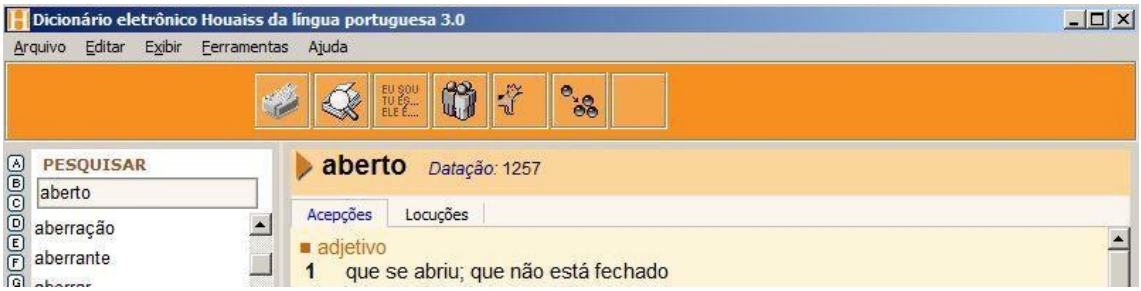
A *definição perifrástica substancial includente negativa* é aquela em que o elemento includente lógico possui um sentido negativo, como se observa no verbete a seguir (figura 16).

Figura 16: Verbete *amoralismo* no Dicionário Houaiss (2009).

Fonte: Houaiss (2009).

Por sua vez, a *definição perifrástica substancial excludente* ou *antonímica* é formada por uma partícula negativa acrescida de um antônimo do definido. Essa definição pode ocorrer por meio de um simples antônimo ou por meio de um enunciado frasal, como apresentado na figura 17:

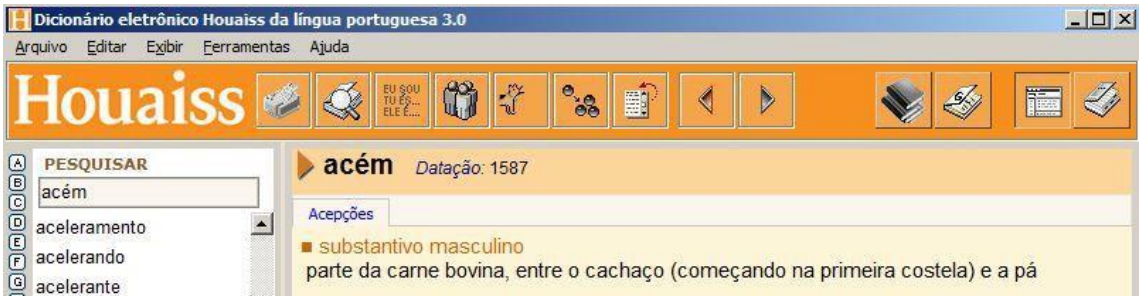
Figura 17: Verbete *aberto* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

A *definição perifrástica substancial participativa* ou *metonímica* é aquela que tem início por um enunciado que busca situar o verbete como a parte de um todo. Essa definição geralmente utiliza expressões como, por exemplo, *parte de*, *cada um dos*, *peça de* etc. O verbete a seguir ilustra tal construção (figura 18).

Figura 18: Verbete *acém* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

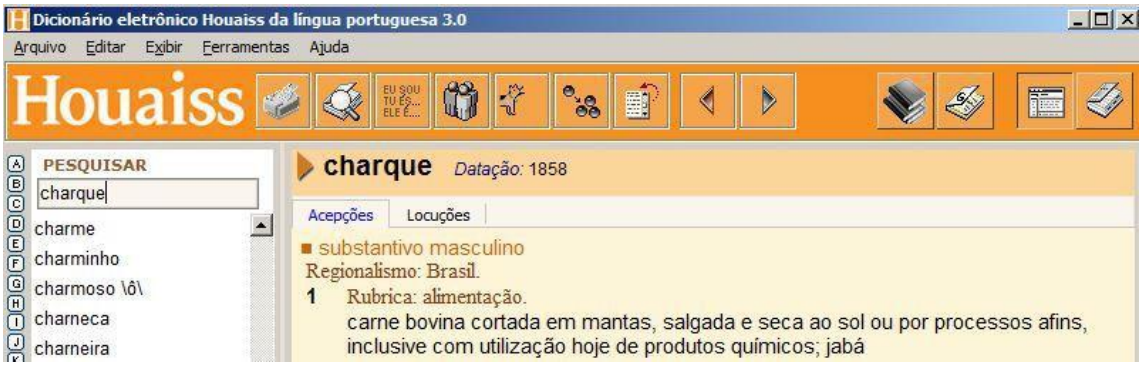
Já a *definição perifrástica substancial aproximativa* ou *analógica* descreve o definido por meio de uma relação de aproximação ou semelhança. Normalmente a definição começa com a expressão *espécie de* ou *tipo de*. Essa construção pode ser observada na figura 19, a seguir:

Figura 19: Verbete *saltério* no Dicionário Houaiss (2009).

Fonte: Houaiss (2009).

Por fim, a *definição perifrástica substancial aditiva* é aquela em que ocorre uma descrição aditiva do verbete por meio de uma associação de vários lexemas, unidos sintaticamente por coordenação copulativa. Ou seja, há a soma de dois ou mais enunciados ligados por preposição de valor aditivo como se pode constatar no verbete *charque* (figura 20):

Figura 20: Verbete *charque* no Dicionário Houaiss (2009).



Fonte: Houaiss (2009).

No verbete apresentado na Figura 20, é possível observar que, além da vírgula após a unidade lexical *mantas*, há uma segunda adição representada sintaticamente pela conjunção *e*, acrescentando o complemento *seca ao sol*.

Conforme o exposto é possível observar que a definição lexicográfica se apresenta ao consulente de variadas maneiras e, nesse sentido, nem sempre o texto definitório é redigido adequadamente. Sabemos que não há obra isenta de falhas e que, na realidade, elas devem ser encaradas como alertas, pois toda crítica construtiva é bem-vinda e deve ser vista como uma oportunidade para melhorar, cada vez mais, o produto lexicográfico.

No que diz respeito às definições do protótipo do *VoDiNorte*, buscamos seguir o modelo aristotélico, em que o definidor é descrito por meio de um gênero próximo e de uma diferença específica como, por exemplo, a definição do verbete "**carapanã:** Inseto de pequeno porte que pica e produz um zumbido agudo enquanto voa". Nessa definição, o gênero próximo representa o hiperônimo *inseto* e as diferenças específicas são formadas pelo restante do enunciado *de pequeno porte que pica e produz um zumbido agudo enquanto voa.* Destacamos, ainda, que a segunda parte desse texto definitório corresponde a uma *definição perifrástica substancial aditiva* em que a presença da conjunção aditiva *e* acrescenta uma segunda informação de cunho específico sobre *inseto,* isto é, *produz um zumbido agudo enquanto voa*.

É possível observar a partir dos tópicos apresentados neste capítulo que léxico e Lexicografia são temas que se entremeiam de forma muito natural e que as obras lexicográficas são instrumentos de consultas lexicais que possuem uma variedade de critérios que norteiam a qualidade, a crítica e a evolução dessas obras. No capítulo seguinte, discutimos o desenvolvimento lexicográfico sob o viés da tecnologia, ou seja, as contribuições que as novas tecnologias computacionais têm propiciado ao fazer dicionarístico.

CAPÍTULO 2 – CAMINHOS PARA UMA LEXICOGRAFIA ELETRÔNICA

Atualmente, a Lexicografia vive uma mudança de paradigma no qual o tradicional, herdado por séculos do labor lexicográfico, cede espaço para o digital em que os dicionários se configuram como ferramentas de consulta a dados tratados lexicograficamente. Neste capítulo, abordamos questões pontuais relacionadas às mudanças em curso na Lexicografia a partir de uma perspectiva computacional, que tem revolucionado o planejamento e a produção de ferramentas de consulta lexicográfica, sem perder o foco no usuário, que precisa ter suas necessidades de pesquisa atendidas satisfatoriamente.

2.1. Dicionários impressos *versus* dicionários eletrônicos

Quando se pensa no contraste existente entre os dicionários impressos e os dicionários eletrônicos algumas diferenças básicas podem ser inferidas, dada as particularidades evidentes que o formato em papel e o meio eletrônico assumem.

A primeira evidência, mais perceptível, são os benefícios dos dicionários eletrônicos em relação ao acesso e à capacidade de armazenamento. Todavia, destacamos que o acesso rápido a uma quantidade de informações lexicográficas que não pesam no bolso ou que não ocupam lugar na estante não são as únicas facetas da Lexicografia Eletrônica. Assim, as discussões que envolvem a evolução dos dicionários impressos para os dicionários eletrônicos perpassam por três questões elementares, mencionadas anteriormente no tópico *1.3. Tipologia lexicográfica* deste trabalho, a saber: a forma, o conteúdo e o propósito (TARP, 2018, p. 246).

Para tanto, a fim de compreender como essas questões são postas como base para a reflexão que discute as mudanças que estão em curso na Lexicografia, em que o dicionário em papel está cedendo espaço para o dicionário eletrônico, recorremos a uma definição de *dicionário* a partir da perspectiva da Lexicografia Impressa:

> [...] el diccionario es una obra de consulta consistente en una descripción atomística del léxico y determinado, a su vez, por cuatro factores variables: a) el número y extensión de sus entradas, b) el modo de estudiarlas, c) la ordenación aplicada a las mismas, y,

> finalmente, d) el soporte de dicha descripción[24] (PORTO DAPENA, 2002, p. 42).

Na primeira parte dessa definição, o autor procura conceituar o dicionário como uma obra de consulta que descreve o léxico de forma atomística, ou seja, levando em consideração que o átomo é a menor partícula de um corpo, a função do dicionário, nessa perspectiva, é a de oferecer um conjunto de informações detalhadas de um conteúdo lexical. Na segunda parte da definição, por sua vez, Porto Dapena (2002) esclarece que as informações lexicográficas podem ser apresentadas ao consulente de formas variadas.

O fato é que as contribuições de Porto Dapena (2002) são ricas para a Lexicografia Impressa e insuficientes para a perspectiva computacional que, por sua vez, demanda uma discussão ampliada que envolve o usuário e o próprio contexto tecnológico no qual vivemos que, por sua vez, tem revolucionado o fazer dicionarístico.

Desse modo, com base na premissa lexicográfica de que um dicionário deve atender as necessidades de consulta de usuário que apresentem demandas distintas de pesquisa e, ao mesmo tempo, considerando as possibilidades que as novas tecnologias computacionais oferecem ao labor lexicográfico, focalizamos um conceito de *dicionário* construído à luz da Lexicografia Eletrônica:

> A dictionary is a utility tool, which is conceived for consultation with the genuine purpose of meeting punctual information needs experienced by specific types of potential user in specific types of extra-lexicographical context, and which is designed to assist its users by providing manual or automatic access to lexicographically prepared data, which can either be used directly by the users in order to retrieve the required information which they can subsequently use to solve specific problems in the context where the needs originally occurred, or by a digital tool in order to make automatic corrections in texts or translations produced by the users of this tool[25] (TARP, 2018, p. 246).

---

[24] "[...] o dicionário é uma obra de referência que consiste em uma descrição atomística do léxico e determinada, por sua vez, por quatro fatores variáveis: a) o número e a extensão de seus verbetes, b) a forma de estudá-los, c) a gestão que lhes é aplicada e, por fim, d) o suporte da referida descrição". (T.N.).

[25] "Um dicionário é uma ferramenta de utilidade, concebida para a consulta com o propósito genuíno de satisfazer necessidades pontuais de informação requeridas por determinados tipos de potenciais usuários em determinados tipos de contexto extralexicográfico, e que se destina a auxiliar os seus utilizadores através da disponibilização manual ou automática de acesso a dados preparados lexicograficamente, que podem ser usados diretamente pelos usuários para recuperar as informações necessárias que podem ser usadas posteriormente para resolver problemas específicos no contexto em que as necessidades ocorreram originalmente, ou por uma ferramenta digital para automatizar correções em textos ou traduções produzidas pelos usuários desta ferramenta". (T.N.).

A definição apresentada por esse autor partilha de um objetivo comum à Lexicografia que é o atendimento a uma dúvida específica do usuário, tendo em vista que uma obra lexicográfica é planejada e concebida para esse fim. Além disso, Tarp (2018) enfatiza que os usuários buscam o dicionário para tentar sanar um problema que surge de modo extralexicográfico e, desse modo, o dicionário é incumbido da missão de ser assertivo no que diz respeito a apresentar uma informação lexicográfica que resolva a dúvida que motivou a pesquisa. O autor também explica que a consulta lexicográfica pode ser realizada de maneira manual, a partir do uso de dicionários impressos ou de maneira eletrônica. Nesse particular, ferramentas digitais oferecem soluções automáticas que facilitam o trabalho de tradutores e estudantes de língua estrangeira.

Observamos, desse modo, que a perspectiva de Tarp (2018) não anula os pressupostos de Porto Dapena (2002), pois o objeto em foco é o mesmo, isto é, os dicionários. No entanto, o diferencial abordado pela perspectiva computacional da Lexicografia é o uso das tecnologias computacionais da atualidade para melhorar, significativamente, a elaboração de produtos inovadores.

Nesse sentido, a Lexicografia tem passado por um movimento de transição causado pela ação contínua e crescente da evolução dos recursos computacionais de última geração que Tarp (2019, p. 225) denomina de *tecnologias disruptivas*[26]. Esses recursos computacionais estão presentes em diversas áreas e representam a *Quarta Revolução Industrial*[27] que tem modificado profundamente a maneira de como os seres humanos produzem e consomem produtos em uma sociedade cada vez mais conectada e automatizada.

Destacamos, ainda, que as tecnologias disruptivas não só oferecem maneiras revolucionárias de desenvolvimento técnico-científico, como também convidam os pesquisadores a expandirem seu campo de visão profissional e a pensarem igualmente de maneira disruptiva. Isso significa que o movimento de transição que a Lexicografia vivencia pede um esforço coletivo para que os lexicógrafos comecem a pensar que uma obra lexicográfica, no âmbito da Quarta Revolução Industrial, deve ser planejada para funcionar, exclusivamente, em plataformas digitais. Para tanto, é preciso que os

[26] Termo utilizado para designar as inovações tecnológicas que têm revolucionado diversas áreas da sociedade. Por funcionarem a partir da conexão com a Internet, proporcionam uma gama de funcionalidades que podem ser executadas remotamente.

[27] Expressão que designa a aplicação das tecnologias disruptivas em diversos campos, sobretudo, na indústria, promovendo a automação de diversas atividades por meio de computadores, de algoritmos e da Internet.

pesquisadores avancem um passo em direção ao ponto de transição que demarca a evolução da Lexicografia Impressa para a Lexicografia Eletrônica e, posteriormente, deem um segundo passo, mais ousado, no sentido de planejar e produzir obras lexicográficas genuinamente digitais.

Destacamos, ainda, que para se tirar o máximo proveito das tecnologias disruptivas uma obra lexicográfica deve ser planejada e produzida com uma mente disruptiva, ou seja, a partir de uma metodologia de trabalho diferenciada e um pensamento desvinculado do conceito de dicionário impresso e tradicional. Esse posicionamento é importante para se desenvolver obras lexicográficas realmente inovadoras do ponto de vista tecnológico, buscando aprimorar a maneira de atender as necessidades de pesquisa de seus usuários, vislumbrando uma futura individualização do acesso[28] ao dado lexicográfico.

Porém, destacamos que o uso de tecnologias disruptivas na Lexicografia ainda é incipiente e o que ocorre atualmente, com mais frequência, é o desenvolvimento de versões eletrônicas de dicionários impressos. É o que ocorre com as obras de Ferreira (2010) e de Houaiss (2009) que, apesar de serem obras de referência na língua portuguesa do Brasil, foram concebidas a partir do formato impresso e receberam, posteriormente, uma adaptação ao meio digital.

Dessa forma, para exemplificar melhor o desalinhamento que dos dicionários impressos que receberam versões digitais têm em relação à verdadeira Lexicografia Eletrônica, apresentamos, na seção seguinte, uma breve análise do Houaiss (2009) tomando como critério de análise as funções de busca que essa versão disponibiliza ao usuário.

2.2. A versão eletrônica do Dicionário Houaiss da Língua Portuguesa (2009)

A consulta ao dicionário Houaiss (2009), para recuperação de exemplos de definições que se enquadram na tipologia proposta por Porto Dapena (2002, p. 266-296), apresentada na seção *1.6 Tipologia das definições lexicográficas* desta Tese, evidenciou algumas limitações das ferramentas de pesquisa desse dicionário que restringem as possibilidades de busca a conteúdos específicos do banco de dados,

[28] Funcionalidade estudada pela Lexicografia Eletrônica com o objetivo de aplicá-la em futuros dicionários on-line. Maiores informações estão apresentadas no item *2.6. Critérios de classificação de dicionários eletrônicos*, desta Tese.

principalmente quando se tratam de pesquisas avançadas, ou seja, necessidades pontuais de usuários especializados.

Dessa forma, apresentamos, nesta seção, uma breve análise das ferramentas de busca oferecidas pelo Houaiss (2009), na versão em CD-ROM[29], a partir de uma perspectiva funcional[30], isto é, tomando como critério a capacidade que as barras de pesquisa avançada têm para recuperar e apresentar dados ao usuário.

O dicionário Houaiss eletrônico (2009) oferece ao usuário as seguintes possibilidades de busca: i) consulta convencional em que é possível buscar por uma entrada a partir do índice alfabético ou escrevendo o verbete desejado em uma caixa de pesquisa; ii) consulta avançada que se divide em três modalidades de busca, a saber: a) pesquisa simples – em que a busca pode ser realizada a partir do início ou do término de uma entrada, sendo um tipo de consulta útil em estudos relacionados aos prefixos e sufixos da língua portuguesa; b) pesquisa combinada – que acrescenta à pesquisa simples a possibilidade de filtrar a busca por meio da classificação gramatical das entradas; c) pesquisa reversa – que permite buscar um item lexical simples que figure na definição dos verbetes.

Para tanto, a ferramenta de pesquisa reversa foi utilizada para identificar as características das definições do Houaiss (2009), com a finalidade de exemplificar a tipologia das definições lexicográficas de Porto Dapena (2002), como mencionado anteriormente. Essa tarefa, no entanto, foi mais morosa do que o imaginado devido a algumas limitações de ordem técnica da ferramenta.

Em primeiro lugar, identificamos que a pesquisa reversa só realiza a busca de unidades lexicais que possuam um mínimo de três caracteres[31]. Com essa limitação o usuário fica impedido de investigar o uso de conectores, de conjunções, de artigos, bem como não consegue buscar no corpo das definições o uso das desinências verbais acompanhas ou não de hífen como, por exemplo, *-ar, -er, -ir* ou *ar, er, ir*.

O segundo aspecto constatado é que, apesar de a ferramenta de pesquisa reversa retornar ao usuário itens lexicais a partir de três caracteres, algumas UL que se

[29] O Houaiss (2009) em formato eletrônico também pode ser acessado por assinantes da plataforma UOL. Em 2017 foi lançada uma versão corporativa desse produto para assinantes, disponível em: [https://www.houaiss.net/corporativo/apps/www2/v6-2/html/index.php]. Reiteramos, porém, que a análise desta seção se refere, exclusivamente, à versão em CD ROM, lançada em 2009 e apenas para computadores com sistema operacional Windows.

[30] Referente à Teoria das Funções Lexicográficas, apresentada na seção 2.8 desta Tese.

[31] O caractere é um termo utilizado na Informática que pode representar uma letra, número, sinal de pontuação ou símbolo.

enquadram nesse padrão não foram recuperadas pela ferramenta como, por exemplo, os itens lexicais *que* e *por*. Como não é possível realizar buscas de UL compostas, ou seja, separadas por um espaço em branco, não foi possível identificar acepções que fazem o uso, por exemplo, das conjunções *por que, para que, a fim de que, visto que*, além de outras combinações compostas que podem ser de interesse de um consulente especializado.

O terceiro ponto que observamos relaciona-se às remissivas que não aparecem nos verbetes em formato clicável, ou seja, o usuário que desejar consultar uma remissiva deverá digitá-la na barra de pesquisa principal do dicionário. Essa característica se configura como outro ponto negativo para a obra, pois os usuários de dicionários eletrônicos esperam que a navegação entre os conteúdos aconteça de maneira automática, isto é, por meio de links de acesso aos dados.

Em quarto lugar, destacamos que muitas abreviações foram preservadas nas definições dos verbetes, o que não faz muito sentido quando se trata de dicionários eletrônicos, pois a limitação de espaço é um problema dos dicionários impressos. Todavia, levando em consideração que o Houaiss eletrônico (2009) é uma obra impressa e que foi adaptada para o meio eletrônico, entendemos que a equipe responsável por essa publicação fez algumas escolhas no sentido de automatizar alguns procedimentos de busca ao *corpus* do dicionário e outros aspectos que poderiam ser melhorados, como os apresentados nesta seção, não foram considerados.

Por fim, destacamos que os dados lexicográficos que constituem o *corpus* dos dicionários impressos e os bancos de dados dos dicionários eletrônicos, se configuram em uma rica fonte de pesquisa que podem servir de base para a construção de produtos diversificados como, por exemplo, assistentes de escrita e de tradução, além de uma gama de ferramentas monofuncionais[32] que têm como objetivo suprir necessidades pontuais de pesquisa de um tipo específico de usuário.

No que diz respeito à elaboração do protótipo do *VoDiNorte*, as ferramentas de pesquisa que foram desenvolvidas têm como objetivo recuperar informações do banco de dados a partir de variados critérios. Desse modo, a ferramenta de busca avançada permite identificar no *corpus* da pesquisa unidades lexicais simples ou compostas e sem limite mínimo de caracteres. Isso significa, por exemplo, que é possível solicitar a

[32] Esse tema é trabalhado com mais profundidade na seção *2.7. Ferramentas monofuncionais e polifuncionais*, desta Tese.

visualização do caractere *à* ou da unidade lexical *chuva branca*, além da possibilidade de se configurar outros critérios de busca avançada que retornam ao usuário um conteúdo filtrado pelas variáveis dialetais sexo, idade e localidade, conforme detalhado no Capítulo 4.

Por meio do exposto, consideramos que o processo de transição que a Lexicografia está passando engloba uma busca constante por melhorias metodológicas que possam resultar em produtos inovadores. Dessa maneira, na seção seguinte, damos mais um passo em direção à Lexicografia Eletrônica apresentando as principais características das obras lexicográficas que são planejadas e desenvolvias no âmbito das tecnologias disruptivas e baseadas na Teoria Funcional da Lexicografia (TARP, 2008).

2.3. Lexicografia Eletrônica

A Lexicografia Eletrônica se desenvolve na via expressa onde corre a tecnologia. Isso significa que quanto mais desenvolvidos forem os recursos computacionais que podem ser aplicados à Lexicografia, melhores serão os produtos desenvolvidos por lexicógrafos e programadores. Assim, a Lexicografia Eletrônica exige uma equipe mais diversificada de especialistas em comparação aos projetos de dicionários impressos, pois, atualmente, o planejamento e a produção de dicionários eletrônicos exigem, além dos critérios advindos da Lexicografia, conhecimentos específicos das áreas relacionadas à Ciência da Computação como, por exemplo, banco de dados, linguagens de programação, web designer, entre outras especialidades.

Destacamos, ainda, que há uma pequena variação terminológica que circunda o termo *Lexicografia Eletrônica* e seu objeto de estudo, o *dicionário eletrônico* que merece ser pontuada.

2.4. Lexicografia Eletrônica ou Lexicografia Digital?

Em língua inglesa o item léxico *e-lexicography* é comumente encontrado em artigos e livros que abordam a Lexicografia Eletrônica. O fato é que o uso do prefixo *e-* está bem difundido na sociedade contemporânea nomeando produtos e/ou serviços como, por exemplo, *e-mail, e-book, e-social, e-tools, e-commerce* etc. Na contramão do uso generalizado desse prefixo, Lew; De Schryver (2014 p. 342-343) pontuam que o adjetivo *digital* é mais adequado para designar os dicionários eletrônicos assim como

ocorre com o termo Humanidades digitais[33]. No entanto, os autores também reconhecem que o uso do prefixo *e-* encontra-se bem difundido e promover uma mudança de terminologia dentro da comunidade lexicográfica não é tarefa fácil.

Sylviane Granger, na introdução do livro *Eletronic Lexicography*, publicado em 2012, defende o uso do termo *Lexicografia Eletrônica* e *dicionário eletrônico* e apresenta uma definição que justifica o sentido dessas nomenclaturas:

> In this volume 'electronic lexicography' is used as an umbrella term to refer to the design, use, and application of electronic dictionaries (EDs), which are in turn defined as primarily human-oriented collections of structured electronic data that give information about the form, meaning, and use of words in one or more languages and are stored in a range of devices (PC, Internet, mobile devices)[34] (GRANGER, 2012, p. 2).

Rodríguez Barcia (2016, p. 139), por sua vez, prefere o uso do termo *digital*, argumentando que esse item léxico abrange todos os tipos de dicionários disponíveis em meio eletrônico incluindo, por exemplo, o dicionário impresso que foi escaneado e disponibilizado na Internet. Desse modo, para a autora, "[…] un diccionario digital es un repertorio lexicográfico cuya consulta exige medios tecnológicos y procedimientos informáticos. Por otro lado, puede tratarse de obras digitalizadas a partir de sus originales físicos" (RODRÍGUEZ BARCIA, 2016, p. 140)[35].

Como é possível observar, os termos *Lexicografia Eletrônica* e *Lexicografia digital* buscam abranger as obras lexicográficas disponíveis ao usuário por meio do suporte eletrônico incluindo, inclusive, a digitalização em PDF de um dicionário por meio de um escâner ou pela câmera fotográfica do celular.

---

[33] Termo utilizado nas no âmbito das Ciências Humanas para pesquisas que têm incorporado em sua metodologia o uso de ferramentas digitais.

[34] "Nesse volume 'Lexicografia Eletrônica' é utilizado como um termo guarda-chuva para se referir ao design, uso e aplicação dos dicionários eletrônicos (DEs) que, por sua vez, são definidos como coleções de dados eletrônicos estruturados com foco nas necessidades dos seres humanos e que fornecem informações sobre a forma, o significado e o uso de palavras em um ou mais idiomas e são armazenados em uma variedade de dispositivos como, por exemplo, computadores pessoais, Internet, dispositivos móveis" (GRANGER, 2012, p. 2). (T.N.).

[35] "[...] um dicionário digital é um repertório lexicográfico cuja consulta exige meios tecnológicos e procedimentos informatizados. Por outro lado, pode se tratar de obras digitalizadas a partir de seus originais físicos" (RODRÍGUEZ BARCIA, 2016, p. 140). (T.N.).

Além disso, os termos *dicionário eletrônico* e *dicionário digital*, no âmbito da Lexicografia Eletrônica, podem assumir denominações como *recurso de referência* ou *ferramentas de consulta*:

> […] the purpose of a dictionary, at the highest level of abstraction, is to assist its users in satisfying punctual information needs. The users retrieve the needed information from the lexicographically prepared data contained in the dictionary. A dictionary can therefore be classified as a reference resource or consultation tool[36] (TARP, 2018, p. 245-246).

Na introdução do livro *e-Lexicography*, Fuertes-Olivera e Bergenholtz (2011) levantam uma discussão em torno do termo *dicionário* no sentido de que, do ponto de vista científico, se trata de uma nomenclatura ultrapassada e que não denota a realidade dos dicionários eletrônicos e, dessa forma, o termo *ferramenta de informação* seria o mais adequado, tendo em vista a capacidade que tais ferramentas têm, a partir das tecnologias disruptivas, de atender as necessidades de pesquisa de seus usuários:

> One of the solutions proposed in this book is to move beyond the term 'dictionary' and introduce the term 'information tool' as a kind of umbrella term with which researchers can design any tool, no matter what we call them, aiming to satisfy the needs users might have in the four use–situations described so far: communicative, cognitive, operative and interpretive[37] (FUERTES-OLIVERA; BERGENHOLTZ, 2011, p. 3).

Outro ponto a ser considerado nessa discussão terminológica é o fato de que a nomenclatura *dicionário* ganhou *status* comercial com o passar dos séculos e tem sido usado, desde então, para designar obras que organizam listas de palavras e seus respectivos sentidos em qualquer área do conhecimento. Nesse sentido, o número de publicações que levam em seu título o termo *dicionário* cresce exponencialmente, mas isso não significa que essas publicações sejam de dicionários linguísticos, confeccionados mediante critérios lexicográficos. Assim, dado o fato que os dicionários

[36] "[...] o propósito de um dicionário, no mais alto nível de abstração, é auxiliar seus usuários na satisfação de necessidades pontuais de informação. Os usuários recuperam as informações necessárias dos dados preparados lexicograficamente contidos no dicionário. Um dicionário pode, portanto, ser classificado como um recurso de referência ou ferramenta de consulta" (TARP, 2018, p. 245-246). (T.N.).

[37] "Uma das soluções propostas neste livro é ir além do termo 'dicionário' e introduzir o termo 'ferramenta de informação' como uma espécie de termo guarda-chuva com o qual os pesquisadores podem projetar qualquer ferramenta, não importa como a chamemos, visando satisfazer necessidades que os usuários podem ter nas quatro situações de uso descritas até agora: comunicativa, cognitiva, operativa e interpretativa" (FUERTES-OLIVERA; BERGENHOLTZ, 2011, p. 3). (T.N.).

podem tratar de temas variados, o uso do termo *ferramentas* que agregam *informações lexicográficas* torna-se mais adequado principalmente no contexto da Lexicografia Eletrônica (LEROYER, 2011, p. 124).

Seguindo a linha de que os dicionários eletrônicos funcionam como ferramentas informatizadas que oferecem ao usuário o acesso a dados tratados lexicograficamente, Tarp (2011, p. 69) questiona até que ponto o termo *dicionário* é relevante e se esta terminologia pode estar em vias de se tornar obsoleta na perspectiva científica, pois, embora o nome *dicionário* tenha força do ponto de vista comercial, na Lexicografia Eletrônica é mais coerente denominar estas obras com termos que as descrevem melhor como, por exemplo, *ferramenta de informação lexicográfica*, *ferramenta de consulta lexicográfica* ou *e-ferramenta lexicográfica*. Segundo esse autor, a tendência é que os projetos lexicográficos busquem utilizar ao máximo as novas tecnologias computacionais disponíveis para inovar a experiência do usuário na resolução de um problema. Além disso, ao fazer uso das tecnologias disruptivas a obra lexicográfica deixa de ser um instrumento de consulta a dados estáticos, como ocorre com os dicionários impressos, e passa a se caracterizar como uma *ferramenta de consulta lexicográfica.* Desse modo, a partir das considerações de Tarp (2011, p. 69), Leroyer (2011, p. 124) e Fuertes-Olivera; Bergenholtz (2011, p. 3) adotamos o termo *ferramenta de consulta lexicográfica* para nos referir ao protótipo do *VoDiNorte*, dado o contexto metodológico que baseia o seu desenvolvimento.

2.5. Características dos dicionários eletrônicos

Diante da variedade de obras lexicográficas que se apropriam do adjetivo eletrônico ou digital sem realmente fazer uso das tecnologias disruptivas para elaborar ferramentas inovadoras de consulta a dados lexicográficos, Granger (2012, p. 2) enumera seis facetas revolucionárias da Lexicografia Eletrônica que se configuram como itens de base a ser explorados pelos lexicógrafos da atualidade, sintetizadas no quadro a seguir:

Quadro 4: Características da Lexicografia Eletrônica.

Item a ser explorado	Descrição
1) Uso de *corpus* integrado	Fazer uso de quantos *corpora* forem necessários para o desenvolvimento de uma obra lexicográfica, incluindo bancos de dados disponíveis da Internet.
2) Quantidade e qualidade de dados	Aumentar a quantidade de dados oferecida ao usuário sem perder de vista a qualidade.
3) Eficiência e rapidez de acesso	Atender ao usuário com rapidez de acesso e ser assertivo ao exibir dados que resolvam, com eficiência, a dúvida que motivou a pesquisa.
4) Customização do dicionário	Desenvolver ferramentas capazes de identificar o tipo de usuário que está acessando a plataforma, a fim de selecionar com maior precisão os dados que serão exibidos.
5) Uso de ferramentas híbridas	Fornecer links externos que levem o usuário a ampliar o assunto a partir de outras ferramentas e/ou plataformas.
6) Colaboração do usuário	Oferecer a possibilidade de os usuários colaborarem por meio do envio de informações adicionais sobre determinado verbete.

Fonte: Granger (2012, p. 2-5).

As informações apresentadas no quadro 4 reforçam a afirmação de que os dicionários eletrônicos se desenvolvem no trilho das inovações tecnológicas, como mencionado no início da seção *2.3. Lexicografia Eletrônica.* No entanto, para que essas novidades possam, de fato, impactar a Lexicografia é preciso lançar mão desses recursos e aplicá-los ao labor lexicográfico.

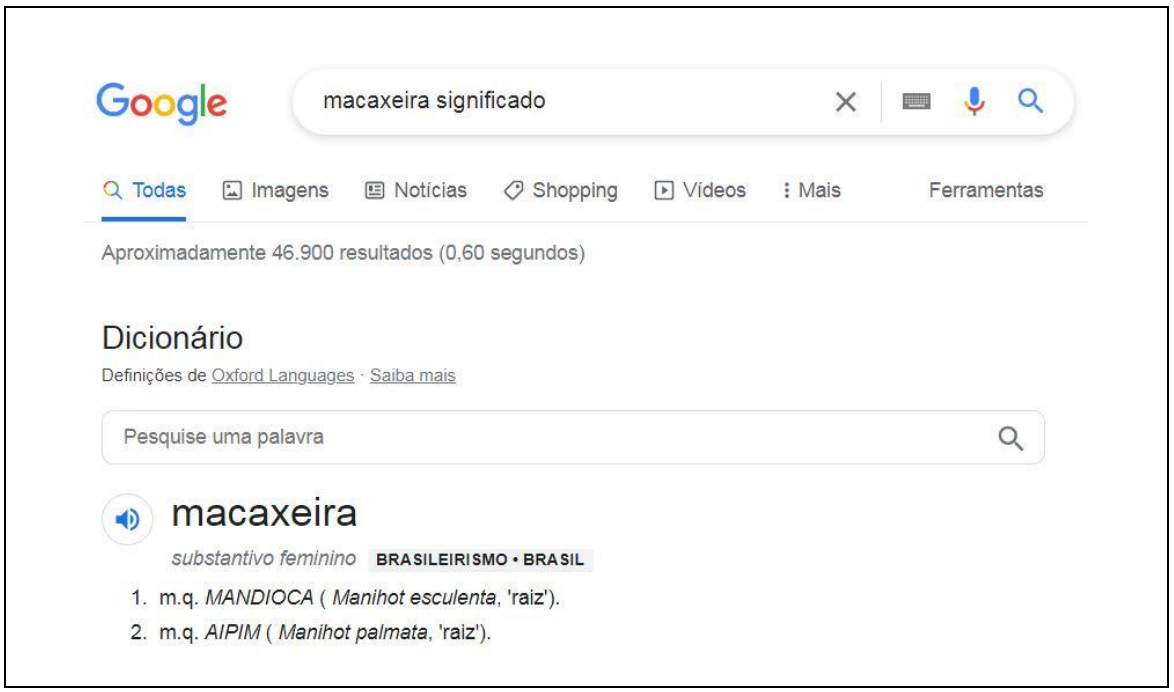
De maneira geral, as seis facetas da Lexicografia Eletrônica elencadas por Ganger (2012) podem ser encontradas em plataformas como o buscador do Google e a Wikipédia, que oferecem aos usuários informações em formato lexicográfico com o objetivo de atender com rapidez e eficiência (item 3 do quadro 4) seus potenciais usuários. Destacamos, ainda, que a Wikipédia e principalmente o motor de pesquisa do Google investem em tecnologia para cativar a preferência do usuário que passa a utilizar esses serviços com maior frequência, o que alimenta a liderança dessas empresas nesse modelo de negócio.

No intuito de aprofundar um pouco mais essa temática, vamos criar uma situação hipotética em que um estudante da educação básica tenha uma dúvida sobre o sentido da UL *macaxeira*. Assim, como de costume, esse aluno vai até o computador ou

o celular e digita diretamente no Google duas unidades léxicas: *macaxeira significado*[38]. Essa configuração de busca não é aleatória, pois durante o manuseio do buscador Google por esse jovem internauta a primeira sugestão que aparece na barra de pesquisa enquanto o usuário está digitando uma palavra é, com base no histórico de navegação, exatamente essa ordem: *macaxeira significado*. Com o passar do tempo e com a experiência desse usuário em utilizar o Google como uma fonte de pesquisa lexicográfica, o aluno acaba internalizando essa configuração de busca, pois percebe que digitar a palavra que se deseja saber o sentido, acrescida do termo *significado*, resulta no que ele está procurando de forma rápida e eficaz. Esse fato, por si só, já demostra o quanto os usuários estão habituados a recorrer ao Google e dão preferência ao que pode ser lido na primeira página exibida pelo motor de busca ao invés de recorrer aos resultados elencados ao acionar a barra de rolagem para baixo.

A figura a seguir mostra o resultado da busca das unidades lexicais *macaxeira significado* digitadas no motor de pesquisa Google:

Figura 21: Resultados para *macaxeira significado* no buscador Google.



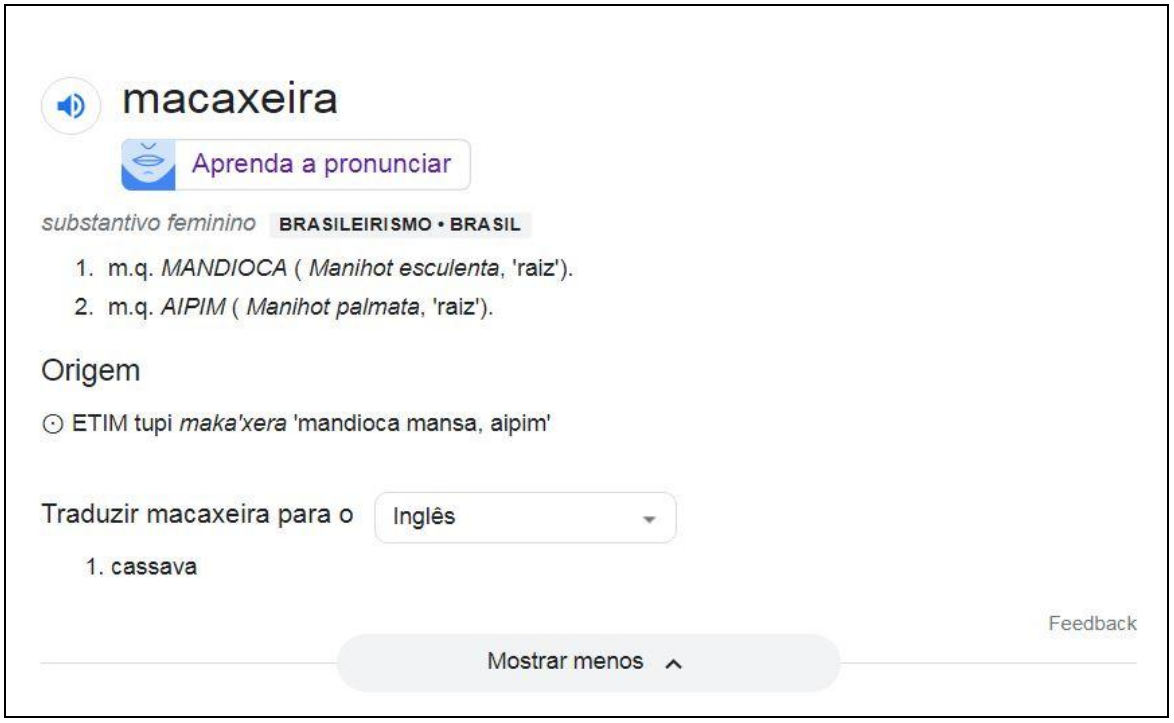
Fonte: Google.com. Acesso em: 15 nov. 2022.

[38] O termo *significado* é comumente utilizado como sinônimo de *sentido* como discutido por Medina Guerra (2003, p. 129-131) e apresentado no subitem *1.5 Definição lexicográfica*, deste trabalho.

Observamos a partir da figura 21 que o primeiro resultado da busca é exibido em formato de verbete lexicográfico, sem a necessidade do usuário clicar em nenhum dicionário on-line. Na verdade, uma barra de pesquisa intitulada *Dicionário*, ao centro da tela, indica que o Google abriu uma porta de comunicação com o banco de dados da *Oxford Languages*, que é a instituição responsável por fornecer os dados lexicográficos que são exibidos diretamente na interface do buscador Google. Desse modo, o usuário tem uma caixa de pesquisa ligada aos dados tratados lexicograficamente pela *Oxford Languages* o que significa que esses resultados possuem qualidade lexicográfica, já que são oriundos de uma editora reconhecida mundialmente pela produção de dicionários.

Destacamos, ainda, que no verbete ilustrado na figura 21, há informações que elucidam ao usuário a classe gramatical, de que se trata de um brasileirismo, além de duas definições que, a princípio, devem satisfazer a necessidade de pesquisa do um aluno da educação básica. Há também uma preocupação em dosar a quantidade de informação que é exibida na tela do usuário. Assim, o recurso utilizado para explorar essa funcionalidade é a inserção de abas que são expandidas e recolhidas ao clicar do mouse, aumentando ou diminuindo a quantidade de informação a ser visualizada, conforme é possível observar na figura a seguir:

Figura 22: Informações adicionais para *macaxeira significado* no buscador Google.



Fonte: Google.com. Acesso em: 15 nov. 2022.

O usuário pode acessar informações adicionais relacionadas à UL *macaxeira* clicando no ícone do alto falante que abrirá a ferramenta de áudio com a pronúncia que pode ser executada em duas velocidades. Além disso, uma integração com a ferramenta de tradução do Google permite traduzir a UL pesquisada em uma quantidade considerável de línguas. Esses recursos adicionais se caracterizam pela hibridização em que outras ferramentas são acessadas a fim de aumentar o leque de possibilidades para a referida pesquisa. Esse leque aumenta à medida que o usuário avança com a barra de rolagem para baixo, conforme é possível identificar na figura a seguir:

Figura 23: Outras possibilidades de acesso para *macaxeira significado* no buscador Google.

Fonte: Google.com. Acesso em: 15 nov. 2022.

Evidentemente, o motor de busca Google recupera uma quantidade muito grande de dados. Todavia, para que essas informações sejam relevantes ao usuário, o navegador de Internet organiza o conteúdo exibido na primeira página por seções distintas. Assim, na figura 23, observamos duas seções que agrupam dados de maneiras diferentes. A primeira é composta por abas intitulada *As pessoas também perguntam* que são alteradas dinamicamente conforme a navegação do usuário, ou seja, novas abas surgem

na tentativa de oferecer resultados semelhantes ao conteúdo que está sendo clicado, sendo um recurso interessante rumo à customização do acesso, pois se baseia nas escolhas do usuário para filtrar o que é exibido na tela. A segunda seção de dados é formada pela lista de links relacionados à busca do usuário.

Um ponto positivo na interface do Google com a *Oxford Language* é a ausência de anúncios, que deixa a navegação mais leve e simplifica o acesso à informação de forma eficiente. Isso é relevante, principalmente, quando o usuário é um aluno da educação básica que possui um celular com acesso à Internet limitado a uma quantidade de dados diários (Internet móvel pré-paga). Dessa maneira, a exibição de um verbete lexicográfico como primeira opção na pesquisa com uma interface limpa, sem propagandas e que dispensa a abertura de outros websites, beneficia o usuário com uma navegação mais leve, além de economizar o fluxo de dados que é um fator a ser considerado quando sem tem uma conexão limitada à Internet.

Desse modo, constatamos que o buscador *Google* tem desenvolvido uma estratégia para organizar o conteúdo de uma busca que se assemelha aos itens que caracterizam a Lexicografia Eletrônica apresentados por Granger (2012) e sintetizados no quadro 4 desta seção.

Levando em consideração que o navegador de Internet é o ponto de partida de qualquer internauta, ou seja, é uma ferramenta de uso obrigatório para se acessar a rede mundial de computadores, as empresas que desenvolvem esses motores de busca investem em tecnologia e estratégias de indexação e exibição de dados com a finalidade de melhor atender seus usuários e, consequentemente, buscam ganhar a preferência desses indivíduos que passarão a utilizar outros serviços dessas empresas.

Isso evidencia o quão valioso podem ser os dados organizados sob critérios da Lexicografia como, por exemplo, aqueles fornecidos pela *Oxford Languages* para alimentar a interface híbrida do buscador *Google* em determinadas pesquisas. Essa estratégia também mostra um novo modelo de negócio a ser explorado pela Lexicografia no intuito de monetizar o acesso a bancos de dados lexicográficos.

Outro ponto a ser destacado é uma parceria feita entre o *Google* e a *Wikipédia* no qual as buscas realizadas pelos usuários no motor de pesquisa *Google* exibem na primeira página do navegador, dados recuperados da *Wikipédia* e sumarizados em um quadro localizado à direita da tela do usuário. Desse modo, ao digitar na barra de pesquisa apenas a UL *macaxeira*, ao invés de *macaxeira significado*, a interface do

Google irá recuperar ao topo da página as informações do banco de dados da *Wikipédia* e não da *Oxford Languages*.

Essa parceria não foi uma eventualidade, pois a *Wikipédia* possui um repositório de dados extremamente extenso, em vários idiomas e que não para de crescer. Fundada em 2001, a organização sem fins lucrativos se caracteriza pela elaboração de conteúdos de forma colaborativa por meio da tecnologia *wiki*[39], que são compartilhados gratuitamente. Com o passar do tempo a *Wikipédia* desenvolveu projetos irmãos como, por exemplo, o *Wikcionário* lançado em 2004 que, diferentemente do caráter enciclopédico da *Wikipédia*, objetiva-se a desenvolver dicionários colaborativos em diversas línguas. No caso da língua portuguesa, no momento da escrita desta Tese, o dicionário colaborativo possui 271.560 entradas. No entanto, há *Wikcinários* construídos em alguns idiomas que chegam a um milhão de entradas como é o caso do alemão, chinês, francês, inglês, malgaxe e russo.

Em relação à questão da qualidade e da veracidade das informações veiculadas no portal da *Wikipédia*, dois pontos devem ser considerados: i) a ação de robôs que atuam contra a inserção de vandalismos, deletando esse tipo de conteúdo da plataforma; ii) a própria atuação da comunidade, formada por usuários que revisam os conteúdos voluntariamente a fim de melhorar a qualidade dos dados publicados mediante a indicações de fontes de pesquisa. Além disso, os colaboradores podem abrir seções de discussões sobre qualquer tema com o objetivo de questionar uma informação publicada no portal.

Destacamos, ainda, que os conteúdos publicados na *Wikipédia* ou no *Wikcionário* devem ser vistos como um conjunto de dados construídos colaborativamente por usuários múltiplos. Isso inclui entusiastas, especialistas e qualquer pessoa que queira criar ou editar uma informação. Assim, ao pesquisar um conteúdo *wiki* não se deve tomar aquilo como a voz de um especialista no assunto, mas como a voz de uma comunidade que trabalha para construir um sistema informatizado de conteúdo livre na Internet. Nesse caso, é importante questionar e consultar outras fontes e, no caso de encontrar informações imprecisas, notificar o fato na aba *Discussão* presente em cada entrada do portal ou até mesmo atualizar o conteúdo na aba *Editar*.

[39] Ferramentas computacionais que permitem a criação e a edição de conteúdos de forma colaborativa por meio de uma página na Internet.

Se, por um lado, é possível encontrar conteúdos na *Wikipédia* de origem duvidosa, por outro lado, é possível reunir dados de uma grande gama de temas e de uma variedade de línguas e localidades que se configura num trabalho bastante complexo de se executar sem o uso da tecnologia *wiki.* Desse modo, acreditamos que os benefícios superam em larga escala os pontos negativos.

Além do *Wikcionário*, outros 14 projetos em formato *wiki* foram criados e estão em desenvolvimento no portal da *Wikipédia,* a saber: *Commons* – repositório de mídia livre; *Wikinotícias* – jornalismo colaborativo; *Wikispecies* – diretório de espécies livre; *MediaWiki* – software wiki; *Incubator* – incubadora de projetos; *Wikidata* – base de conhecimentos livre; *Wikiquote* – coletânea de citações livre; *WikiTech* – projetos técnicos; *MetaWiki* – projeto de coordenação; *Wikilivros* – livros e manuais livres; *Wikisource* – biblioteca livre; *Wikivoyage* – guia livre de viagem; *Wikispore* – projeto experimentais.

Com base no exposto, é possível observar que o conceito *wiki* tem se difundido entre usuários de todo o mundo, formando um verdadeiro exército de colaboradores que criam, editam e policiam o conteúdo armazenado e disponibilizado, gratuitamente e livre de propagandas, nos projetos encabeçados pela *Wikipédia*.

Outro exemplo da colaboração ativa do usuário a dados lexicográficos é o *Dicionário inFormal*, que possui a proposta de publicar verbetes redigidos pelos usuários. Nesta plataforma, o internauta envia uma definição que, inclusive, pode conter palavras de baixo calão. Dessa forma, essa ferramenta se configura num acervo sobre a língua portuguesa, registrado pelos próprios falantes.

Um ponto que merece destaque no uso de plataformas que publicam conteúdo lexicográfico a partir da colaboração dos usuários é o registro de unidades lexicais regionais que, em alguns casos, não ocorrem em outras localidades. Isso foi atestado quando nos deparamos com a unidade lexical *tarisca* durante as transcrições dos áudios do Projeto ALiB, que não está dicionarizada nas principais obras lexicográficas da língua portuguesa. Após uma busca mais aprofundada na Internet, o item *tarisca* foi encontrado apenas no *Dicionário informal*, que a definia como ripas de madeira, como aquelas que existem no estrado da cama, por um usuário do estado da Paraíba. Essa definição é verdadeira, porém, nas gravações do ALiB, o item lexical *tarisca* é utilizado para nomear as lâminas do moinho que trituram a mandioca durante o processo de fabricação de farinha de mandioca.

Em síntese, a colaboração do usuário deve ser vista com bons olhos, já que é possível delimitar critérios para atestar a qualidade e a veracidade do conteúdo enviado por um internauta. Desse modo, no âmbito da Lexicografia Eletrônica há possibilidades do usuário ser mais do que um consulente e atuar como coadjuvante no intuito de enriquecer os dados tratados lexicograficamente em um dicionário on-line.

Além da colaboração do usuário, outro ponto importante que caracteriza a Lexicografia Eletrônica é a customização do acesso, conforme apontado por Granger (2012, p. 5). Nesse sentido, é possível utilizar mecanismos que relacionam o histórico de navegação para distinguir tipos de usuário. Basicamente, a plataforma guarda o acesso, itens pesquisados e configurações predefinidas para nortear o acesso e a visualização de dados em uma segunda visita.

Uma segunda alternativa para adequar o acesso aos dados em um dicionário on-line é definir, na primeira página do website, os tipos de usuários que a ferramenta de consulta lexicográfica atende. Assim, ao clicar em um tipo específico de usuário o internauta saberá que está acessando dados que foram pensados exclusivamente para esse tipo de consulente.

Esse foi a caminho traçado na aplicação web do protótipo do *VoDiNorte* no qual o consulente, antes de ter acesso às ferramentas de busca, deve clicar em um dos três tipos de usuários, a saber: i) usuário comum – leigo e estudantes do Ensino Fundamental; ii) usuário intermediário – leigo, estudantes do Ensino Médio e estudantes do Ensino Superior; iii) usuário avançado – pesquisadores em geral e professores universitários. Essa configuração permite oferecer dados e ferramentas de busca voltadas para um público específico de modo a não ser complexo em demasia para o usuário comum e nem superficial para os usuários especializados. Desse modo, conteúdo, forma e ferramentas foram planejados tendo em vista estes três tipos de usuários, que serão descritos com mais detalhes no capítulo destinado à metodologia.

Destacamos, ainda, que as características dos dicionários eletrônicos citadas por Granger (2012, p. 2-5) pavimentam a estrada que leva a um conjunto de critérios estabelecidos por Fuertes-Olivera; Tarp (2014, p. 13) para classificar as obras lexicográficas no âmbito da Lexicografia Eletrônica.

2.6. Critérios de classificação de dicionários eletrônicos

Como mencionado anteriormente, no âmbito da Lexicografia Eletrônica, se faz necessário explorar as tecnologias disruptivas no intuito de oferecer aos usuários ferramentas de pesquisa lexicográfica inovadoras. Desse modo, Lexicografia e tecnologia computacional caminham juntas e, partindo dessa interface, Fuertes-Olivera; Tarp (2014, p. 13) apresentam uma classificação dos dicionários eletrônicos a partir do critério tecnológico.

Essa classificação dos dicionários eletrônicos foi inspirada em Henri Ford, que revolucionou a indústria automobilística, há cerca de 100 anos, com o veículo Ford Modelo T. Essa invenção representa um salto tecnológico que vai do cavalo para o automóvel. Há uma anedota que descreve o espírito inovador de Henry Ford: perguntaram ao visionário inventor se ele tinha indagado às pessoas a respeito do que elas queriam antes de apresentar o Ford modelo T. A resposta de Henry Ford foi a seguinte: *Se eu tivesse perguntado às pessoas o que elas queriam, responderiam que desejariam cavalos mais rápidos e não um automóvel*. Logo, Ford foi um inventor audacioso que soube satisfazer as necessidades das pessoas de uma maneira completamente nova (FUERTES-OLIVERA; TARP, 2014, p. 13).

Assim, analogamente às contribuições de Henri Ford no campo do automobilismo, Fuertes-Olivera; Tarp (2014, p 13) propõem uma classificação das obras no âmbito da Lexicografia Eletrônica que abrange os dicionários existentes e, ainda, uma categoria de dicionários que não existem atualmente, ou seja, uma classificação de obras lexicográficas que serão desenvolvidas no futuro, a saber: 1) *Copycats*; 2) *Faster Horses*; 3) *Stray Bullets*; 4) *Model T Fords*; 5) *Rolls Royces* que foram sintetizadas no quadro a seguir:

Quadro 5: Tipologia dos dicionários eletrônicos.

Tipologia	Descrição
Copycats	Dicionários impressos que foram escaneados e disponibilizados na Internet.
Faster Horses	Dicionários impressos que receberam uma versão eletrônica.
Stray Bullets	Dicionários eletrônicos que carecem de critérios lexicográficos, sobretudo, da Lexicografia Eletrônica.
Model T Fords	Dicionários eletrônicos desenvolvidos a partir dos pressupostos da Lexicografia Eletrônica.
Rolls Royces	Dicionários que ainda não foram produzidos e que oferecerão ao usuário um acesso personalizado aos dados.

Fonte: Fuertes-Olivera; Tarp (2014, p. 13-16).

Com base na classificação apresentada no Quadro 5, constatamos que é o tipo de tecnologia empregada na elaboração dessas obras que determina sua classificação de modo que os primeiros fazem um uso elementar dos recursos informáticos disponíveis atualmente, ao passo que os últimos recorrem às tecnologias disruptivas de maneira mais produtiva.

Dessa forma, os dicionários do tipo C*opycats* configuram-se numa classe, como o próprio nome sugere, das cópias de dicionários impressos que têm se multiplicado ao longo dos anos a partir do desenvolvimento da tecnologia *OCR* (*Optical Character Recognition*), que permite a um escâner copiar um livro impresso e armazená-lo em formato eletrônico, normalmente com extensão *.pdf*. Além disso, o barateamento desse tipo de equipamento, bem como a sua comercialização de forma integrada às impressoras de uso doméstico, facilitam o processo de digitalização de uma obra impressa e seu compartilhamento por meio da Internet.

Destacamos, entretanto, que as cópias de obras impressas sem a autorização dos autores violam os direitos autorais. Mas, se a obra em questão for de domínio público, não há problemas em sua reprodução por meios eletrônicos. No campo da

Lexicografia, por exemplo, a digitalização de dicionários antigos, que são encontrados apenas em algumas bibliotecas ao redor do mundo e que apresentam um valor histórico, significa ganho de tempo e redução de custos, por exemplo, para pesquisadores que têm esses dicionários como *corpus* de estudo (FUERTES-OLIVERA; TARP, 2014, p. 14). Dessa forma, a digitalização legal de obras raras e de acesso restrito é uma prática que contribui para o desenvolvimento científico.

Os dicionários do tipo *Faster Horses*, por sua vez, representam uma categoria em que se enquadram a maioria dos dicionários que levam em seu nome o termo *eletrônico*. Vale ressaltar que a nomenclatura *Faster Horses* remete ao criador do *Ford modelo T*, Henri Ford, que revolucionou os meios de transportes de carga e de pessoas de sua época ao produzir, em escala industrial, um veículo inovador a partir da tecnologia que ele dispunha na época ao invés de oferecer aos consumidores apenas um *cavalo mais rápido*.

Dessa maneira, um dicionário do tipo *Faster Horse* é planejado nos moldes da Lexicografia Impressa, publicado em papel e, posteriormente, adaptado para uma versão eletrônica como, por exemplo, é o caso dos dicionários Houaiss (2009) e Ferreira (2010), entre outros. Assim, os denominados *cavalos mais rápidos* oferecerem ferramentas de busca que permitem o acesso aos verbetes de forma eletrônica, isto é, rápida. Vale destacar que, além da busca por ordem alfabética e da busca por um verbete específico, alguns dicionários oferecem a pesquisa das entradas a partir da classe gramatical ou a partir de um item léxico[40] presente no texto da definição. Essas possibilidades conferem um acesso diferenciado, embora limitado, aos dados lexicográficos que, normalmente, são de interesse de usuários mais especializados.

Os dicionários do tipo *Stray Bullets*, ao seu turno, representam um número pequeno de obras lexicográficas dedicadas a incorporar o uso das novas tecnologias no processo de elaboração de dicionários. Todavia, segundo Fuertes-Olivera e Tarp (2014, p. 15) esses trabalhos estão caminhando na direção errada, pois, embora façam o uso de ferramentas computacionais, privilegiam o campo da Informática deixando de lado os critérios lexicográficos que visam a atender as necessidades dos consulentes. Ainda de acordo como os pesquisadores, outros exemplos dos

[40] No Houaiss (2009) e no Ferreira (2010) a busca avançada no texto da definição dos verbetes se limita a uma única unidade léxica, não sendo possível a pesquisa de unidades lexicais complexas.

dicionários do tipo *Stray Bullets* são aqueles que priorizam a quantidade em relação à qualidade. Desse modo, essas obras se caracterizam por oferecer opções que podem ampliar a quantidade de dados exibida para o usuário, conferindo um salto quantitativo, ao passo que deixa a desejar no quesito qualitativo.

Para tanto, a Lexicografia Eletrônica não deve focar apenas na apresentação de dados com finalidades lúdicas, pois o uso de arquivos multimídia em dicionários deve ter um propósito maior de que o simples entreter, ou seja, é preciso que essas obras atendam as necessidades de pesquisa de seus usuários de maneira inovadora:

> In this respect, it is necessary – as happened in the case of Henry Ford – to leave old habits behind and make full use of the available technology in order to invent new advanced solutions to old problems, in other words, to satisfy people's lexicographical needs in a completely new way (FUERTES-OLIVERA; TARP, 2014, p. 16) [41].

É nesse contexto de inovação lexicográfica que os dicionários do tipo *Model T Fords* e *Rolls Royces* surgem e se configuram como soluções lexicográficas que visam a atender as necessidades dos usuários de maneira revolucionária.

Desse modo, as obras lexicográficas do tipo *Model T Fords* têm como objetivo principal oferecer uma experiência de consulta com dados dinâmicos, isto é, as informações exibidas pelo dicionário podem variar a partir do tipo de usuário que está acessando a obra. Isso é possível mediante uma configuração prévia na página principal do dicionário on-line, no qual o consulente seleciona o tipo de usuário em uma lista de links que podem, por exemplo, diferenciar um internauta comum de outros que podem ser intermediários e especializados.

Em síntese, os dados dinâmicos são aqueles que podem mudar de acordo com as especificações configuradas antes do usuário acessar o dicionário. Por sua vez, os dados estáticos são aqueles que sempre permanecerão os mesmos, consulta após consulta, pois a obra não está equipada para oferecer resultados dinâmicos. Além do mais, links podem ser acrescentados a fim de oferecer informações complementares aos usuários, como fazem os dicionários do tipo *Faster horses* e *Stray Bullets*.

---

[41] "Nesse sentido, é necessário – como aconteceu no caso de Henry Ford – deixar para trás velhos hábitos e fazer pleno uso da tecnologia disponível para inventar novas soluções avançadas para velhos problemas, ou seja, para satisfazer as necessidades lexicográficas das pessoas de uma maneira completamente nova" (FUERTES-OLIVERA; TARP, 2014, p. 16) (T.N.).

De acordo com Fuertes-Olivera e Tarp (2014, p. 16), poucos dicionários especializados se encaixam na classificação *Model T ford*. Entre esses poucos, os autores citam o *Diccionarios de Contabilidad*, produzido em parceria entre pesquisadores do *International Centre for Lexicography* da Universidade de Valladolid, Espanha, e o *Centre for Lexicography*, da Universidade de Aarhus, Dinamarca. Vale destacar que essas afirmações datam de 2014 e que esse não deve ser o panorama atual dos dicionários que se enquadram na tipologia *Model T Fords.*

Os dicionários do tipo *Rolls Royces*, como mencionado anteriormente, ainda não existem e, portanto, representam uma categoria vazia que será preenchida num futuro próximo.

Essas obras lexicográficas serão capazes de atender as necessidades dos usuários de maneira customizada, ou seja, os acessos ao website serão personalizados. Vale destacar que esse tipo de tecnologia já existe e está presente nos algoritmos que oferecem sugestões de conteúdos no *YouTube*, na *Netflix* e nas redes sociais em geral. Outro exemplo de aplicação desenvolvida para oferecer serviços personalizados aos usuários é a ferramenta *Discover* que carrega notícias e artigos variados quando uma nova aba do *Google Crome* é aberta. Os resultados são criados com base no histórico de navegação da Internet e nos tipos de aplicativos utilizados. O usuário também pode inserir manualmente suas preferências para que o *Discover* otimize, ainda mais, a filtragem de informações sugeridas na tela. No entanto, essas tecnologias precisam ser testadas e adaptadas para o meio lexicográfico e é nesse caminho que o futuro da Lexicografia Eletrônica tende a caminhar.

Outro aspecto a ser mencionado sobre os dicionários do tipo *Rolls Royces* é que essas obras poderão percorrer páginas na Internet e acessar variados bancos de dados, com o objetivo de minerar informações relevantes que possam ser processadas em soluções dinâmicas ao usuário. Nesse sentido, um trabalho em conjunto com o campo da Inteligência Artificial pode auxiliar na escrita automática de verbetes com base em grandes quantidades de dados extraídos da Internet.

Em suma, a proposta dos tipos de dicionários eletrônicos apresentada por Fuertes-Olivera e Tarp (2014, p. 13) não só amplia as possibilidades de pesquisa de um consulente, como também convida linguistas e lexicógrafos a se engajarem em projetos que tenham a participação expressiva de programadores e web designers, pois este é o futuro do labor lexicográfico.

Destacamos, ainda, que a classificação tipológica de Fuertes-Olivera e Tarp

(2014, p. 13) tem um caráter didático, pois auxilia os pesquisadores a compreenderem os rumos que a Lexicografia Eletrônica está tomando, além de nortear os projetos que almejam explorar as tecnologias disruptivas no planejamento e na elaboração de dicionários eletrônicos como é o caso do protótipo do *VoDiNorte*. Dessa maneira, ao aplicar essa classificação ao produto desta Tese, nos deparamos com uma mescla que envolve características de duas tipologias distintas, a saber: i) *Faster Horses* – caracterizada pela apresentação estática de dados lexicográficos ao usuário comum e intermediário; ii) *Model T Fords* – representada pela apresentação de dados dinâmicos ao usuário avançado, exibidos mediante a configuração de filtros na ferramenta de pesquisa lexicográfica.

Essa dupla classificação também apresenta duas justificativas: a primeira relacionada com o pensamento focado na Lexicografa Impressa que nos impede de vislumbrar as possibilidades que as tecnologias disruptivas têm a oferecer no âmbito da Lexicografia Eletrônica e, a segunda, diz respeito a uma organização dos dados lexicográficos pensando nas necessidades de três tipos distintos de usuários, isto é, o comum, o intermediário e o avançado. Assim, o protótipo do *VoDiNorte* contará com ferramentas monofuncionais que atenderão aos usuários comuns e uma ferramenta polifuncional que poderá ser utilizada por consulentes intermediários. Além disso, o usuário avançado conta com uma ferramenta de busca sofisticada que pode recuperar dados dinâmicos e exibi-los na tela de acordo com a configuração de filtros. Como tais funcionalidades serão detalhadas no capítulo 4 apresentamos, na seção seguinte, a diferença entre ferramentas monofuncionais e polifuncionais.

2.7. Ferramentas monofuncionais e polifuncionais

Os dicionários impressos, em sua maioria, são ferramentas polifuncionais, pois têm o objetivo de reunir uma gama de informações em seus verbetes a fim de atender a variadas situações de pesquisa. Dessa forma, esse modelo de estruturar os dados lexicográficos exige que o consulente saiba identificar os tipos de informações contidas em cada entrada. Nesse sentido, as instruções sobre como utilizar o dicionário são importantes, pois somente com essas orientações é que o usuário poderá compreender a estrutura de cada verbete e, assim, fazer uso produtivo do dicionário.

Porém, quando se trata de usuários leigos, o uso de ferramentas polifuncionais pode oferecer um obstáculo no que diz respeito ao atendimento de necessidades

específicas. Imaginemos, a título de exemplo, um aluno da educação básica que tenha uma dúvida quanto ao uso do acento grave e precisa pesquisar sobre a regência de um verbo. Essa informação poderá ser de difícil visualização em um dicionário polifuncional e de fácil acesso em um dicionário de regência verbal. O mesmo princípio poderá ser aplicado em outras necessidades específicas do usuário como, por exemplo, na pesquisa por sinônimos, por antônimos, por homônimos entre outras possibilidades. Nesses casos, ferramentas monofuncionais, ou seja, que oferecem uma única funcionalidade são mais assertivas no que diz respeito ao atendimento e à satisfação do usuário.

Tendo em vista que os dicionários são concebidos como ferramentas e que, por sua vez, toda ferramenta é construída para auxiliar pessoas na execução de diversificadas tarefas se faz necessário, também, compreender que para cada tipo de atividade há um tipo de ferramenta apropriada. Desse modo, para se cortar uma grande árvore é preciso utilizar uma serra elétrica. Embora não seja o ideal, essa mesma ferramenta poderia ser utilizada para cortar uma viga de madeira. No entanto, esse tipo de serra é inadequada para cortar uma folha fina de madeira em três pedaços iguais. O mesmo ocorre com os dicionários polifuncionais, que são construídos similarmente aos canivetes suíços, ou seja, um artefato que pode ser utilizado em várias situações. Todavia, um dicionário monofuncional poderá atender a necessidade de consulta de um usuário de maneira rápida e eficiente (BERGENHOLTZ; BERGENHOLTZ, 2011, p. 187).

Vele destacar que no âmbito da Lexicografia Eletrônica o desenvolvimento de obras lexicográficas polifuncionais tem uma tendência de ser descontinuada, especialmente no que diz respeito ao atendimento do usuário comum. Desse modo, uma alternativa é o investimento na elaboração de dicionários monofuncionais ou até mesmo, a depender da situação, é válido considerar transformar um dicionário polifuncional em várias ferramentas de consulta on-line monofuncionais, ou seja, trabalhar para segmentar os dados lexicográficos e organizá-los por categoria, de modo a facilitar o acesso e a compreensão dessas informações pelos usuários que são plurais em suas necessidades e prezam pela clareza e pela objetividade. Nesse sentido, a Teoria Funcional da Lexicografia (TARP, 2008), que será discutida na seção seguinte, contribui para nortear o labor lexicográfico voltado para os dicionários eletrônicos, prezando sempre pela satisfação das necessidades do usuário.

2.8. Teoria Funcional da Lexicografia

A Teoria Funcional da Lexicografia é resultado de muitos anos de pesquisa. Em 1992, Sven Tarp faz a primeira tentativa de formular essa teoria, em sua tese de doutoramento, partindo do pressuposto de Wiegand (sem data) de que os dicionários são objetos de uso com propósito genuíno. Nesse contexto, surge a necessidade de se discutir os vários tipos de usuários que têm necessidades distintas de consultas lexicográficas (TARP, 2008, p. 33-34) que, por sua vez, precisam ser atendidas de maneira satisfatória. Esse é o eixo central que Tarp busca discutir em sua teoria, que passou por três reformulações até ser publicada, em 2008, no livro *Lexicography in the Borderland Between Knowledge and Non-Knowledge*[42].

Os pilares dessa teoria podem ser sumarizados da seguinte forma:

> Lexicography is a separate science whose object of study is dictionaries and their production and use. Consequently, there is a need to develop a general theory for this object of study. The theory of lexicographical functions is just a such a theory and is based on the idea that dictionaries are objects of use which are produced or should be produced to satisfy specific types of social need. These needs are not abstract – they are lined to specific types of users in specific types of social situation. Attempts are made available in specific types of dictionaries[43] (TARP, 2008, p. 43).

A partir desses pressupostos é possível identificar a preocupação do autor em elaborar uma teoria que auxilie linguistas e lexicógrafos na atividade prática de elaboração de dicionários, sem perder o foco no usuário.

Para alcançar esses objetivos, a Teoria Funcional da Lexicografia sistematizou possíveis demandas de pesquisa em quatro grupos, denominadas de situações lexicográficas, a saber:

---

[42] "Lexicografia na Fronteira do Conhecimento e Não-Conhecimento" (T.N.).

[43] "Lexicografia é uma ciência autônoma cujo objeto de estudo é o dicionário abrangendo sua produção e uso. Consequentemente, existe a necessidade de desenvolver uma teoria geral para esse objeto de estudo. A Teoria Funcional da Lexicografia é exatamente uma dessas teorias e se baseia na ideia de que os dicionários são objetos de uso que são produzidos ou deveriam ser produzidos para satisfazer tipos específicos de necessidades sociais. Essas necessidades não são abstratas – elas se alinham a tipos específicos de usuários em tipos específicos de situações sociais. As tentativas são feitas por meio de tipos específicos de dicionários" (TARP, 2008, p. 43) (T.N.).

> 1. Situaciones comunicativas donde puede presentarse la necesidad de resolver un problema de comunicación. Estas situaciones son las más estudiadas por la lexicografía y pueden subdividirse en producción, recepción, traducción y revisión de textos.
> 2. Situaciones cognitivas donde puede presentarse la necesidad de obtener conocimientos sobre algún tema o disciplina, p.ej. la economía, el comercio o la teoría lingüística. También pueden subdividirse en varias situaciones.
> 3. Situaciones operativas donde puede presentarse la necesidad de tener instrucciones para realizar una acción física, cultural o mental.
> 4. Situaciones interpretativas donde puede presentarse la necesidad de interpretar y comprender un signo, señal, símbolo o sonido que no es lingüístico[44] (TARP, 2015, p. 36).

Desse modo, a elaboração de dicionários deve tomar como referência essas situações lexicográficas com o objetivo de promover o desenvolvimento de ferramentas que atendam a funções específicas de consulta, ou seja, obras lexicográficas com funções comunicativas, cognitivas, operativas e interpretativas.

Em síntese, o conceito de função lexicográfica, bem como sua aplicação de modo prático na Lexicografia Eletrônica é descrita da seguinte forma:

> […] 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[45] (TARP, 2015, p. 36).

Partindo das contribuições de Tarp (2008, 2015) é possível afirmar que linguistas, lexicógrafos e programadores têm em mãos um norte teórico que fundamenta

[44] "1. Situações comunicativas em que pode surgir a necessidade de resolver um problema de comunicação. Essas situações são as mais estudadas pela lexicografia e podem ser subdivididas em produção, recepção, tradução e revisão de textos.
2. Situações cognitivas em que pode haver necessidade de obter conhecimento sobre algum assunto ou disciplina, por exemplo, economia, negócios ou teoria linguística. Também podem ser subdivididos em várias situações.
3. Situações operativas em que possa surgir a necessidade de instruções para realizar uma ação física, cultural ou mental.
4. Situações interpretativas em que possa surgir a necessidade de interpretar e compreender um signo, sinal, símbolo ou som que não seja linguístico". (TARP, 2015, p. 36) (T.N.).

[45] "[...] uma função lexicográfica pode ser definida como a assistência fornecida por uma obra lexicográfica para satisfazer os tipos específicos de necessidades pontuais de informação que um tipo específico de usuário potencial pode ter em um tipo específico de situação extralexicográfica. A referida assistência é realizada por meio de dados lexicográficos cuidadosamente elaborados e disponibilizados para consulta" (TARP, 2015, p. 36) (T.N.).

o labor lexicográfico, convidando esses profissionais a desenvolverem ferramentas de consulta lexicográfica alinhadas ao rigor metodológico da Lexicografia, além de explorar ao máximo as tecnologias disruptivas com a intenção de atender, satisfatoriamente, a multiplicidade de tipos de pesquisas de usuários diversificados. Destacamos, ainda, que esses esforços podem contribuir para um aumento na elaboração de obras do tipo *Model T Fords* e uma redução de compilações lexicográficas do tipo *Faster Horses*.

Na seção seguinte discutimos as áreas que se relacionam, de maneira bastante harmoniosa, com a Lexicografia e que também fazem parte do embasamento teórico desta Tese.

CAPÍTULO 3 – INTERFACES DA LEXICOGRAFIA

> Ao refletir sobre o aporte teórico-metodológico que embasa esta Tese, percebemos que a intersecção da Lexicografia com outras áreas do conhecimento constrói os alicerces desta pesquisa.

Em uma análise metafórica, esses alicerces podem ser comparados com a estrutura de uma árvore. Desse modo, a Lexicografia pode ser considerada como o tronco da árvore, pois dá o suporte físico que viabiliza o desenvolvimento da planta. A Linguística Computacional, por sua vez, representa os galhos. As suas bifurcações crescem e em suas extremidades surgem importantes folhas e, ao seu tempo, frutos que podem ser identificados pelos verbetes e pelos resultados gerados na ferramenta de pesquisa lexicográfica. No entanto, para que esse fruto venha a nascer, é preciso que a árvore esteja devidamente nutrida. Para tanto, a Dialetologia representa o emaranhado de raízes e os dados dialetais coletados e, posteriormente, mapeados pela Geolinguística estão presentes no solo, que é tão vasto quanto o mundo. A Linguística de *Corpus* representa o fluxo desses dados que nutrem a árvore e percorrem seu interior até chegar às extremidades, onde se encontram os frutos. Cada parte dessa estrutura arbórea é essencial e, mesmo apresentando características diferentes, são igualmente valiosas para a manutenção e frutificação da planta.

Desse modo, apresentamos nas seções deste capítulo cada uma das áreas que se correlacionam com a Lexicografia, no âmbito desta Tese, com a finalidade de compreender o estabelecimento de tais interfaces.

3.1 Dialetologia

Antes de focalizarmos essa área dos estudos linguísticos se faz necessário esclarecer que, no âmbito das pesquisas dialetais, há uma tendência, de alguns estudiosos, em tomar a Geolinguística como sinônimo de Dialetologia (CHAMBERS; TRUDGILL, 1994, p. 37). Nesse sentido, é importante destacar que a Geolinguística é o método da Dialetologia, ou seja, um conjunto de práticas metodológicas estabelecidas criteriosamente para a coleta de dados *in loco* que são utilizados, posteriormente, para estudar os dialetos de uma língua partindo de uma perspectiva espacial, pois:

> O espaço geográfico evidencia a particularidade de cada terra, exibindo a variedade que a língua assume de uma região para a outra, como forma de responder à diversidade cultural, à natureza da formação demográfica da área, à própria base linguística preexistente e à interferência de outras línguas que se tenham feito presentes naquele espaço no curso da história (CARDOSO, 2010, p. 15).

Vale destacar que o senso comum, frequentemente, define dialeto como toda forma da língua que difere da norma padrão. Assim, o falar de prestígio[46] é chamado de *estandar* e todas as demais variedades, de *subestandar*, que é associado ao falar rústico, campestre, da classe trabalhadora (CHAMBERS; TRUDGILL, 1994, p. 19).

O fato é que a maneira de falar de uma pessoa pode revelar informações que podem ser interpretadas de maneira preconceituosa. Dessa forma, o problema reside no olhar das variedades *subestardar* focado apenas no que se considera *certo* ou *errado* na língua, ao invés de observar o *adequado* e o *inadequado* na comunicação oral.

Superando essa visão preconceituosa em relação às variedades linguísticas, é preciso pontuar que, na realidade, cada indivíduo carrega consigo pelo menos um dialeto de uma língua. Desse modo, não há espaço para se classificar um dialeto como superior a outro (CHAMBERS; TRUDGILL, 1994, p. 19).

De fato os dialetos são formas muito ricas e peculiares da expressão oral que evidenciam a origem do indivíduo. Ora, quantas vezes não identificamos a terra natal de uma pessoa ao observar a sua fala? Isso porque em algumas regiões há características muito marcantes que se conservam, por tanto, no falar de nordestinos, mineiros, paulistas, cariocas, sulistas entre outros.

Esse falar peculiar, que denuncia a origem de um indivíduo, também é encontrado em textos bíblicos que Cardoso (2010, p. 27) considera "[...] uma exemplificação assaz distanciada no tempo e inteiramente desprovida de preocupação linguístico-científica" em que uma variante linguística é utilizada, por exemplo, para identificar fugitivos de Efraim, junto ao rio Jordão. O episódio é

[46] No Brasil, a variedade comumente considerada de prestígio é o falar encontrado no eixo Rio de Janeiro – São Paulo.

descrito na Bíblia Sagrada, no livro dos Juízes, capítulo 12, versículos 4 – 6, num contexto em que os efraimitas estão infiltrados entre os galaaditas. Assim, os vaus do Jordão foram ocupados por forças militares galaaditas e, para identificar se um indivíduo era efraimita, lhe pediam para pronunciar *chibolet* (espiga de trigo). Se o sujeito pronunciasse *sibolet*, ou seja, de maneira que não condizia com o falar dos galaaditas, era preso e degolado ali mesmo. Naquele tempo morreram 42 mil efraimitas (BÍBLIA SAGRADA, 1993, p. 239).

Para melhor compreender o funcionamento dos dialetos, tendo em vista que eles representam uma espécie de identidade linguística, é preciso considerar que o eixo dessa questão gira em torno de dois fatores: a liberdade de que cada falante dispõe para fazer suas escolhas lexicais e o sistema de possibilidades da língua que proporciona tais escolhas (COSERIU, 1980, p. 101-102).

Vale acrescentar que as escolhas lexicais, mediadas pelo sistema de possibilidades de uma língua, sofrem influências externas. Dessa forma, o estudo da norma lexical se depara com a relação existente entre as opções lexicais e as coisas, isto é, o mundo extralinguístico que, por sua vez, pode ser exemplificado pelo uso de objetos, ferramentas, dentre outras coisas do dia a dia dos falantes e que acabam se relacionando com algumas UL, formando uma espécie de rede semântica.

Assim, é importante distinguir o saber linguístico do saber das coisas, ou seja, do extralinguístico. Por exemplo, o item lexical *boi* remete a *vaca*, *touro, bezerro, chifres, ruminar, mugir, arado, canga.* Além disso, *boi* também é um elemento que integra determinadas expressões idiomáticas como, por exemplo, *colocar o carro diante dos bois, trabalhar que nem um boi* etc. No caso da relação entre *boi* e *arado* ocorre uma relação extralinguística, pois advém de uma experiência que se tem com o boi no trato da terra. É um saber cultural e próprio do labor realizado no campo. Porém, em culturas em que o boi é sagrado, ligados a sacrifícios e têm valores religiosos, como na Índia ou Egito antigo, essas associações seriam diferentes (COSERIU, 1980, p. 102-103).

Desse modo, a norma lexical de uma determinada comunidade de falantes carrega consigo um saber extralinguístico que se diferencia de região para região. Por exemplo, a item lexical *rapariga* evoca ideias distintas sobre a mulher em Portugal e no Brasil. Vale acrescentar que essas associações não são puramente linguísticas, já que são construções que nascem e se propagam em meio a contextos espaciais, sociais, econômicos, históricos, culturais e religiosos.

Se, por um lado, a Dialetologia é o campo de estudo dos dialetos partindo de uma perspectiva diatópica, por outro, a Geolinguística se constitui numa criteriosa metodologia de representação de dados por meio de mapas.

Sendo um método da Dialetologia, a Geolinguística busca, portanto, criar uma base empírica de dados com o objetivo de refletir e elaborar conclusões sobre a variedade linguística que ocorre em um certo lugar (CHAMBERS; TRUDGILL, 1994, p. 45).

Vale destacar que a metodologia utilizada, atualmente, pela Geolinguística é fruto de anos de evolução, afinal, essa é uma prática antiga. A primeira pesquisa dialetal da história foi realizada por Wenker, em 1876, que buscou representar a realidade dialetal alemã em 40.736 localidades. No entanto, esse trabalho sofreu severas críticas por gerar poucos resultados que foram publicados muito tempo após a coleta de dados. Concomitantemente aos problemas metodológicos dos estudos de Wenker, o *Atlas Linguistique de la France (ALF)*, liderado por Jules Gilliéron, estava em andamento acelerado. Desse modo, mesmo não sendo o primeiro ao iniciar um trabalho dialetológico na história da Dialetologia, a pesquisa de Gilliéron recebe o status de pioneira por conta de sua metodologia, consolidada por ser criteriosa e consistente (CARDOSO, 2010, p. 41-42).

A tarefa de percorrer a França, de bicicleta, para realizar os inquéritos dialetológicos para o *ALF* recaiu sobre Edmond Edmont, que possuía um ouvido muito aguçado para perceber as variações fonéticas dos entrevistados e foi treinado por Gilliéron, a fim de realizar adequadamente a transcrição fonética. Assim, Edmond percorreu 639 localidades que resultaram em uma média de 700 entrevistas. Os dados eram enviados a Gilliéron e seus ajudantes, periodicamente, agilizando assim o processo de análise linguística. Desse modo, as análises puderam ser publicadas em um curto espaço de tempo sendo possível estrear o primeiro volume em 1902 e o décimo terceiro em 1910 (CHAMBERS; TRUDGILL, 1994, p. 41).

A pesquisa de Julles Gillierón se transforma na pedra de toque para as pesquisas dialetais e para a produção de futuros atlas linguísticos. Desse modo, em 1931, ocorre a publicação dos primeiros volumes sobre os dialetos da Itália e do Sul da Suíça, dirigidos por dois discípulos de Gillierón, Karl Jaberg e Jakob Jud. Mais tarde, Jakob Jud e um dos três entrevistadores do projeto italiano treinam entrevistadores nos Estados Unidos para dar início aos trabalhos do Atlas Linguístico

dos Estados Unidos e Canadá. Em síntese, Gillierón exerceu um papel de mentor, orientando direta e indiretamente trabalhos dialetais desenvolvidos na Espanha, na Roménia e na Inglaterra (CHAMBER; TRUDGILL, 1994, p. 41-42).

Mais tarde, uma pesquisa sobre os dialetos ingleses tem início a partir de uma coleta de dados realizada entre 1950 e 1961. Coordenado por Eugen Dieth, de Zurich, e Harold Orton, de Leeds, os primeiros resultados aparecem em publicações realizadas entre 1962 e 1978 (CHAMBER; TRUDGILL, 1994, p. 44).

Constatamos, portanto, que o pioneirismo de Wenker e a qualidade metodológica desenvolvida por Gillierón motivaram a realização de diversas pesquisas dialetais, nacionais e regionais, em vários lugares do mundo.

No que tange à Dialetologia brasileira três obras representam o marco inicial dos estudos dialetais no país, a saber: i) *O dialeto caipira* de Amadeu Amaral, em 1920; ii) *A língua do Nordeste* de Mário Marroquim, em 1934; iii) *O linguajar carioca* de Antenor Nascentes, em 1953.

Em relação à produção de atlas linguísticos brasileiros, destacam-se os estaduais produzidos entre as décadas de 60, 70, 80 e 90, do século XX, como: o *Atlas prévio dos falares baianos* (APFB) coordenado por Nelson Rossi, em 1963; o *Esboço de um atlas linguístico de Minhas Gerais* (EALMG) elaborado por Mário Roberto Lobuglio Zágari, José Ribeiro, José Passini e Antônio Gaio, em 1977; o *Atlas linguístico da Paraíba* (ALPB) coordenado por Maria do Socorro Silva de Aragão, em 1984; o *Atlas Linguístico de Sergipe* (ALS) produzido por Nelson Rossi, Carlota Ferreira, Judith Freitas, Nadja Andrade, Suzana Cardoso, Vera Rollemberg e Jacyra Mota, em 1987 e o *Atlas linguístico do Paraná* (ALPR) de Vanderci de Andrade Aguilera, em 1994.

A produção de atlas linguísticos continua se desenvolvendo e ganha fôlego nas duas primeiras décadas do século XXI de modo que é possível destacar: o *Atlas Linguístico de Sergipe II* (ALS II) desenvolvido por Suzana Alice Marcelino Cardoso, em 2002; o *Atlas Linguístico-Etnográfico da Região Sul do Brasil* (ALERS) coordenado por Walter Koch, em 2002; o *Atlas Lingüístico Sonoro do Pará (ALISPA)* coordenado por Abdelhak Rasky, em 2004; o *Atlas Linguístico do município de Ponta Porã/MS* elaborado por Regiane Coelho Pereira Reis, em 2006; o *Atlas Geolinguístico do Litoral Potiguar* (ALiPTG) produzido por Maria das Neves Pereira, em 2007; o *Atlas Linguístico de Mato Grosso do Sul* (ALMS) organizado

por Dercir Pedro de Oliveira, em 2007; o *Atlas Semântico-Lexical da Região do Grande ABC* elaborado por Adriana Cristina Cristianini, em 2007; o *Atlas Linguístico da Mata Sul de Pernambuco* (ALMASPE) desenvolvido por Edilene Maria de Oliveira Almeida, em 2009; o *Atlas Linguístico da Mesorregião de Mato Grosso* (ALMESEMT) confeccionado por Marigilda Antônio Cuba, em 2009; o *Atlas Linguístico do Estado do Ceará* (ALECE) coordenado por José Rogério Fontenele Bessa, em 2010; o *Atlas Semântico-Lexical de Caraguatatuba, Ilhabela, São Sebastião e Ubatuba* – municípios do Litoral Norte de São Paulo, elaborado por Márcia Teixeira da Encarnação, em 2010; o *Atlas Geossociolinguístico de Londrina* (AGeLO) produzido por Valter Pereira Romano, em 2012; o *Atlas Linguístico de Pernambuco (ALiPE)* desenvolvido por Edmilson José de Sá, em 2013; o *Atlas Linguístico-Contatual da Fronteira entre Brasil/Paraguai (ALF-BR-PY)* produzido por Regiane Coelho Pereira Reis, em 2013; o *Atlas Linguístico de Corumbá e Ladário,* confeccionado por Beatriz Aparecida Alencar, em 2013 e o *Atlas Linguístico Pluridimensional do Português Paulista:* níveis semântico-lexical e fonético- fonológico do vernáculo da região do Médio Tietê, elaborado por Selmo Ribeiro Figueiredo Junior, em 2019.

É preciso salientar que a metodologia dialetal foi sendo aperfeiçoada ao longo dos anos. A coleta de dados tradicional era realizada, basicamente, com um tipo específico de informante: homem, idoso e sedentário. A tradição justifica essa escolha afirmando que o homem é aquele que guarda a língua vernácula com mais frequência, pois a mulher tende a ser mais reflexiva. Por sua vez, o idoso reflete a fala de uma época passada e o sedentário garante que aquele falar realmente pertence ao lugar onde vive o informante.

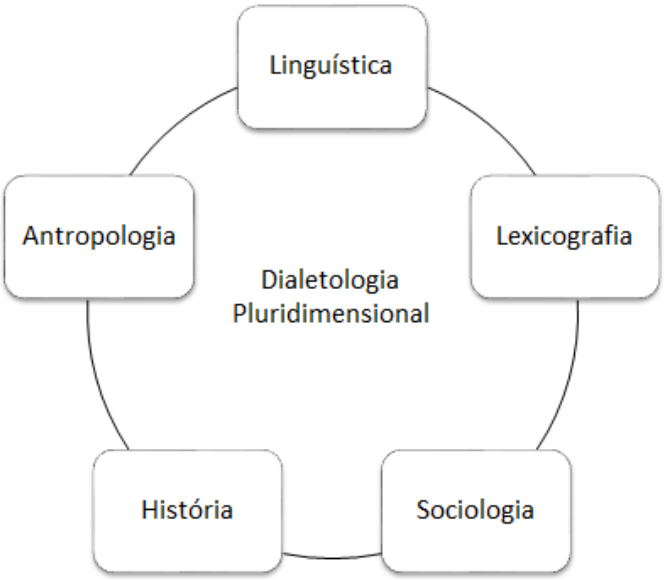
No entanto, essa visão foi mudando com o tempo e os critérios relacionados à seleção dos informantes, atualmente, contemplam variáveis como sexo (masculino e feminino), idade (idoso e jovem), escolaridade, e pessoas nascidas nos meios rural e urbanos.

Outro aspecto que complementou a base teórica da Dialetologia foi a inclusão do meio social como critério para coleta e análise dos dados. Ou seja, a variável social foi contemplada na pesquisa dialetológica sem perder de vista a sua essência diatópica (CARDOSO, 2010, p. 45).

Por tanto, a Dialetologia se beneficiou das pesquisas no campo da Sociolinguística ampliado seu escopo teórico. Assim, ao incorporar variáveis

diassexuais, diageracionais e todo o contexto social que caracteriza uma dada região geográfica, a Dialetologia transforma-se numa disciplina linguística de natureza pluridimensional, além de fornecer dados para pesquisas em outras áreas do conhecimento, conforme ilustra a figura a seguir:

Figura 24: A Dialetologia Pluridimensional e interfaces com outras áreas do conhecimento.



Fonte: Elaboração do autor.

O esquema da figura 24 ilustra o processo de evolução da Dialetologia, ao incorporar o aspecto social, difundido pela Sociolinguística na década de 1960. Além do mais, exemplifica que outras áreas do conhecimento se beneficiam dos estudos dialetológicos como é o caso desta Tese, que parte do dado dialetal, para desenvolver um produto lexicográfico.

Vale destacar que o dado dialetal coletado por meio de entrevistas com informantes deve ser de qualidade. Para tanto, é preciso realizar testes com o questionário antes da entrevista efetiva. Esse procedimento pode identificar a necessidade de ajustes em relação ao conteúdo das perguntas feitas aos informantes, evitando possíveis invalidações de perguntas por problemas na formulação para o entrevistado. Além disso, é importante estabelecer parâmetros em relação ao perfil do informante a ser entrevistado e aos locais que serão escolhidos para a coleta dos dados.

Nesse sentido, a metodologia do Projeto ALiB se destaca no cenário nacional, pois tem servido de parâmetro para diversas pesquisas dialetais em nível de iniciação científica, mestrado, doutorado e pós-doutorado. O Projeto ALiB teve início em 1996 e a coleta de dados começou em 2001, se estendendo até 2013. Os entrevistadores percorreram uma ampla rede de pontos com 250 localidades, que contemplam todas as regiões do país. Essas localidades foram escolhidas levando em consideração a extensão territorial, a densidade demográfica, além dos aspectos culturais e históricos que refletem no processo de povoamento de uma área (COMITÊ NACIONAL..., 2001, p. vii-viii).

Até o estágio atual o *Atlas Linguístico do Brasil* (CARDOSO et al, 2014a; 2014b) possui dois volumes publicados que reúnem dados de 25 capitais brasileiras. Os próximos volumes, que estão sendo preparados para publicação, trarão os resultados das 225 localidades do interior, além de dados adicionais das capitais.

É preciso destacar que o dado dialetal é uma fonte de pesquisa que pode subsidiar o desenvolvimento de diversificados produtos como, por exemplo, os próprios atlas linguísticos, além de dicionários gerais e especializados interessados em registrar as marcas de uso e a norma lexical de um idioma.

Em suma, a interface da Lexicografia com a Dialetologia resulta em trabalhos lexicográficos que buscam nos dados dialetais produzir obras lexicográficas que versam sobre os falares regionais. Nesse sentido, apresentamos e discutimos, na próxima seção, algumas considerações sobre a Lexicografia Dialetal.

3.2. Lexicografia Dialetal

A Lexicografia Dialetal ocupa-se em documentar as particularidades regionais de uma língua em dicionários e/ou vocabulários dialetais. Esse importante trabalho tem o potencial de oferecer aos usuários informações sobre a norma lexical de uma dada região, que são identificados ao serem comparados com a norma padrão (GRANJA; SEONE, 2018, p. 9) e, principalmente, por meio de critérios estabelecidos pela metodologia da Geolinguística.

É importante destacar que uma UL é reconhecida como dialetal por meio da junção desses critérios, especialmente o diatópico, pois, isoladamente, tais critérios não são suficientes para determinar a norma lexical: “Cosa bien distinta es que luego muchas de estas obras incluyan voces, como popularismos o vulgarismos, que no

pueden considerarse en sentido estricto dialectalismos en la medida en que no presentan una distribución territorial restringida"[47] (GRANJA; SEONE, 2018, p. 10).

Outro ponto de suma relevância no âmbito da Lexicografia Dialetal é a fonte das informações que precisam, de fato, apresentar o caráter regional. Desse modo, a coleta de dados realizada pelo Projeto ALiB tem se tornado em um rico material para a elaboração de obras lexicográficas dialetais como, por exemplo, os vocabulários desenvolvidos no âmbito da pós-graduação baseados nos dados do ALiB e que visam a contribuir com o Projeto do *Dicionário Dialetal Brasileiro* (MACHADO FILHO, 2010), conforme mencionado na introdução desta Tese.

Se faz necessário, ainda, destacar que a coleta de dados realizada pelo Projeto ALiB não foi exaustiva, mas, baseada na seleção de áreas lexicais (GRANJA; SEONE, 2018, p. 10) como, por exemplo, o Questionário Semântico-lexical (QSL), em que as 202 questões foram organizadas a partir de 14 áreas semânticas[48].

Tendo em vista que um dicionário dialetal necessita de uma fonte confiável de dados para que possa registrar, com veracidade, o falar regional de uma dada localidade, Esquerra (1997, p. 79) esclarece que:

> […] la Lexicografía ha necesitado acudir frecuentemente a la Dialectología para tomar sus informaciones, así como la Dialectología se ha dirigido a la Lexicografía para comprobar sus datos. Son dos los ámbitos en los que nuestras disciplinas entran en contacto: el de la presencia de voces dialectales, regionales, locales, etc., en los diccionarios, y el de los repertorios dedicados a esos tipos de palabras[49].

A posição desse lexicógrafo justifica o caráter interdisciplinar da Lexicografia, além de mostrar a interface dessa disciplina com outros campos de estudo em uma espécie de *via de mão dupla*, isto é, a Lexicografia se beneficia da Dialetologia na consulta de dados geolinguísticos, ao passo que a Dialetologia se

[47] "Coisa muito distinta é que logo muitas destas obras incluem vozes, como popularismos ou vulgarismos, que não podem considerar-se, em sentido estrito, dialetalismos na medida em que não apresentam uma distribuição territorial restringida" (T.N.).

[48] As 14 áreas semânticas do QSL-ALiB são: 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.

[49] "[...] a Lexicografia tem precisado recorrer frequentemente à Dialetologia para tomar suas informações, assim como a Dialetologia tem se dirigido à Lexicografia para comprovar seus dados. São dois os âmbitos em que nossas disciplinas entram em contato: o da presença de vozes dialetais, regionais, locais, etc., nos dicionários, e dos repertórios dedicados a esses tipos de palavras" (T.N.).

utiliza dos repertórios da Lexicografia para testar suas hipóteses.

Dessa forma, os dados dialetais não são importantes apenas porque fornecem informações sobre o falar de determinada região, pois o uso desse tipo de *corpus* confere, acima de tudo, rigor metodológico que solidifica a elaboração de um produto lexicográfico dialetal:

> [...] la geografía lingüística es un campo virgen para la lexicografía, y es donde sus frutos poden resultar sorprendentes. Desde luego, siqueremos una lexicografía rigurosa, hoy en día, no hay más remedio que echar mano de los atlas lingüísticos[50] (NAVARRO CARRASCO, 1993).

Fica explícito, a partir do exposto, a importância dos dados geolinguísticos na elaboração de obras lexicográficas dialetais de qualidade. Nesse sentido, o acesso a tais informações pode ser feito por meio dos atlas linguísticos já publicados ou pelo acesso aos bancos de dados construídos mediante critérios geolinguísticos, com é o caso do *corpus* do Projeto ALiB. Nessa perspectiva, a constituição desses conjuntos textuais se configura como a base, o ponto de partida para a extração de informações que objetivam servir aos propósitos de qualquer pesquisa científica. Assim, apresentamos, na seção a seguir, a Linguística de *Corpus* que modificou, em larga escala, as pesquisas científicas em todas as áreas do conhecimento.

3.3. Linguística de *Corpus*

John Sinclair, trinta anos atrás, escreve na introdução de seu livro intitulado *Corpus, Concordance, Collocation* que:

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular[51] (SINCLAIR, 1991, p. 1).

[50] "[...] a geografia linguística é um campo virgem para a lexicografia, no qual seus frutos podem se tornar surpreendentes. Dessa maneira, se queremos uma lexicografia rigorosa, atualmente, não há mais remédio a não ser lançar mãos dos atlas linguísticos" (T.N.).

[51] "Trinta anos atrás, quando esta pesquisa começou, era considerado impossível processar textos de vários milhões de palavras. Vinte anos atrás, era considerado marginalmente possível, mas lunático. Dez anos atrás, era considerado bastante possível, mas ainda lunático. Hoje é muito popular" (T.N.).

Com base no exposto por Sinclair (1991), o que era considerado muito popular, há trinta anos, hoje é essencial. A constituição de um *corpus* que, na prática, é formado por uma grande quantidade de textos, permite a manipulação de dados textuais com finalidades diversificadas.

Vale destacar que a Linguística de *Corpus* é uma metodologia anterior à era do computador, que tem como marco os anos 1940. Desse modo, pesquisas comparativas com um grande volume de textos já eram realizadas no século XVIII por equipes de escribas. Esses religiosos utilizavam métodos para criar listas de palavras com a indicação de onde elas ocorriam na Bíblia, juntamente com a frequência de cada item lexical. Esse laborioso serviço, realizado manualmente por centenas de estudiosos, é semelhante às listas de palavras que podem ser geradas, automaticamente, por um software de concordância, que pode ser operado por pessoas que saibam processar dados em um concordanceador (O'KEEFFE; McCARTHY, 2010, p. 3).

O fato é que o desenvolvimento computacional alavancou a Linguística de *Corpus* de tal maneira que, atualmente, é difícil pensar nesse método sem ligá-lo a um software executando alguma tarefa em um determinado *corpus*. Esse fato também ampliou o leque de atuação da Linguística de *Corpus* sendo possível realizar estudos em diversas áreas como, por exemplo, ensino e aprendizagem de línguas, análise do discurso, estilística, linguística forense, pragmática, sociolinguística entre outras (O'KEEFFE; McCARTHY 2010, p. 7). Outro exemplo pode ser visto em ferramentas de tradução automática como o *Linguee*[52] que mostra ao usuário alguns contextos de uso da palavra pesquisada nos dois idiomas, isto é, na língua de partida e de chegada.

Desse modo, a Linguística de *Corpus* é compreendida como um método de compilação de dados autênticos, ou seja, retirados de seu contexto de uso, com a finalidade de servir aos objetivos de uma pesquisa científica (BERBER SARDINHA, 2004, p. 18-19). Esses dados podem ser de origem textual ou oral. No uso desses últimos, precisam ser transcritos para que o computador possa reconhecê-los e executar o processamento e a recuperação das informações relevantes ao pesquisador.

É importante destacar, todavia, que a seleção de um conjunto de textos precisa seguir critérios que garantam a representatividade do *corpus*, ou seja, precisa reunir uma quantidade de textos que atendam ao objeto da pesquisa. Desse modo, um estudo que

[52] A ferramenta pode ser acessada em <https://www.linguee.com.br/>.

pretenda analisar a terminologia da agricultura, por exemplo, necessita de um *corpus* balanceado com textos de diferentes gêneros textuais que versem sobre essa área de especialidade[53]. Recorrer a textos jornalísticos, nesse caso, refletiria apenas no acréscimo ao *corpus* de dados que apresentam baixo valor terminológico, pois esse gênero é voltado para o público leigo o que diminui as chances de se utilizar uma linguagem mais técnica.

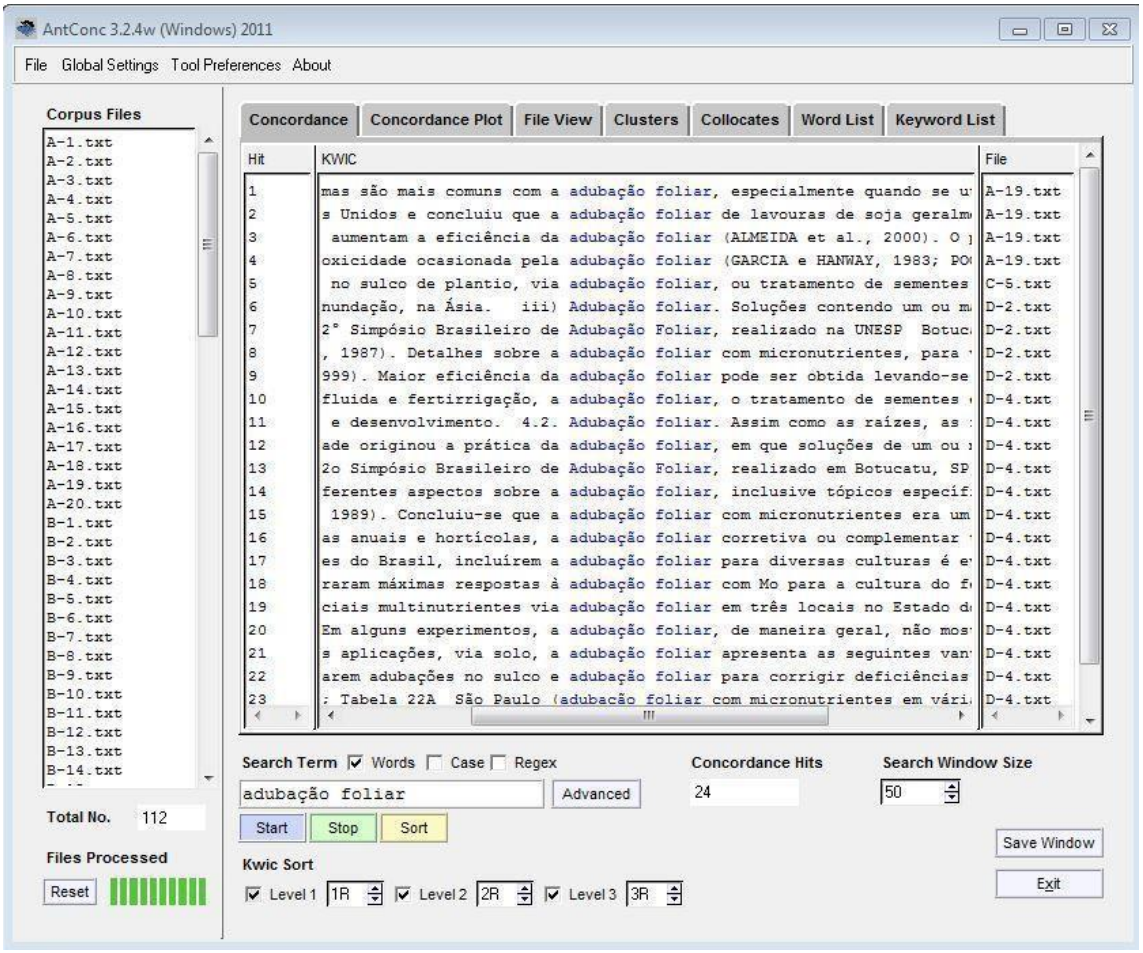
Para McEnery; Hardie (2012, p. 1), a Linguística de *Corpus* é "[…] an area which focuses upon a set of procedures, or methods, for studying language[54]." Esse conjunto de métodos envolve não apenas a compilação de *corpora* a partir de critérios predefinidos, como também abrange o uso de ferramentas computacionais capazes de realizar o Processamento de Linguagem Natural como, por exemplo, o uso de um concordanceador[55], que permite a extração do contexto de uso de uma determinada UL e a exibição do resultado para o usuário de maneira alinhada e destacada, como demostra a figura a seguir:

[53] Santos Junior (2015, p. 53-55), ao publicar o *Glossário de termos da agricultura: um estudo terminológico sobre o manejo do solo*, compilou um *corpus* composto de textos de diferentes gêneros , a saber: i) científicos; ii) científicos de divulgação; iii) instrucionais; iv) boletins técnicos.

[54] "[...] uma área que se concentra num conjunto de procedimentos, ou métodos, para estudar a linguagem". (T.N.).

[55] O concordanceador é uma ferramenta que pode ser encontrada em pacotes de software proprietário como o *WordSmith Tools* <https://lexically.net/wordsmith/>, ou em software livre, a exemplo do *AntConc* <https://www.laurenceanthony.net/software/antconc/>.

Figura 25: Contexto de uso de uma unidade lexical extraída a partir de um concordanceador.



Fonte: Santos Junior (2015, p. 72).

A figura 25 focaliza o funcionamento do concordanceador que é disponibilizado pelo software *AntConc*. Desse modo, o termo *adubação foliar* aparece em destaque e centralizado na tela, permitindo que o usuário consulte o seu contexto. Para ampliação do conteúdo desse contexto, basta clicar no termo desejado para que uma nova janela seja aberta, exibindo todo o arquivo de onde esse item lexical foi extraído.

Além dos concordanceadores, outra ferramenta essencial para a manipulação de dados a partir de *corpora* é aquela que realiza a contagem das unidades lexicais e as ordena por frequência. Essa ferramenta é denominada de *word list* e também está disponível nos pacotes do *AntConc* e *WordSmith Tools*.

Esses são apenas dois exemplos de funcionalidades possíveis de se operar quando juntamos ferramentas computacionais e um *corpus*. Há outras possiblidades

de manipulação, que não iremos detalhar, como por exemplo, ferramentas que realizam o agrupamento de unidades lexicais para encontrar termos compostos por mais de um vocábulo ou fraseologias, além de ferramentas que combinam a *word list* a um conjunto de parâmetros predefinidos, a fim de gerar uma lista de palavras-chave com a porcentagem de chavissidade de cada termo, sendo uma ferramenta muito útil na elaboração de glossários terminológicos.

Nesse sentido, destacamos que o uso de *corpora* em pesquisas linguísticas, sobretudo no labor lexicográfico, contribui em larga escala para agilizar as etapas metodológicas envolvidas na produção de dicionários que, em tempos muito remotos, eram realizadas manualmente, demandando muito tempo e esforço do dicionarista. Contemporaneamente, entretanto, há muitos softwares que realizam diversas tarefas no âmbito das pesquisas em Linguística e em Lexicografia conferindo, não apenas maior rapidez no desenvolvimento de etapas metodológicas, como também permitindo a manipulação de enormes quantidades de dados.

Em síntese, o uso de *corpus* em uma parcela das pesquisas científicas se tornou imperativo. Além de abrir novos horizontes (SINCLAIR, 1991, p. 14) promove um enorme ganho quantitativo que se reverte em análises qualitativas. Além do mais, a uso integrado de *corpus/corpora* na Linguística pode proporcionar a reorientação de abordagens metodológicas, refinando e reorientando uma série de teorias da linguagem (McENERY; HARDIE, 2012, p. 1).

Na realidade, a Linguística de *Corpus* representa o primeiro passo de uma revolução que transformou e continua transformando, em grandes proporções, os estudos linguísticos, à medida que o campo da Ciência da Computação se desenvolve e novas soluções computacionais são criadas e absorvidas pelas Ciências Humanas e Sociais promovendo, dessa maneira, reorientações teóricas e metodológicas em diversas áreas como é o caso desta Tese, que teve um redirecionamento a partir de um contato mais aprofundado coma Linguística Computacional, que será apresentada na seção seguinte.

3.4. Linguística Computacional

O crescimento acelerado do universo informático tem proporcionado grandes avanços na área do desenvolvimento de softwares e, consequentemente, tem surgido um número maior de programas capazes de processar a linguagem natural. Esse panorama

beneficia pesquisadores interessados em manipular dados textuais permitindo, dessa maneira, realizar análises, testar e comprovar hipóteses, criar produtos entre outras possibilidades.

No âmbito da Linguística Computacional é muito comum o uso da expressão Processamento de Linguagem Natural (PLN). Kedia e Rasu (2020, p. 7) definem o PLN como uma "interdisciplinary area of research aimed at making machines understand and process human languages[56]." Isso significa que é por meio do PLN que o computador passa a compreender uma língua natural para que seja possível o processamento desses dados pela máquina. Os autores também ilustram que é por meio do PLN que a indústria computacional tem desenvolvido produtos que estão se popularizando rapidamente na sociedade como, por exemplo, *Alexia*, *Google Tradutor* e *chatbots*.

Destacamos, ainda, que o PLN é uma metodologia que utiliza recursos computacionais para desenvolver pesquisas voltadas para a manipulação da linguagem humana em diferentes frentes, ou seja, além de ser possível observar sua aplicação no ramo da Linguística, o PLN também é muito presente em pesquisas no campo da Ciência da Computação, tendo em vista que pertencente a uma subárea da Inteligência Artificial (KEDIA; RASU, 2020, p. 9).

Para além do conceito de PLN se faz necessário compreender e distinguir o que vem a ser a Linguística Computacional. Desse modo, uma elucidativa definição desse ramo de investigação é a de Srinivasa-Desikan (2018, p. 11): "[…] 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[57]."

O autor menciona como exemplo de estudos linguísticos assistidos por computadores a marcação morfossintática[58] de um texto, tarefa comumente realizada por seres humanos. Assim, é possível anotar um *corpus* automaticamente por meio da comparação de *corpora*, ou seja, alimentamos o computador com o conjunto textual que deverá ser anotado e o software, por meio de comparação estatística, irá proceder com a anotação. Dessa forma, quanto maiores forem os *corpora* anotados no computador,

[56] "...área interdisciplinar de pesquisa que visa a fazer com que as máquinas entendam e processem as linguagens humanas" (T.N.).
[57] "[…] 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" (T.N.).
[58] Para maiores informações acesse: <https://www.cis.lmu.de/~schmid/tools/TreeTagger/>.

maior será a eficiência das anotações automáticas, pois os à medida que os modelos são treinados[59], uma considerável quantidade de parâmetros sobre a morfossintaxe de um idioma são armazenados.

É relevante destacar que no âmbito da Linguística Computacional é possível utilizar os recursos informáticos, basicamente, de três maneiras gerais em um projeto científico: i) por meio de softwares que foram desenvolvidos para executarem um determinado tipo de tarefa; ii) a partir do desenvolvimento das próprias ferramentas computacionais; iii) mesclando as duas opções anteriores. A primeira opção é bastante utilizada pelo fato de não exigir conhecimentos de programação do pesquisador. Porém, o estudioso fica limitado a executar tarefas eletronicamente para a qual o software foi projetado. A segunda opção abre um leque de possibilidades para o linguista criar, a partir de suas necessidades de pesquisa, ferramentas computacionais de maneira personalizada. Todavia, exige conhecimentos de programação que podem ser terceirizados a um profissional da área ou, se for o perfil do estudioso em questão, iniciar um processo de aprendizagem de programação voltado para seu campo de pesquisa. Finalmente, a última opção é a mais equilibrada do ponto de vista da Informática, pois, por mais que se tenha habilidade em programar para desenvolver as próprias aplicações computacionais, sempre será preciso utilizar softwares já existentes, afinal, não há necessidade de se *reinventar a roda*.

Desse modo, um programa muito utilizado para a análise lexical é o *AntConc*, mencionado na seção anterior, que consiste em um pacote de softwares desenvolvido pelo professor Laurence Anthony, na Faculdade de Ciência e Engenharia da Universidade de Waseda, no Japão e reúne um conjunto de ferramentas que permite ao linguista realizar tarefas de PLN a partir de um *corpus* já definido. Assim, é possível gerar uma lista de unidades lexicais organizadas em ordem de frequência, consultar o contexto de cada vocábulo por meio de um concordanceador, gerar combinações de unidades lexicais compostas, além de contrastar um *corpus* de estudo com um *corpus* de referência, a fim de gerar uma lista de palavras-chave.

Outro exemplo de aplicação muito utilizada na elaboração de pesquisas acadêmicas e na construção de dicionários de pequeno porte é o *FieldWorks Language*

[59] No ramo da das Ciências Computacionais, sobretudo na área da Inteligência Artificial, um *corpus* é denominado *treinado* quando possui uma grande quantidade de informações sobre alguma particularidade da língua o que reflete em uma porcentagem alta de êxito na execução de uma atividade automaticamente como, por exemplo, na anotação morfossintática de outro *corpus*.

Explorer (*FLEx*), desenvolvido pelo *Summer Institute of Linguistic* (*SIL International*). Esse programa possibilita a criação de um banco de dados lexicográficos e, posteriormente, esses dados podem ser organizados em forma de verbetes. Há, ainda, a possibilidade de publicar o dicionário compilado no *FLEx* em um website sem a necessidade de conhecimentos de programação, por meio do uso de uma ferramenta adicional[60]. No entanto, o linguista fica à mercê das funcionalidades que esse programa oferece e não poderá acrescentar tarefas para além daquilo que o software foi projetado.

Essa limitação, que qualquer software apresenta, é compreensível tendo em vista que os programas de computador são planejados para executarem um número finito de tarefas. Desse modo, muitas vezes é preciso recorrer a outros programas durante as etapas metodológicas de uma pesquisa científica e, ainda, considerar a possibilidade de que algumas tarefas específicas, inicialmente planejadas pelo pesquisador, poderão ficar sem opção de automação.

Uma possível solução para resolver o problema da limitação de determinados softwares, já que as tarefas previstas por um programador podem diferir das necessidades de um usuário em específico, é o linguista se apropriar de linguagens de programação com a finalidade de criar suas próprias aplicações informáticas atendendo, dessa forma, aos objetivos da sua pesquisa.

Tendo em vista que a aquisição desse conhecimento técnico demanda tempo, é preciso que o pesquisador centralize seus esforços na aprendizagem de linguagens de programação que sejam úteis, já que no universo computacional há uma infinidade de temas que podem ser explorados. Nesse sentido, a orientação de um profissional da área é importante para nortear o pesquisador em relação aos caminhos que podem ser tomados para o desenvolvimento de aplicações específicas que atendam aos objetivos de um determinado projeto.

Destacamos, ainda, que desenvolver os próprios softwares não implica apenas se libertar das limitações que os programas de computador podem apresentar. Lançar mão dessa investida é, sobretudo, ampliar as possibilidades de pesquisa no ramo da Linguística abrindo caminhos para um novo horizonte a ser explorado, pois:

> Les instruments permettent de voir de nouveaux phénomènes. Leur emploi effectif ne va pas cependant sans ajustement nécessaire des données à

[60] Denominada *webonary*, a ferramenta tem publicado até o momento da escrita desta Tese, 295 trabalhos lexicográficos desenvolvidos em diversos países por meio da plataforma *FLEx*. Para maiores informações acesse: <https://www.webonary.org/>.

> observer. A l'inverse, les instruments se prêtent au détournement qui leur permet de traiter des données pour lesquelles ils n'avaient pas été prévus initialement, ouvrant la voie à de nouveaux progrès[61] (HABERT, 2005, sem páginas).

O termo *instruments* (instrumento) utilizado por Habert (2005) refere-se ao ato de equipar-se com os recursos computacionais necessários ao desenvolvimento da pesquisacientífica, à medida que vai além da simples utilização de um software que foi construído para realizar um par de tarefas. Assim, o autor faz um convite aos linguistas para que aprendam a lidar com linguagens de programação, a fim de ampliar os seus horizontes de pesquisa.

Em uma publicação intitulada *Outiller la linguistique: de l'emprunt de techniques aux rencontres de savoirs*[62], em 2004, Benoît Habert esclarece que muitas pesquisas no ramo da Linguística não exigem conhecimentos detalhados de Informática. Todavia, o trabalho linguístico será muito mais proveitoso se incorporar em sua metodologia o uso de instrumentos que a Linguística Computacional pode oferecer. Desse modo, o autor exorta para que o pesquisador desconstrua possíveis medos, que se mostrem como dificuldades frente ao aprendizado necessário ao domínio dos conteúdos relacionados à Informática que, por sua vez, podem enriquecer a pesquisa linguística (HABERT, 2004, sem página).

Não se tem a intenção, contudo, de advogar que o linguista ou lexicógrafo deva ser um especialista em programação. Todavia, queremos mostrar que o estudioso interessado poderá utilizar em sua metodologia soluções informáticas próprias. Além do mais, referimo-nos aqui às pesquisas em nível de pós-graduação, que não possuem recursos para contratar uma equipe de trabalho com especialistas de outras áreas como ocorre, frequentemente, em projetos de grande porte como é o caso das editoras voltadas para o mercado lexicográfico.

Nesse sentido, o pesquisador que deseje desenvolver suas próprias soluções computacionais pode se apropriar apenas das técnicas que utilizará em sua pesquisa. Para tanto, como já mencionado, é importante consultar um especialista do campo da Ciência da Computação a fim de que esse profissional indique os possíveis caminhos a

---

[61] "Os instrumentos nos permitem ver novos fenômenos. Seu uso real, entretanto, não ocorre sem o ajuste necessário dos dados a serem observados. Por outro lado, os instrumentos se prestam ao redirecionamento que lhes permite processar dados para os quais não foram inicialmente previstos, abrindo caminho para novos avanços." – (T.N.).

[62] "Equipando a linguística: do empréstimo de técnicas ao conhecimento" (T.N.).

serem trilhados e os conhecimentos que deverão ser mobilizados.

Desse modo, no capítulo seguinte, detalharemos os procedimentos metodológicos desta Tese, que pavimentaram a estrada que nos levou ao alcance de resultados bastante satisfatórios.

CAPÍTULO 4 – PERCURSO METODOLÓGICO

O percurso metodológico desta pesquisa foi marcado por momentos em que o tradicional foi confrontado com o inovador, ou seja, o habitual processo de enxergar o labor lexicográfico a partir da Lexicografia Impressa teve que ceder espaço para um ângulo de visão que contemplasse a Lexicografia Eletrônica. Assim, algumas decisões foram necessárias e importantes frente às descobertas que se seguiram durante os quatro anos de pesquisa. Dentre as escolhas realizadas, podemos destacar dois momentos cruciais que modificaram, substancialmente, a trajetória desta Tese, a saber: i) a elaboração do banco de dados em *XML* e o desenvolvimento de uma solução para tarefas de PLN; ii) o alinhamento do protótipo do *VoDiNorte* aos pressupostos teórico-metodológicos da Lexicografia Eletrônica.

Dessa forma, o primeiro momento se deu a partir de duas disciplinas que foram oferecidas, na Universidade Federal de Mato Grosso do Sul, em cumprimento a um convênio internacional celebrado por essa instituição com a *Université Sorbonne Nort*. Esse convênio possibilitou que o professor Fabrice Charles Bernard Issac viesse ao Brasil, na categoria de Professor Visitante no Programa de Pós-Graduação em Estudos de Linguagens, da UFMS, Faculdade de Artes, Letras e Comunicação, para ministrar duas disciplinas sobre o uso da Informática e das linguagens de programação aplicadas aos estudos linguísticos. Esses cursos possibilitaram a abertura de novos horizontes em relação aos estudos que usam um *corpus* eletrônico para a investigação linguística.

O segundo momento, por sua vez, teve início durante o processo de escrita do texto submetido ao Exame de Qualificação e foi, aos poucos, tomando fôlego a partir do acesso às novas leituras relacionadas ao planejamento e ao uso de dicionários eletrônicos. Dessa maneira, alguns ajustes foram realizados no que diz respeito à apresentação dos dados lexicográficos aos usuários e, acima de tudo, percebemos que não é possível desenvolver um produto lexicográfico inovador, no âmbito da Lexicografia Eletrônica, sem descontruir o modelo de verbete tradicional que guardamos, inconscientemente, em nossa memória habituada ao consumo de dicionários impressos.

Dessa maneira, descrevemos nas seções seguintes cada uma das etapas metodológicas realizadas nesta Tese. Buscamos assumir uma atitude de descrição pormenorizada e didática, pois acreditamos que esse conteúdo poderá interessar a outros pesquisadores que queiram explorar, como maior profundidade, a temática aqui exposta.

4.1. O *corpus* oral do ALiB

O ALiB é um projeto interinstitucional iniciado em 1996, cuja sede fica na Universidade Federal da Bahia. Seu objetivo principal é descrever os falares das cinco regiões do Brasil por meio de mapas linguísticos, entre outros estudos. Para tanto, uma minuciosa coleta de dados foi realizada entre 2001 e 2013 que resultou em um *corpus* oral com perguntas e respostas de informantes moradores de 250 localidades espalhadas por todo o país.

A escolha das localidades, denominadas rede de pontos do ALiB, levou em conta critérios como, por exemplo, densidade demográfica, aspectos históricos e culturais.

Os informantes, por sua vez, foram definidos segundo critérios que singularizam o nível de escolaridade dos das capitais em relação aos das localidades do interior. Dessa forma, oito informantes foram entrevistados nas capitais, quatro deles com nível universitário de escolaridade e quatro com Ensino Fundamental incompleto. Ao seu turno, nas cidades do interior, as entrevistas são feitas com quatro informantes com o nível fundamental de escolaridade.

Destacamos, ainda, que os entrevistados são distribuídos igualmente entre os sexos feminino e masculino, além de abranger duas faixas etárias, a saber, 18 a 30 anos e 50 a 65 anos. Os entrevistados também devem ter uma profissão definida, que não requeira grande mobilidade, que sejam nascidos na localidade investigada e que sejam filhos de pais dessa mesma região (COMITÊ NACIONAL ..., 2001, p. viii).

Como mencionado anteriormente, a fatia de entrevistas realizadas pelo Projeto ALiB que receberam um tratamento eletrônico e lexicográfico no âmbito desta Tese corresponde aos dados coletados junto a informantes nascidos e criados nos 18 municípios do interior, da rede de pontos do ALiB, na região Norte do Brasil, como é possível observar no quadro a seguir:

Quadro 6: Rede de pontos do Projeto ALiB referente à região Norte[63].

ESTADO	CIDADE
Amapá	1. Oiapoque
	2. Macapá (capital)
Roraima	3. Boa Vista (capital)
Amazonas	4. São Gabriel da Cachoeira
	5. Tefé
	6. Manaus (capital)
	7. Benjamin Constant
	8. Humaitá
Pará	9. Soure
	10. Óbidos
	11. Almeirim
	12. Belém (capital)
	13. Bragança
	14. Altamira
	15. Marabá
	16. Jacareacanga
	17. Conceição do Araguaia
	18. Itaituba
Acre	19. Cruzeiro do Sul
	20. Rio Branco (capital)
Rondônia	21. Porto Velho (capital)
	22. Guajará Mirim
Tocantins	23. Pedro Afonso
	24. Natividade

Fonte: Elaboração do autor.

Multiplicando-se os 18 municípios do interior pelo número de entrevistas realizadas em casa localidade, isto é, por quatro, temos um total de 72 inquéritos gravados em áudio que foram transcritos no banco de dados em *XML*.

[63] Esta Tese trabalhou apenas com os dados as cidades do interior. Assim, as capitais figuram no quadro 6 apenas para ilustrar a numeração adotada pelo Projeto ALiB referente à rede de pontos.

A tarefa de transcrição se configura em uma atividade morosa e os softwares de reconhecimento de voz não foram eficientes no auxílio desse labor, já que era preciso revisar o que a máquina escrevia acarretando um gasto maior de tempo em relação à escrita de modo manual. Desse modo, a transcrição dos áudios foi realizada de forma manual com auxílio de um software para executar o arquivo de áudio e o outro para escrever o *XML*.

4.2. A construção do primeiro banco de dados em *XML*

A primeira etapa correspondente ao tratamento eletrônico dos dados do Projeto ALiB utilizados nesta pesquisa foi a criação de um arquivo *XML* para armazenar as perguntas e respostas das 202 questões do QSL. Destacamos, nesse contexto, que o *corpus* em *XML* pode ser compreendido como um banco de dados[64], tendo em vista que se trata de um conjunto de informações organizadas em uma estrutura arbórea.

Desse modo, o banco de dados em *XML* foi pensado para estruturar os dados provenientes das gravações do ALiB de modo que fosse possível acessar, individualmente ou em conjunto com outros pares de dados, informações relacionadas às falas de cada informante, além de filtrá-las por meio das variáveis sexo, idade, escolaridade e localidade. Ao mesmo tempo, o banco de dados deve oferecer a consulta de informações tratadas lexicograficamente, ou seja, o usuário poderá visualizar um conjunto de dados em formato de verbete lexicográfico e, para que isso ocorra, as informações devem estar devidamente organizadas para figurarem em campos específicos do verbete. Esse procedimento é compreendido como o tratamento lexicográfico dos dados.

A melhor forma de descrever o funcionamento do *XML* é com o ato de etiquetar gavetas de um grande armário. Nessa analogia, o armário contém milhares de gavetas que recebem um rótulo de acordo com o tipo de informação armazenada garantindo,

[64] No campo da Ciência da Computação o termo *banco de dados* está fortemente ligado à *Linguagem de Consulta Estruturada*, originário do termo em inglês *Structured Query Language* (*SQL*). A linguagem *SQL*, por sua vez, é utilizada para realizar a recuperação de informações em bancos de dados relacionais por meio de softwares como, por exemplo, *Oracle Database*, *MySQL*, *QSL Server* entre outros, que são considerados *Sistemas de Gerenciamento de Bancos de Dados* (SGBD). Todavia, o conceito de banco de dados também pode ser utilizado na construção de arquivos que armazenam informações em *XML*, tendo em vista que os dados podem ser manipulados por meio de softwares como o *BaseX*. Para maiores informações acesse: <https://basex.org/>. Acesso em: 15 abr. 2023.

assim, um rápido acesso aos dados. No arquivo *XML* também é possível criar grupos de gavetas, a fim de subclassificar um conjunto de informações.

É preciso ressaltar que a estrutura de um banco de dados *XML* deve ser planejada e testada previamente antes de implementar o armazenamento das informações em definitivo. Estabelecer uma etapa de ajustes é muito importante, pois permite ao pesquisador verificar, por meio de testes com uma pequena amostra de dados, se existe algo a ser mudado na estrutura do documento e se os objetivos da pesquisa serão alcançados com a estrutura atual do *XML*. Esse alerta é válido para evitar a realização de futuros ajustes na estrutura do projeto. No entanto, caso seja preciso alterar a composição do *XML* e o processo de alimentação do banco de dados estiver bastante avançado, o pesquisador deve ter em mente que terá um trabalho longo de edição, já que deverá percorrer todas as linhas do arquivo para realizar os ajustes necessários.

Outro ponto fundamental que foi planejado para a estruturação do banco de dados em *XML* se relaciona ao tratamento lexicográfico das informações dialetais presentes nos áudios. Dessa maneira, as transcrições foram feitas diretamente no interior das *tags*[65] do *XML* que estão organizadas e nomeadas a partir dos elementos de uma microestrutura lexicográfica, como é possível observar na figura a seguir.

[65] As *tags* podem ser compreendidas como as gavetas de um grande armário que armazenam dados e são escritas a partir de uma sintaxe específica como, por exemplo, <lema>igarapé</lema> no qual *igarapé* está etiquetado pelo elemento<lema> de abertura e </ lema> de fechamento.

Figura 26: Estrutura do arquivo *XML*.

```
1  <?xml version="1.0" encoding="utf8" ?>
2  <!DOCTYPE dicio SYSTEM "corpus-1.dtd">
3
4  <dicio>
5       <entrada id="acid.geo.água.1" abc="i">
6           <lema>igarapé</lema>
7           <perg campo="Acidentes geográficos" ref="QSL-1">Como se chama um rio pequeno, de uns dois metros de largura?</perg>
9           <ex>Aqui é garapé, né...(Tem outros nomes?) Não. Aqui é garapé só.</ex>
10          <obs></obs>
11          <fone>garapé</fone>
12          <aud src="nome-do-arquivo" type="mp3">001_01_QFF01_QSL051_A == 48:56</aud>
13          <ver name="igarapé" ref="acid.geo.água.1"/>
14          <info sexo="M" escolaridade="F" idade="J" > 28 anos</info>
15          <lg ponto="1" cidade="Oiapoque" estado="AP" />
16          <gram></gram>
17          <def></def>
18         <map src="" type="jpg"/>
19     </entrada>
20 </dicio>
```

Fonte: Elaboração do autor.

A figura 26 exibe as primeiras linhas do *XML* que foram escritas com o software jEdit[66]. As linhas 1 e 2 exibem uma declaração que identifica o tipo de arquivo que estamos construindo, no qual a linha 1 exibe a versão do *XML* (1.0) e a codificação dos caracteres (utf8), e a linha 2 informa que o arquivo *XML* está sendo executado juntamente com um arquivo de validação (*corpus*-1.dtd) que padroniza a estrutura do banco de dados. Essa validação ocorre por meio de um arquivo denominado *Document Type Definition (DTD)* que é responsável por monitorar a estrutura do *XML*, garantindo a integridade do documento em relação à composição das *tags*.

Em linhas gerais, o *DTD* representa um conjunto de regras para a formatação do *XML*. Esse documento é construído no momento que se estabelece a estrutura do banco de dados e quando o *DTD* detecta algum erro de digitação que comprometa a arquitetura das *tags* do *XML*, uma mensagem de erro aparece na tela solicitando a atenção do usuário.

Na linha 4, da figura 26, temos a *tag* de abertura do documento *XML* (<dicio>) e na linha 20 a *tag* de fechamento (</dicio>). Isso significa que todas as informações do banco de dados estão dentro dessas duas *tags* que podem ser entendidas, em nossa

[66] Para maiores informações acesse: <http://www.jedit.org/>.

analogia com as gavetas, como as *tags* que representam o armário. Dessa forma, todas as demais *tags* representam as gavetas desse armário.

Nessa mobília as gavetas estão organizadas em grupos identificados pela *tag* de abertura <entrada> (linha 5) e pela *tag* de fechamento </entrada> (linha 19). Por sua vez, dentro de cada grupo existem 12 gavetas representadas pelas *tags* que armazenam os dados referentes a uma das 202 perguntas do QSL, que foram feitas a um determinado tipo de informante e em um determinado município da região Norte do Brasil (linhas 6 a 19). Vale reiterar que os dados dessas 12 *tags* estão organizados de maneira lexicográfica. Assim, ao traçar um paralelo com a Lexicografia Impressa, que se utiliza das fichas lexicográficas para armazenar dados referentes a cada candidato a verbete, podemos denominar esse conjunto de dados como uma espécie de ficha lexicográfica eletrônica, dada as suas características como ilustrado no quadro a seguir:

Quadro 7: Organização lexicográfica dos dados dialetais.

Elemento	Descrição
1) lema	Resposta do informante como denominação de determinado referente/conceito.
2) pergunta	Número e pergunta do QSL.
3) exemplo	Fala do informante que ilustra o uso de uma UL em uma situação real de comunicação.
4) observação	Informações adicionais relacionadas ao momento de interação entre entrevistador e entrevistado.
5) fonética	Registro da variação fonética do referente.
6) áudio	Execução do áudio e tempo de gravação.
7) remissiva	Variação das respostas para o mesmo referente.
8) informante	Idade, sexo e escolaridade.
9) legenda dialetal	Localidade em que os dados foram coletados.
10) classe gramatical	Indicação gramatical do referente.
11) definição	Texto da definição.
12) mapa	Legenda dialetal

Fonte: Elaboração do autor.

Cada um desses 12 elementos, descritos no quadro 7, figuram no *XML* por meio de suas respectivas *tags* que permitem a identificação e recuperação das informações no banco de dados. Para melhor compreensão do funcionamento das *tags* se faz necessário explicitar que esses mecanismos de indexação de dados assumem dois formatos distintos, a saber: i) *tags* do tipo *elemento*: armazenam dados textuais sem nenhum tipo de restrição em relação à quantidade de caracteres; ii) *tags* do tipo *atributo*: armazenam dados com finalidades distintas pois, geralmente, são utilizadas para registrar

especificidades relacionadas a um conjunto textual que está inserido em uma *tag* do tipo *elemento*. É importante frisar que não há uma regra para o uso das *tags* do tipo *elemento* ou do tipo *atributo*. O que ocorre, frequentemente, é uma combinação das duas modalidades de etiquetagem com vistas a construir um banco de dados que atenda aos propósitos de um determinado projeto.

No caso do protótipo do *VoDiNorte*, as *tags* foram estabelecidas de acordo com o tipo de dado armazenado e podem ser do tipo *elemento*, *atributo* ou uma mistura dos dois tipos. A partir das informações apresentadas na figura 26 especificamos, no quadro a seguir, o tipo e a função de cada *tag* do banco de dados da pesquisa:

Quadro 8: Descrição das *tags* do banco de dados em *XML*.

Tag	Descrição
<entrada id="**acid.geo.água.1**" abc="**i**">	*Tag* do tipo atributo com a indicação de dois tipos de informação, ou seja, um atributo para indicar uma identificação (id) que está expressa no texto escrito entre aspas, além de um atributo referente à ordem alfabética do lema em questão.
<lema>**igarapé**</lema>	*Tag* do tipo elemento que armazena a resposta do informante.
<perg campo="Acidentes geográficos" ref="QSL-1">**Como chama um rio pequeno, de uns dois 8 metros de largura?**</perg>	*Tag* mista com dois atributos sendo o primeiro destinado a indicar a área semântica a que pertence a pergunta do QSL e o outro utilizado para armazenar o número da pergunta do QSL. A pergunta do QSL está armazenada em formato de *elemento.*
<ex>**Aqui é garapé, né...(Tem outros nomes?) Não. Aqui é garapé só.**</ex>	*Tag* do tipo *elemento* utilizada para transcrever a fala do entrevistado.
<obs></obs>	*Tag* do tipo elemento reservada para registrar observações diversas referentes à entrevista.
<fone>**garapé**</fone>	*Tag* do tipo *elemento* para indicar as variações fonéticas dos entrevistados.
<aud src="**nome-do-arquivo**" type="**mp3**">**001_01_QFF01_ QSL051_A == 48:56**</aud>	*Tag* mista com dois atributos relacionados ao arquivo de áudio que é executado pela aplicação web. Além disso, na porção do tipo *elemento* há o nome do arquivo de áudio utilizado na transcrição, bem como a indicação do tempo em que aquela fala ocorre.
<ver name="**igarapé**" ref="**acid.geo.água.1**"/>	*Tag* do tipo atributo com a especificação do nome do verbete remissivo e sua respectiva identificação.
<info sexo="**M**" escolaridade="**F**" idade="**J**" > **28 anos**</info>	*Tag* do tipo mista com dados relacionados ao informante em formato de *atributo*. Informações adicionais como a idade do entrevistado foram adicionadas em forma de *elemento*.
<lg ponto="**1**" cidade="**Oiapoque**" estado="**AP**" />	*Tag* do tipo atributo com a indicação do número do ponto de inquérito, município e estado onde a entrevista foi realizada.
<gram></gram>	*Tag* do tipo elemento reservada para armazenar a classe gramatical do lema.
<def></def>	*Tag* do tipo elemento utilizada para escrever a definição

	lexicográfica.
<map src="" type="jpg"/>	*Tag* do tipo atributo que armazena o link que deve ser acessado pela ferramenta responsável por exibir a legenda dialetal do verbete por meio de uma representação cartográfica.

Fonte: Elaboração do autor.

É possível observar, a partir do quadro 8, que o conteúdo de cada *atributo* foi escrito entre aspas e que as informações textuais, presentes em *tags* do tipo *elemento*, foram escritas sem o uso desse sinal de pontuação. Essa característica é própria de cada tipo de *tag*, ou seja, uma *tag* do tipo *atributo* necessita das aspas para delimitar o dado que será armazenado, enquanto os dados textuais, presentes em *tags* do tipo *elemento* não exigem tal indicação. Vale destacar que a falta de uma das aspas em *tags* do tipo *atributo* acarreta um erro de formatação do *XML* que é detectado pelo *DTD*.

Tendo em vista a importância do *Document Tipe Definition* apresentamos, na figura a seguir, as regras que foram formuladas para o arquivo *XML:*

Figura 27: Regras escritas no *DTD*.

```
<!-- DTD para dicionarios -->
<!ELEMENT dicio (entrada+)>

<!ELEMENT entrada   (lema,perg,ex,obs,fone,aud,ver+,info,lg,gram,def,map)>
<!ATTLIST entrada
        id              ID              #REQUIRED
        abc             CDATA           #REQUIRED
>
<!ELEMENT lema      (#PCDATA)>
<!ELEMENT perg      (#PCDATA)>
<!ATTLIST perg
        campo  CDATA        #REQUIRED
        ref    CDATA        #REQUIRED
>
<!ELEMENT ex        (#PCDATA)>
<!ELEMENT obs (#PCDATA)>
<!ELEMENT fone (#PCDATA)>
<!ELEMENT aud (#PCDATA)>
<!ATTLIST aud
   src      CDATA     #REQUIRED
   type     (mp3|wav) #REQUIRED
>
<!ELEMENT ver       EMPTY>
<!ATTLIST ver
        name CDATA    #REQUIRED
        ref     IDREF    #REQUIRED
>
<!ELEMENT info (#PCDATA)>
```

```
<!ATTLIST info
   sexo      (M|F)          #REQUIRED
   idade     (J|I|X)        #REQUIRED
   escolaridade (F|U|X) #REQUIRED
>
<!ELEMENT lg          EMPTY>
<!ATTLIST lg
         ponto   CDATA          #REQUIRED
         cidade  CDATA          #REQUIRED
         estado  CDATA           #REQUIRED
>
<!ELEMENT gram      (#PCDATA)>
<!ELEMENT def       (#PCDATA)>
>
<!ELEMENT map EMPTY>
<!ATTLIST map
   src       CDATA      #REQUIRED
   type      (jpg|jpeg)    #REQUIRED
>
```

Fonte: Elaboração do autor.

É possível observar, a partir da figura 27, que umas das primeiras diretrizes estabelecidas no *DTD* é a configuração da *tag* <entrada></entrada>, que agrega outras 12 *tags* a saber: <lema></lema>, <per></perg>, <abo></abo>, <obs></obs>, <fone></fone>, <audio></audio>, <ver></ver>, <info></info>, <lg></lg>, <gam></gram>, <def></def>, <map</map>. Além do mais, é possível perceber que as regras delimitam o tipo da *tag* e sua estrutura. Assim, quando uma *tag* for do tipo *elemento*, será acompanhada do comando *!ELEMENT* e quando for do tipo *atributo*, terá o comando *!ATTLIST*. Destacamos, ainda, que após a indicação do tipo de *tag* as regras especificam quais dados serão armazenados em forma de *atributo* e quais dados serão inseridos em formato de texto. Em síntese, um documento *XML* em conformidade com o *DTD* garante uma estrutura sem falhas e, por sua vez, a compatibilidade do banco de dados com outras ferramentas computacionais.

Destacamos, também, que durante o planejamento da estrutura do banco de dados uma subclassificação foi adicionada à *tag* entrada de cada candidato a lema do protótipo do *VoDiNorte*. Essa categorização se dá por meio da identificação (*id*) composta por três elementos, a saber: i) indicação da área semântica estabelecida pelo Projeto ALiB; ii) acréscimo de uma ou mais subáreas; iii) identificação numérica. Essa organização está visualizada no quadro a seguir:

Quadro 9: Subclassificação do *corpus* de pesquisa.

Id	Subárea	Entrada
1. Acidentes geográficos acid.geo.agua.1	água	córrego, pinguela, foz, redemoinho, onda de mar, onda de rio
2. Fenômenos atmosféricos fen.atm.nuvem.10	vento	redemoinho
	nuvem	relâmpago, raio, trovão
	chuva	temporal, tempestade, tromba d'água, chuva forte, chuva de pedra, estiar, arco-íris, garoa
	terra	terra úmida
	ar	sereno, neblina
3. Astros e tempo ast.temp.tempo.22	astros	estrela matutina, estrela vespertina, estrela cadente, correr uma estrela, via láctea
	tempo	amanhecer, nascer do sol, pôr do sol, alvorada, crepúsculo, entardecer, anoitecer, meses do ano, meses com nomes especiais, ontem, anteontem, transanteontem
4. Atividades agropastoris ativ.agro.alimentos.39	alimentos	tangerina, castanha, amendoim, camomila, penca, banana dupla, coração, espiga, sabugo, touceira, girassol, vagem do feijão, macaxeira, mandioca
	objetos	carinho de mão, hastes do carrinho de mão, cangalha, canga, balaio, bruaca
	animais	borrego, perda da cria
	terra	trabalhador de enxada em terra alheia, picada, trilho
5. Fauna fauna.faun.aves.64	aves	urubu, colibri, joão-de-barro, galinha-d'angola, papagaio, sura
	mamíferos	cotó, gambá, patas dianteiras do cavalo, crina do pescoço, crina da calda, lombo, anca, chifre, boi sem chifre, cabra sem chifre, úbere, rabo, manco
	insetos	mosca varejeira, libélula, pernilongo
	larvas	bicho de fruta, coró, sanguessuga
6. Corpo humano corpo.hum.cabeça.93	cabeça	pálpebras, cisco, cego de um olho, vesgo, míope, terçol, conjuntivite, catarata, dentes caninos, dentes do siso, dentes molares, desdentado, fanho, meleca, soluço, nuca, pomo de adão,
	tronco	clavícula, corcunda, axila, cheiro nas axilas, seios, vomitar, útero
	membros	canhoto, perneta, manco, pernas arqueadas, rótula, tornozelo, calcanhar, cócegas

7. Ciclos da vida ciclo.vida.mulher.129	mulher	menstruação, menopausa, parteira, dar à luz, gêmeos, aborto, abortar
	parentesco	mãe de leite, irmão de leite, filho adotivo, filho mais moço, menino, menina, madrasta
	morte	finado
8. Convívio e comportamento social conv.comp.adjetivos.150	adjetivos	tagarela, pessoa pouco inteligente, sovina, mau pagador, assassino pago, marido enganado, prostituta, xará
	bebida	bêbado
	cigarro	cigarro de palha, toco de cigarro
9. Religião e crenças rel.cren.entidades.170	entidades	diabo, fantasma
	feitos	feitiço
	pessoas	benzedeira, curandeiro
	objetos	amuleto, medalha, presépio
10. Jogos e diversões infantis jog.inf.grupo.184	individual	cambalhota, estilingue, pipa, balanço
	grupo	bolinha de gude, esconde-esconde, cabra-cega, pega-pega, pique, lenço-atrás, gangorra, amarelinha
11. Habitação hab.hab.cozinha.199	paredes	tramela, veneziana, interruptor
	banheiro	vaso sanitário
	cozinha	fuligem, borralho
	objetos	isqueiro, lanterna
12. Alimentação e cozinha alim.coz.alimentos.205	matinal	café da manhã
	alimentos	geleia, carne moída, curau, canjica, mungunzá, bala, pão francês, pão bengala
	bebida	aguardente
	pessoas	empanturrado, glutão
13. Vestuário e acessórios vest.aces.mulher.226	mulher	sutiã, calcinha, rouge, grampo, tiara
	homem	cueca
14. Vida urbana vida.urb.ruas.232	ruas	sinaleiro, lombada, calçada, meio-fio, rotatória
	transporte	ônibus urbano, ônibus interurbano
	localidade	lote
	estabelecimentos	bar

Fonte: Elaboração do autor.

No quadro 9 é possível observar o procedimento de nomeação de cada *Id*. Assim, a entrada *redemoinho*, por exemplo, é nomeada pela *Id fen.atm.vento.9*, no qual os caracteres *fen* e *atm* indicam que a entrada pertence à área semântica dos *Fenômenos Atmosféricos* e o caractere *vento* sinaliza que ela está alocada nesta subárea. Já o caractere *9* é responsável pela identificação numérica que não se repete no *corpus*, garantindo que cada *Id* seja única no banco de dados.

A organização dessa ontologia permite acessar dados por meio das 14 áreas semânticas que formam parte do QSL-ALiB. Além disso, uma segunda camada de filtragem foi adicionada aos candidatos a verbetes do protótipo do *VoDiNorte* que é acionada por meio das subáreas ilustradas no quadro 9. Na prática essa subclassificação auxilia o usuário a fazer uma busca específica na nomenclatura do vocabulário por meio de uma área e subárea. Por exemplo, digamos que um usuário queira visualizar todos os verbetes sobre aves. Para tanto, o consulente deverá acessar a ferramenta que filtra os dados por meio das áreas semânticas, clicar em *Fauna* e, posteriormente, clicar em *Aves* para que seja exibida todas as entradas com nomes de aves do vocabulário.

Após a realização de testes em relação à estrutura do *XML*, alguns ajustes ainda foram executados para, finalmente, proceder com a alimentação dos dados em definitivo. Como mencionado anteriormente, esse procedimento é bastante custoso e moroso, pois é o momento de ouvir a gravação de áudio e transcrever a fala do entrevistador e do entrevistado nas *tags* do banco de dados.

No entanto, não é preciso ter todas as 72 entrevistas transcritas no banco de dados para avançar nas etapas metodológicas da pesquisa. Desse modo, o desenvolvimento das primeiras ferramentas computacionais de análise e recuperação de informações no banco de dados foram produzidas a partir de uma amostra de dados transcritos no *XML*, já que a transcrição completa dos dados levou vários meses para ser concluída.

4.3. A construção de ferramentas computacionais para a recuperação de dados

Para poder acessar as informações armazenadas em cada uma das milhares de *tags* do banco de dados é preciso instruir o computador, em uma linguagem específica, para que o dado seja recuperado e exibido na tela. Para tanto, foi preciso se familiarizar com o funcionamento do software *BaseX*[67], bem como aprender sobre como utilizar a linguagem de consulta *X-Query*.

O *BaseX*[68] é um software que permite, entre outras coisas, visualizar informações específicas em arquivos *XML* por meio da escrita de expressões *X-Query*,

[67] Software gratuito desenvolvido por Chistian Grün. Para mais informações acesse: <https://basex.org>. Acesso em 10 mar. 2023.

[68] Além de ser uma ferramenta capaz de manipular bancos de dados em *XML* o *BaseX* também pode ser entendido como um *framework*, ou seja, uma plataforma que permite reunir linguagens diferentes para o

que podem ser compreendidas como um conjunto de comandos executados no editor do *BaseX* com a função de recuperar dados nas *tags* do *XML*.

É importante destacar que à primeira vista, o *BaseX* nos pareceu ininteligível e progredir com a manipulação do *XML* no software levou vários meses. O mesmo processo ocorreu com as expressões *X-Query* que exigiram várias pesquisas e testes para a compreensão dos comandos elementares. No entanto, todo o esforço foi recompensado e, atualmente, é possível realizar manipulações de dados de modo satisfatório a partir das possibilidades que o *BaseX* oferece.

As manipulações de dados realizadas por meio do *BaseX* podem ser consideradas como pequenos programas ou, como denominamos no âmbito desta Tese, ferramentas computacionais, pois elas são escritas para executar uma tarefa específica em um ambiente informatizado. Podemos chamar de pequenos programas porque o nível de complexidade é baixo, tendo em vista que, no mundo da programação, um software pode ter duas, três ou até centenas de linhas de códigos[69] em sua composição.

Desse modo, a construção das ferramentas computacionais que recuperam informações específicas no banco de dados é feita de acordo com a demanda da pesquisa, ou seja, para cada recuperação de dados, uma ferramenta é escrita no editor do *BaseX*. Em suma, é possível formular perguntas para o *corpus* que serão respondidas por meio da escrita de um pequeno programa, que irá percorrer o banco de dados, selecionar itens específicos e mostrar na tela conforme as orientações do usuário. A seguir, descrevemos o funcionamento das expressões *X-Query* a partir da escrita de algumas ferramentas computacionais.

4.3.1. Localizando uma unidade lexical específica

Para realizar a busca de uma UL no banco de dados é preciso delimitar os critérios da pesquisa por meio de uma expressão *X-Query*[70] que deve ser escrita no editor do *BaseX*. Normalmente, a *X-Query* é formada por três linhas de código e a

---

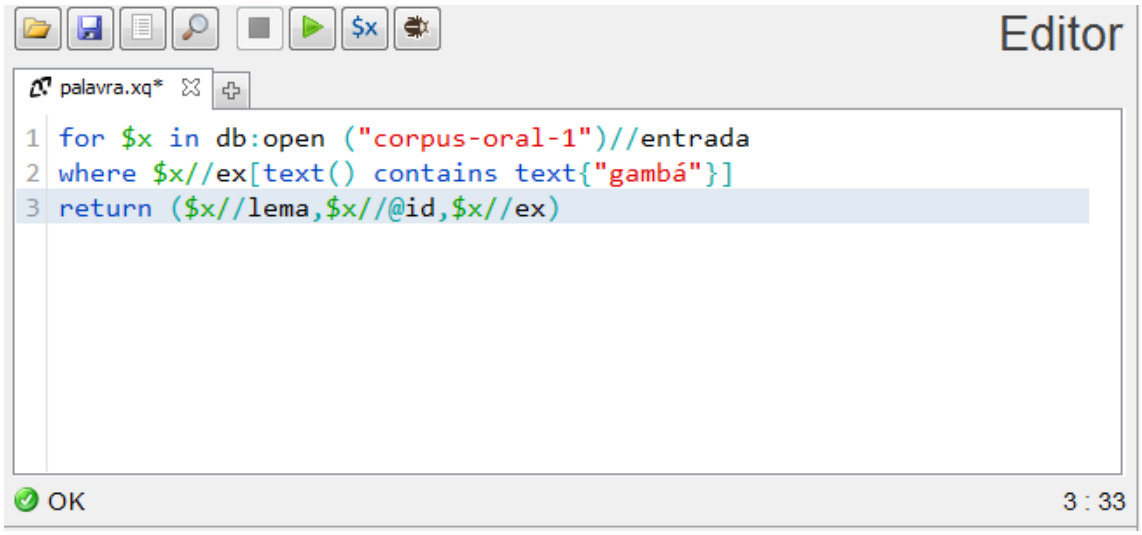
desenvolvimento de aplicações web. Para maiores informações consultar <https://docs.basex.org/wiki/Main_Page>. Acesso em 16 abr. 2023.

[69] Comandos dados ao computador para executar tarefas específicas. As linhas de código são utilizadas nas linguagens de programação, de marcação e de consulta.

[70] As exemplificações e explicações a respeito dessa linguagem de consulta visam a fornecer uma ideia de como é possível desenvolver as ferramentas de recuperação de informação no banco de dados. Caso o pesquisador queira criar as suas próprias ferramentas poderá se apoiar nesses exemplos para se aprofundar, posteriormente, no uso das expressões X-Query.

solicitação é processada ao clicar no botão *run*, que corresponde ao ícone de seta verde. Os resultados da solicitação são exibidos em uma janela abaixo do editor. A figura a seguir ilustra a composição de uma *X-Query*:

Figura 28: Expressão *X-Query* para a recuperação da UL *gambá*.



Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

É possível observar, por meio da figura 28, que o editor do *BaseX* possui algumas ferramentas que são úteis para realizar a recuperação de dados do *corpus* de pesquisa como, por exemplo, a possibilidade de salvar a expressão *X-Query* que foi criada em um arquivo com extensão.*xq*, para poder usá-la em uma futura recuperação de dados.

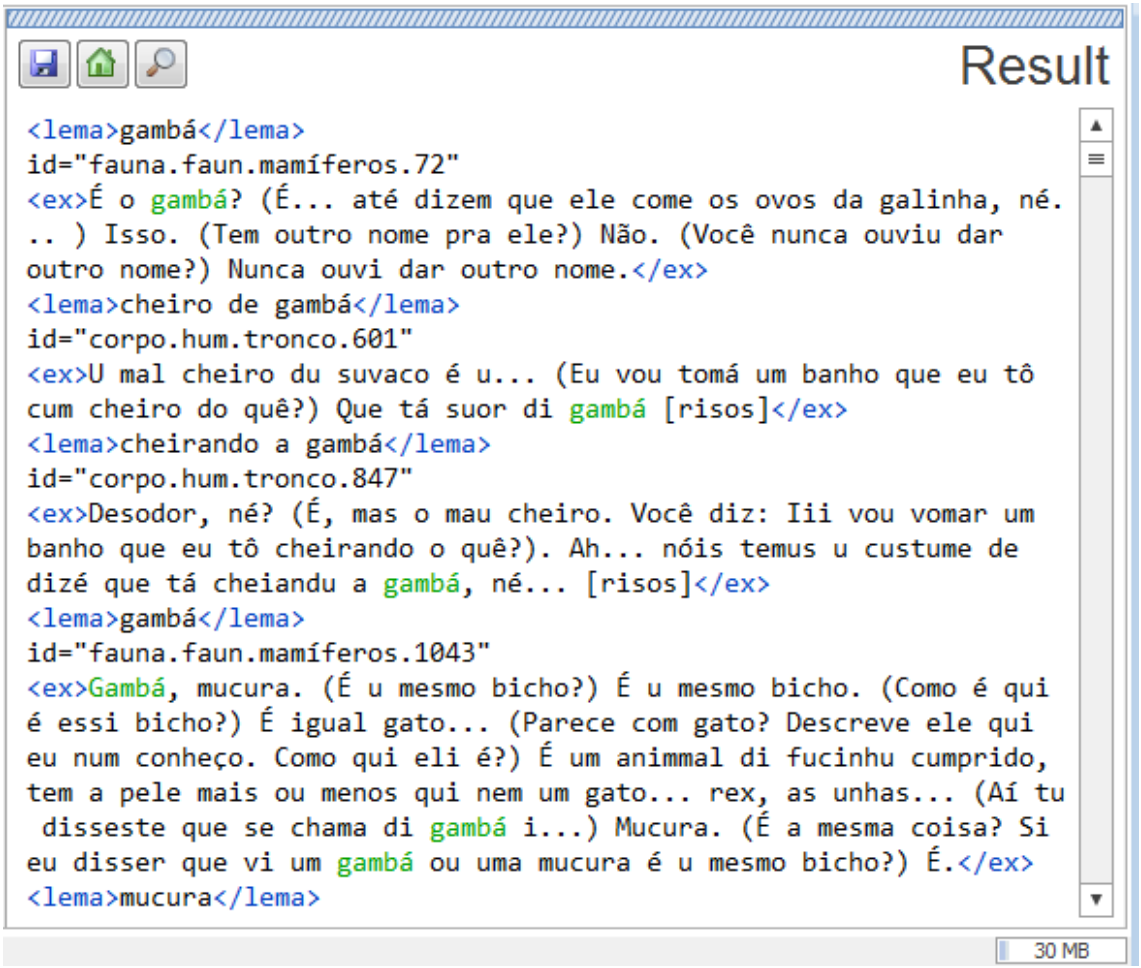
Assim, para que a ferramenta possa buscar no banco de dados a UL *gambá* e mostrar os resultados em uma segunda janela, três linhas de código foram escritas, a saber:

Linha 1: Iniciada pelo comando *for* indica que foi criada a variável *$x* para armazenar a informação a ser extraída do banco de dados *corpus-oral-1*. A busca abrange todas as entradas do banco de dados a partir do comando *//entrada*;

Linha 2: O comando *where* 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* <ex></ex> (abreviação de exemplo);

Linha 3: O comando *return* delimita o resultado a ser exibido na segunda janela. Nesse caso, o retorno deverá conter a identificação da entrada, ou seja, a *id*, o *lema* e o contexto da fala (*ex*), conforme pode ser visto na figura a seguir:

Figura 29: Resultado da expressão *X-Query* para a recuperação da UL *gambá*.



Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

Os dados exibidos na figura 29 representam apenas as primeiras linhas da recuperação dos dados que foi executada, tendo em vista que não é possível exibir todas as informações por meio da figura 28, cabendo ao usuário rolar com a roda do mouse para baixo a fim de visualizar os demais resultados. Ressaltamos, ainda, que as informações são organizadas segundo a ordem estabelecida pelo comando escrito na linha 3da *X-Query*, apresentada na figura 28. Desse modo, temos um conjunto de dados que seguem a seguinte ordem: lema, id (identificação) e ex (exemplo) em que é possível observar que a UL pesquisada aparece destacada na cor verde.

É possível comparar a escrita de uma *X-Query*, assim como o próprio ato de programar, ao uso de bloquinhos de montar como, por exemplo, o brinquedo *Lego* em que as peças vão se encaixando para formar objetos. Nessa atividade criadora, muito comum entre as crianças, algumas partes já montadas são reaproveitadas e dão

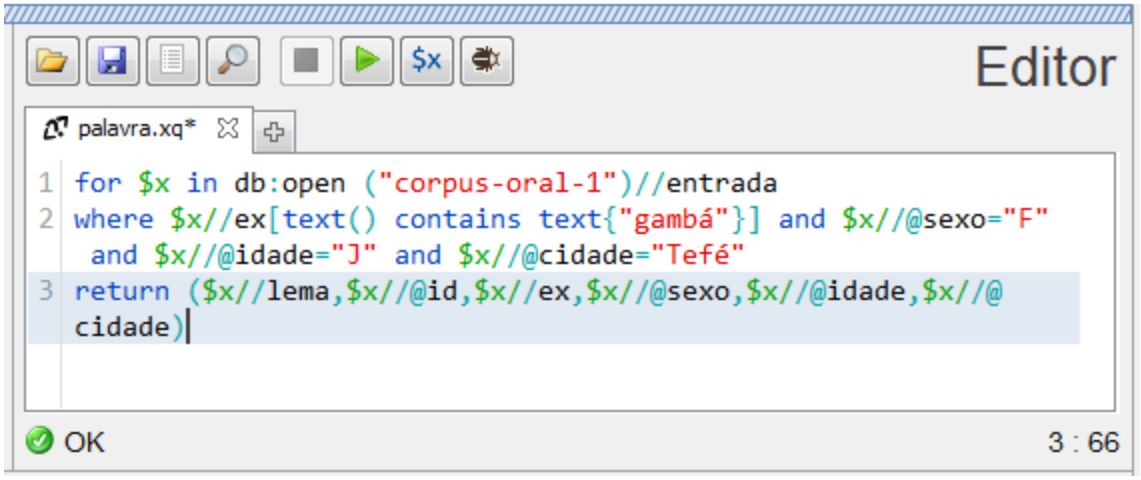
origem a outros objetos. Igualmente acontece com a escrita das expressões *X-Query* e é por isso que salvar cada expressão na pasta disponibilizada no editor do *BaseX* é importante, já que é possível reaproveitar pedaços de uma *X-Query* na composição de uma nova ferramenta de pesquisa.

No caso dos exemplos aqui expostos, a linha 1 não irá mudar a cada nova ferramenta construída, pois estamos realizando buscas em um único arquivo *XML*. Isso significa que devemos editar apenas as linhas 2 e 3 de acordo com o tipo de informação a ser recuperada, conforme ilustrado na próxima recuperação de informações no *corpus* de pesquisa.

4.3.2. Recuperação de dados filtrados por meio das variáveis sexo, idade e localidade.

Tendo em vista que a primeira linha da *X-Query* permanecerá a mesma, podemos utilizar as linhas da expressão anterior e substituir e/ou acrescentar apenas os dados referente ao sexo, à idade e à localidade na linha 2, além de reconfigurar o resultado da pesquisa na linha 3, como é possível constatar na figura a seguir:

Figura 30: Expressão *X-Query* para a recuperação de UL com controle das variáveis *sexo*, *idade* e *localidade*.

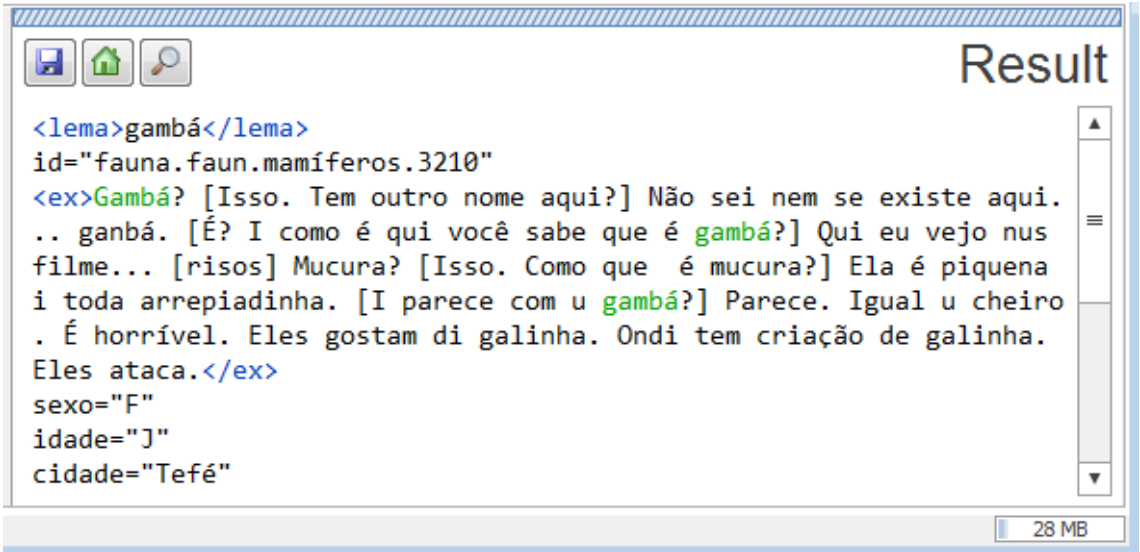


Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

Observamos, na figura 30, que foram adicionadas à linha 2 do editor três dados a serem recuperados pela *X-Query*, especificados pelos atributos *@sexo*, *@idade* e *@cidade* e que fazem parte do comando *where* e estão ligados entre si por

meio do comando *and*. Desse modo, essa solicitação irá filtrar a UL *gambá*, mencionada por informantes do sexo feminino, que sejam jovens e moradores do município de Tefé. O resultado, na linha 3, está configurado para exibir o lema, a id, o exemplo, o sexo, a idade e a cidade. Vale lembrar que esse resultado pode ser configurado para obter qualquer dado que venha a compor a microestrutura do protótipo do *VoDiNorte*. No caso da solicitação expressa na linha 2, poderíamos indicar que a ferramenta exibisse apenas o exemplo, a fim de verificar se houve ou não a ocorrência da UL pesquisada. No entanto, para confirmar a solicitação feita na linha 2 e deixar a exemplificação dessa *X-Query* mais didática, solicitamos para que o software também mostre os dados referente ao *lema*, *sexo*, *idade* e *cidade*, com ilustrado na figura a seguir:

Figura 31: Resultado da expressão *X-Query* para a recuperação de UL com controle das variáveis *sexo*, *idade* e *localidade.*



Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

Como é possível observar, por meio da figura 31, o resultado da pesquisa retornou o lema *gambá*, seguido por sua identificação no banco de dados. Em seguida temos a exibição do exemplo com a UL *gambá* destacada em verde, já que esse foi o elemento central da busca. Para complementar esse conjunto de dados temos a indicação do sexo (F), da idade (J) e da cidade (Tefé).

Como mencionado anteriormente, os dados do ALiB referentes às localidades do interior reúnem informantes que possuem o nível fundamental de escolaridade e,

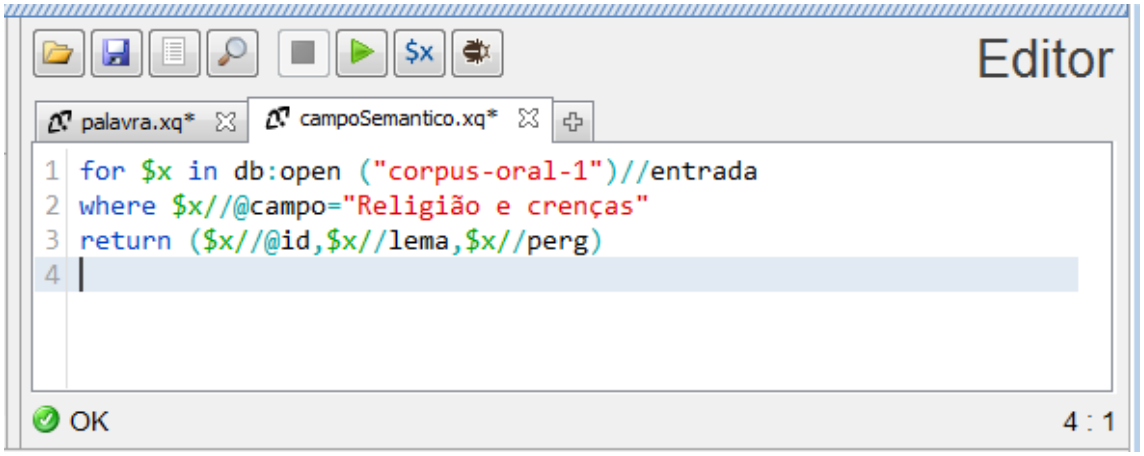
por essa razão, não há a necessidade de filtrar os dados usando a variável escolaridade. Porém, caso fosse necessário a recuperação desse tipo de dado em uma situação em que se esteja trabalhando também com as capitais brasileiras, um comando *and*, na linha 2 do editor, deverá ser inserido juntamente com a especificação da *tag* que armazena esse tipo de dado, para que a expressão *X-Query* processe os resultados levando em consideração também o nível de escolaridade.

4.3.3. Seleção de informações a partir de uma área semântica

É possível realizar buscas no banco de dados em *XML* também a partir de uma área semântica. Isso porque a estrutura do arquivo *XML* foi construída para oferecer essa funcionalidade a partir das categorias e subcategorias, que foram adicionadas em cada *id* conforme exposto no item *4.2. A construção do primeiro banco de dados em XML*.

Desse modo, para construir uma ferramenta que recupere dados relacionados a uma das 14 áreas semânticas do *corpus* de estudo é preciso especificar a área semântica e indicar o tipo de informação que se deseja visualizar na janela dos resultados. No exemplo a seguir, solicitamos a visualização de informações da área semântica *Religião e crenças*:

Figura 32: Expressão *X-Query* para a recuperação de informações a partir da área semântica *Religião e crenças*.



Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

Na figura 32, observamos que os dados requeridos são aqueles que apresentam o

atributo *@campo= "Religião e crenças"* que é especificado na linha 2 do editor. O resultado dessa solicitação está configurada, na linha 3 da expressão *X-Query*, para exibir a *id*, o *lema* e a *pergunta*. Esses elementos podem ser visualizados na figura a seguir:

Figura 33: Resultado da expressão *X-Query* para a recuperação de informações a partir da área semântica *Religião e crenças*.

Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

É possível observar, por meio da figura 33, que os resultados estão organizados conforme a ordem expressa na linha 3 da figura 32. Assim, a partir de cada *id* recuperada é possível visualizar a área e a subárea, ou seja, podemos verificar que o primeiro resultado recuperado pertence ao grupo *Religião e crenças* que está representado por meio dos caracteres *rel.cren* presentes na *id* da entrada, além de apresentar uma subcategoria nomeada de *entidades*. Desse modo, os lemas *diabo, lúcifer, satanás, visagem* e *visura* estão etiquetados como sendo do subgrupo das *entidades*, enquanto o lema *despacho* se enquadra no subgrupo dos *feitos*. Reiteramos que essa organização das *tags id* em forma de áreas e subáreas é a

responsável por tornar possível o funcionamento[71] da pesquisa por uma área semântica, no perfil de usuário intermediário do protótipo do *VoDiNorte*, como apresentado no item *5.3. Usuário intermediário*.

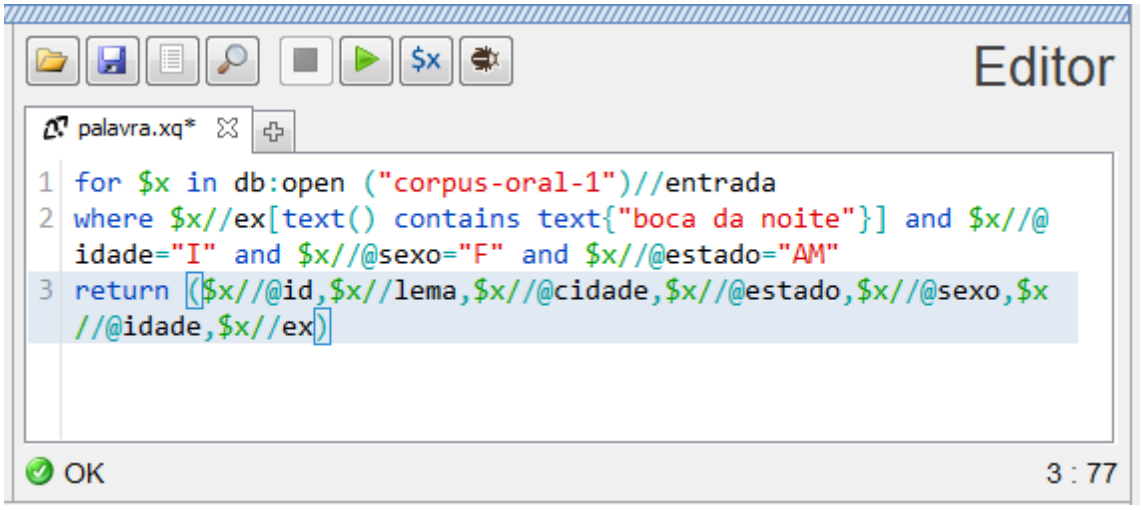
Acrescentamos, ainda, que cada recuperação de dados pode exibir qualquer informação presente no *corpus* em *XML* bastando, para isso, adicionar os elementos que se deseja visualizar na linha 3 da expressão *X- Query*.

Vale frisar que essa manipulação de dados pode ser alterada, na linha de código 2, para exibir os resultados de outras áreas semânticas do banco de dados a partir da substituição da especificação "Religião e crenças" por uma das outras 13 áreas semânticas que formam o QSL do Projeto ALiB.

4.3.4. Acrescentando mais filtros na recuperação dos dados

Semelhantemente a expressão *X-Query* exibida na figura 30, para adicionar mais filtros em uma recuperação de informação no banco de dados em *XML* é preciso agrupar as especificações das *tags* por meio do comando *and*, na segunda linha de código, conforme ilustrado na imagem a seguir:

Figura 34: Exemplo de uso do comando *and* na filtragem de dados para a UL *boca da noite*.

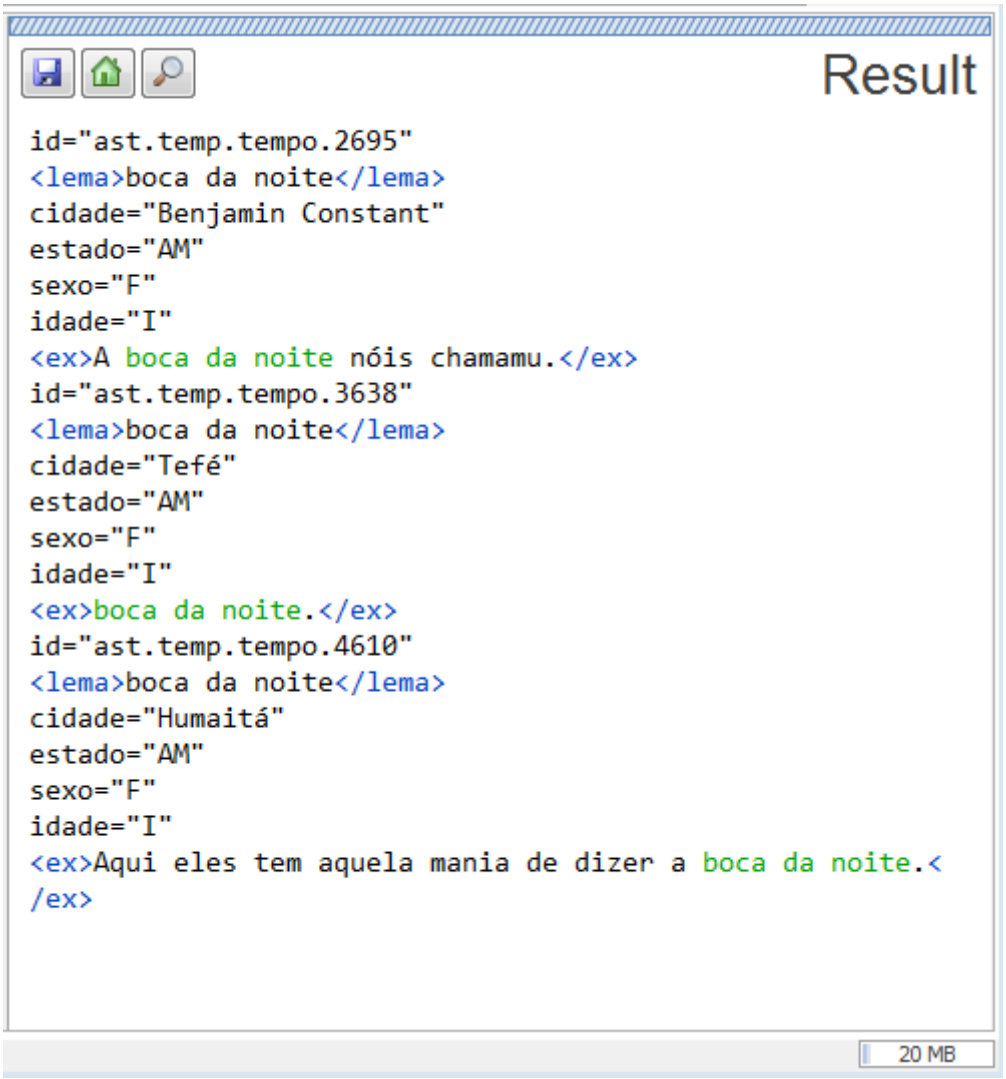


Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

[71] A ferramenta *pesquisar por uma área semântica* funciona por meio de um código *JavaScript* e *CSS* que seleciona os dados para a visualização do usuário acessando uma área e uma subárea da id de cada verbete.

Notamos, na figura 34, que a linha 2 do editor especifica que a informação requerida é a UL *boca da noite* e há, ainda, outras três especificações que são acrescentadas por meio do comando *and*, relacionadas à idade (I), ao sexo (F) e ao estado (AM). O resultado dessa pesquisa está configurada, na linha 3 do editor, para exibir a *id*, o *lema*, a *cidade*, o *estado*, o *sexo*, a *idade* e o *exemplo*, como mostra a figura a seguir:

Figura 35: Resultado do uso do comando *and* na filtragem de dados para a UL *boca da noite*.



Fonte: Dados do ALiB em *XML* e acessados por meio do software *BaseX*.

Como é possível observar, por meio da figura 35, os resultados exibidos foram poucos por se tratar de uma busca ao *corpus* de pesquisa bastante restrita. Assim, de

acordo com as especificações escritas na *X-Query* apresentada na figura 34, três grupos de dados foram recuperados para a UL *boca da noite* que foram mencionadas por informantes idosas, do sexo feminino e residentes nas cidades de Benjamin Constant/AM, Tefé/AM e Humaitá/AM.

Como foi possível constatar, os exemplos aqui apresentados têm a finalidade de mostrar a composição básica de uma expressão *X-Query*, além de ilustrar como é possível manipular alguns grupos de comandos para recuperar informações específicas em um banco de dados em *XML*.

Em suma, as ferramentas computacionais descritas nesta seção possuem dois propósitos básicos, a saber: i) mostrar aos pesquisadores interessados em desenvolver suas próprias ferramentas a possibilidade de recuperar dados de um *corpus* com o uso de *XML*, *BaseX* e *X-Query*; ii) apresentar e exemplificar que essas ferramentas são personalizadas e podem ser modificadas de acordo com os objetivos da pesquisa.

Outro ponto importante a ser destacado é a possibilidade de aplicar os conhecimentos adquiridos com a construção dessas ferramentas na elaboração de uma aplicação web, tendo em vista que o *BaseX* fornece os recursos necessários para escrever, dentro de seu editor, as linhas de código de um website que se comunica diretamente com a base de dados *XML*. Também destacamos que esses procedimentos serão melhor detalhados na seção *4.5. A construção da aplicação web*.

4.4. A construção do segundo banco de dados em *XML*

Criadas as várias ferramentas computacionais, com vistas a conceber um modelo de aplicação web para o protótipo do *VoDiNorte*, percebemos que os caminhos que estávamos trilhando nos obrigavam a realizar uma triagem das informações para compor um segundo banco de dados, destinado a compor os verbetes lexicográficos que recuperam dados estáticos ao usuário.

Essa necessidade foi percebida devido ao fato de que a ferramenta que exibe dados em formato de verbete precisar ler as informações armazenadas em um arquivo *XML* que não apresente multiplicidade de candidatos à entrada. Como no banco de dados os candidatos à entrada se repetem a cada nova entrevista transcrita, é preciso construir um segundo arquivo *XML* no qual não ocorra tal repetição, pois ela acarreta um erro no momento em que a ferramenta recupera dados sobre uma determinada entrada. Ocorre que, no momento de consulta a um determinado verbete, se o

computador encontrar duas entradas com o mesmo nome o conteúdo não será exibido, porque essa ferramenta trabalha com nomes de verbetes únicos e, assim, não pode exibir dois verbete com o mesmo nome, salvo os casos de homonímia em que um expoente numérico é adicionado para diferenciá-los. Desse modo, criamos o banco de dados 2 que contém uma seleção de informações devidamente organizadas para a exibição de dados em formato de verbete lexicográfico. Essa maneira de apresentação de dados ao usuário se caracteriza como uma obra lexicográfica polifuncional, em que as 12 *tags* que armazenam os dados lexicográficos ficam todas visíveis ao usuário.

Destacamos, ainda, que o banco de dados 2 também será utilizado para alimentar as ferramentas lexicográficas monofuncionais que, todavia, ainda não foram desenvolvidas e exibirão dados que atendam a uma única função lexicográfica.

Vale acrescentar que a estrutura do arquivo *XML*, bem como as regras do *DTD* utilizadas no banco de dados 1 foram reutilizadas no banco de dados 2. Desse modo, a única diferença entre os dois arquivos *XML* é que o primeiro armazena todas as transcrições referentes as 72 entrevistas baseadas nas 202 perguntas do QSL, enquanto o segundo apresenta uma seleção de dados organizados para que as ferramentas polifuncionais e monofuncionais funcionem adequadamente.

Essa modificação, em que dois bancos de dados acabaram sendo organizados, possibilitou um aprofundamento em relação aos conceitos de dados estáticos e dados dinâmicos que contribuiu para o desenvolvimento de uma ferramenta de pesquisa avançada que, alimentada pelo banco de dados 1, é capaz de recuperar dados dinâmicos por meio da configuração de filtros no protótipo do *VoDiNorte*.

No item a seguir, buscamos demonstrar como a aplicação web foi construída bem e como as ferramentas de pesquisa lexicográfica foram implementadas no protótipo do *VoDiNorte*. Salientamos, ainda, que houve o cuidado para não ocorrer aprofundamentos de questões técnicas e apresentar, de maneira objetiva, os conhecimentos no campo da Computação que foram mobilizados, a fim de guiar o pesquisador interessado em se aprofundar nessa temática.

4.5. A construção da aplicação web

Embora esta seja uma das últimas etapas metodológicas da Tese, a aplicação web foi prevista desde os primeiros meses da pesquisa e passou por várias reformulações até chegar ao estágio atual. Foram muitos os problemas de ordem técnica

que surgiram durante o processo do seu desenvolvimento e aperfeiçoamento e, todo esse trabalho que reuniu tentativas, erros e acertos, resultou em aprendizados importantes que serão aplicados em projetos futuros.

Ao observar a trajetória investida ao longo dos últimos quatro anos de pesquisa identificamos que muitas das melhorias implementadas na aplicação web ocorreram pela necessidade da mudança de paradigma em relação à Lexicografia, isto é, um processo de desvinculação do pensamento teórico-metodológico herdado pelo uso e estudo de dicionários impressos, para compreender como utilizar as tecnologias disruptivas na construção de ferramentas lexicográficas inovadoras.

Nesse sentido, um avanço importante que reorientou a forma de apresentação dos dados lexicográficos no protótipo do *VoDiNorte* está relacionado ao conceito de dicionário *polifuncional* e de dicionário *monofuncional*, que expressa a necessidade da objetividade no que tange ao atendimento das necessidades de consulta de consulentes variados e que apresentam demandas específicas de pesquisa, de modo que

> The original polyfunctional nature of the dictionaries affected access to and presentation of data. When they searched for words, users would be presented with full articles resembling those in printed dictionaries with data types intended to support a plurality of usage situations. This meant that users had to read, or at least skim, entire articles to find the answers they were looking for because the text contained relatively little relevant data among a large collection of data[72] (NIELSEN; FUERTES-OLIVERA, 2013, p. 335).

Dessa forma, o protótipo do *VoDiNorte* que, a princípio foi imaginado apenas como uma obra lexicográfica polifuncional configura-se, atualmente, em um conjunto de ferramentas capazes de exibir dados lexicográficos e dialetais de maneira polifuncional e monofuncional. Para tanto, a última reformulação da aplicação web reorganizou o modo de acesso às ferramentas de busca lexicográfica e abriu espaço para o desenvolvimento de ferramentas monofuncionais, que não haviam sido previstas anteriormente. Assim, o acesso ao protótipo do *VoDiNorte* pode ocorrer por meio de três perfis de usuários, a saber: i) usuário comum: representado por um público leigo e

[72] "A natureza polifuncional original dos dicionários afetou o acesso e a apresentação dos dados. Ao pesquisar palavras, os usuários eram apresentados a artigos completos semelhantes aos de dicionários impressos com tipos de dados destinados a suportar uma pluralidade de situações de uso. Isso significava que os usuários tinham que ler, ou pelo menos folhear, artigos inteiros para encontrar as respostas que procuravam, porque o texto continha relativamente poucos dados relevantes entre uma grande coleção de dados" (T. N.).

estudantes do Ensino Fundamental; ii) usuário intermediário: em que se encontram os estudantes do Ensino Médio e os estudantes do Ensino Superior; iii) usuário avançado: caracterizado pelo consulente especializado como, por exemplo, professores universitários e pesquisadores em geral interessados em estudar os dados dialetais dos municípios do interior da região Norte do Projeto ALiB.

Para atender as necessidades lexicográficas do usuário comum que, na maioria das vezes, busca por uma informação específica, ferramentas monofuncionais são utilizadas para a recuperação de dados específicos e mostram o conteúdo ao usuário de maneira rápida e de fácil compreensão, tendo em vista que esse é um usuário leigo. Assim, a partir das 12 *tags* que armazenam dados tratados lexicograficamente, é possível desmembrar as funcionalidades presentes na ferramenta de pesquisa polifuncional e desenvolver aplicações que atendam a uma única demanda de pesquisa, garantindo rapidez e objetividade em relação ao encontro da informação requerida pelo usuário. Assim, cada situação de pesquisa conta com uma ferramenta de busca que será acessada na área destinada ao usuário comum do protótipo, a saber: i) vocabulário de definições; ii) vocabulário de variação fonética; iii) vocabulário de exemplos em áudio; iv) vocabulário de exemplos em texto; v) vocabulário de remissivas; vi) vocabulário de legenda geolinguística.

O usuário intermediário, por sua vez, compreendido por estudantes de graduação e estudantes do ensino médio terão acesso ao modelo de verbete polifuncional, pois esse tipo de consulente pode se interessar por todas as informações lexicográficas presentes nas 12 *tags* que armazenam os dados referentes a uma determinada entrada. Todavia, uma gradação dessas informações foi aplicada para reduzir a quantidade de dados exibidos na tela ao clicar em uma entrada. Assim, a informação armazenada na *tag* relacionada à representação cartográfica será apresentada apenas se o usuário clicar no seu link de exibição, que abrirá um mapa com as informações referentes à legenda dialetal do verbete em outra aba do navegador.

Na área do usuário intermediário há, ainda, a possibilidade de acessar uma entrada de três formas distintas: i) por meio de uma lista com todas as entradas organizadas alfabeticamente; ii) com o auxílio de uma caixa de pesquisa em que o usuário deve escrever a entrada que deseja consultar e, caso faça parte da nomenclatura do vocabulário, o verbete será exibido; iii) a partir de uma das 14 áreas semânticas do QSL-ALiB.

Já o usuário avançado dispõe de uma ferramenta de pesquisa avançada capaz de realizar buscas diversificadas em todo o *corpus* da pesquisa. Trata-se do acesso a dados dinâmicos que são exibidos ao usuário mediante a seleção de filtros. Essa ferramenta de pesquisa lexicográfica busca romper com o modelo de Lexicografia tradicional na medida em que deixa de exibir dados estáticos ao usuário e passa a mostrar resultados dinâmicos que variam a cada nova configuração de filtros. Por ser capaz de recuperar partes de uma UL esse motor de busca é útil para visualizar unidades morfológicas e expressões formadas por um conjunto de dois ou mais itens lexicais nas entradas do protótipo.

No que diz respeito ao aspecto técnico, destacamos que a construção da aplicação web foi realizada na proporção em que a compreensão do funcionamento básico dos conteúdos relacionados à programação foi se consolidando. Assim, utilizamos um arquivo com extensão *.xqm* para armazenar todas as linhas de código que são executadas pelo editor do *BaseX* e, nesse particular, as expressões *X-Query* foram fundamentais para a construção das ferramentas de pesquisa lexicográfica. Utilizamos, ainda, a linguagem *HTML*[73] *(Hiper Text Markup Language)* para escrever as páginas do protótipo do *VoDiNorte*, além da linguagem *CSS*[74] *(Cascading Style Sheet)* para estilizar minimamente o website. Além disso, a linguagem *JavaScript*[75] foi usada para gerar o efeito de movimento da ferramenta de pesquisa lexicográfica que busca entradas por meio das áreas semânticas.

Essas linguagens de programação formam a base de sustentação do protótipo do *VoDiNorte*. A partir delas que o website existe e é por meio desses recursos que o pesquisador pode melhorar as funcionalidades e a aparência de uma aplicação web. Vale destacar, sobretudo para aqueles interessados em desenvolver uma aplicação web semelhante, que o arquivo com extensão .xqm, que armazena todas as linhas de comando da aplicação web, deve ser armazenado em uma pasta de arquivos específica do *BaseX*. Essa ação é necessária para que a ferramenta possa ser acessada por meio do navegador de Internet. Desse modo, é preciso acessar, no *Microsoft Windows*, a pasta que contém os arquivos de programas, localizar o diretório do *BaseX*, abrir a pasta *Webapp* e colar o arquivo *.xqm* nesta pasta. Vale destacar que existe um arquivo modelo

[73] Linguagem de marcação utilizada para construir páginas de websites ou de aplicativos web.
[74] Linguagem de marcação que adiciona um estilo ao conteúdo construído em HTML. Este estilo pode modificar substancialmente o aspecto visual de um website.
[75] Linguagem de marcação que permite acrescentar funcionalidades variadas em uma página HTML como, por exemplo, fazer com que uma lista de links se movimente ao clicar do mouse.

com extensão *.xqm* dentro da pasta *Webapp* que deve ser apagado, pois a aplicação web funciona com apenas um arquivo *xqm*, ou seja, manter ou dois acarretaria um erro de programação que impede o funcionamento do website.

Desse modo, uma vez que o arquivo com extensão *.xqm* esteja devidamente alocado na pasta *Webapp* é possível acessar o website por meio de um navegador de internet, pois o *BaseX* possui um servidor web com *local host*, ou seja, um recurso capaz de projetar a aplicação web que está sendo desenvolvida em um navegador de internet sem a necessidade do website estar hospedado em um servidor de internet.

Todavia, iniciar o *localhost*[76] para acessar o projeto pelo navegador de internet pode ser um pouco complexo, além de apresentar obstáculos de ordem técnica já que o funcionamento correto desse recurso depende da configuração do computador, do sistema operacional do computador e da versão do *BaseX*, bem como a versão do *Java*[77] instalados na máquina. Isso também significa que as maneiras de iniciar o *localhost* em um computador podem variar, o que obriga o pesquisador a estudar a documentação do *BaseX*, disponibilizado no site do software, no intuito de solucionar eventuais problemas.

Na experiência que tivemos, o *localhost* foi iniciado, no *Microsoft Windows 7*, por meio da linha de comando *basexhttp.bat* digitada diretamente no *Prompt de comando*. Após a mensagem de que o servidor foi iniciado (*HTTP Server was started <port: 8984>*) a aplicação web pode ser acessada digitando no navegador de internet o endereço *localhost:8984*. Vale destacar que a janela do *Prompt de comando* deve permanecer aberta enquanto o website é acessado por meio do *localhost*, pois caso o usuário feche o *Prompt de comando* o servidor local irá parar de funcionar e, consequentemente, a aplicação web ficará inacessível.

Em síntese, o desenvolvimento da aplicação web fecha um ciclo de trabalho que começou com a construção do banco de dados em *XML* e, posteriormente, com o desenvolvimento do arquivo *xqm.* Isso nos leva a destacar, novamente, a importância da construção desses arquivos para o funcionamento do protótipo do *VoDiNorte* podendo, até mesmo, relacionar o *XML* como sendo a espinha dorsal do vocabulário e o *xqm* como sendo o cérebro que comanda todas as funcionalidades da aplicação web.

---

[76] O *localhost* é iniciado por meio de orientações dadas ao computador no *Prompt de comando* do sistema operacional *Windows* ou no *Terminal* do sistema operacional *Linux*. Para maiores informações acesse: <https://docs.basex.org/wiki/Web_Application>.

[77] Software essencial para o funcionamento de variados tipos de aplicativos. Para maiores informações acesse: <https://www.java.com/pt-BR/>.

Como mencionado anteriormente, o protótipo do *VoDiNorte* se configura como um conjunto de ferramentas capazes de exibir dados dialetais tratados lexicograficamente. Algumas dessas ferramentas se encontram acessíveis e funcionais ao usuário e outras, todavia, ainda estão em fase de desenvolvimento e implementação. Desse modo, o usuário especializado dispõe de uma ferramenta de busca avançada pronta para o uso e o usuário intermediário possui o acesso a ferramentas que realizam a busca aos verbetes polifuncionais e que não estão totalmente finalizados. Já o usuário comum dispõe apenas da ferramenta que exibe a legenda dialetal dos verbetes *carapanã* e *jacinta* por meio de um mapa interativo, pois não houve tempo hábil para o desenvolvimento dos demais recursos computacionais destinados ao processamento de outros tipos de dados de maneira monofuncional.

Além disso, a aplicação web do protótipo do *VoDiNorte* necessita de um servidor para ser disponibilizado pela Internet. Como se trata de um projeto prototípico, estamos realizando testes, provisoriamente, com um servidor da *Universidad Paris 13*. Destacamos, ainda, que a principal vantagem de uma aplicação web é que a atualização, edição ou qualquer outro tipo de intervenção é realizada remotamente. Assim, do ponto de vista lexicográfico, cada vez que um verbete precisar ser editado ou cada vez que novas entradas necessitarem ser incorporadas à nomenclatura de um dicionário, bastará acessar a conta do servidor em que a aplicação web estiver hospedada e realizar as devidas edições. Essa vantagem também se reflete em economia financeira, já que, sem a necessidade de lançar volumes impressos, não há mais custos de impressão e reimpressão de novas edições lexicográficas com atualizações da obra.

CAPÍTULO 5 – APRESENTAÇÃO DO PROTÓTIPO

Este capítulo é destinado a uma apresentação do protótipo do *VoDiNorte*, apresentando as possibilidades de interação que o usuário pode realizar por meio da aplicação web, bem como apontando as tarefas que ainda deverão ser empreendidas futuramente. Nesse sentido, é importante frisar que uma mudança de paradigma se estabeleceu durante os estudos realizados no âmbito da Lexicografia Eletrônica e a prática metodológica aplicada ao protótipo do VoDiNorte de modo que, progressivamente, percebemos que planejar um dicionário com base na Lexicografia Impressa e veicular a obra em uma plataforma digital é bem diferente de planejar um dicionário com base na Lexicografia Eletrônica desde o início. Essa diferença se dá, principalmente, por conta das funções planejadas pelo lexicógrafo que demandam ferramentas computacionais que devem ser pensadas desde o estágio inicial do projeto.

No caso do protótipo do *VoDiNorte* essa mudança de paradigma propiciou duas escolhas importantes em relação ao tratamento dado ao *corpus* de estudo e a sua apresentação ao usuário, a saber: i) a distinção de ferramentas monofuncionais e polifuncionais; ii) adequar o acesso ao protótipo a partir da escolha de um perfil de usuário.

Destacamos, ainda, que no âmbito da Lexicografia Eletrônica um produto pode ser disponibilizado ao público mesmo que apresente um número de funcionalidades reduzidas, tendo em vista a possibilidade de atualização da base de dados e de melhora no desenvolvimento das ferramentas de pesquisa lexicográfica a qualquer momento. Assim, uma obra lexicográfica on-line pode estar em constante processo de aperfeiçoamento no intuito de disponibilizar um produto inovador e em constante atualização.

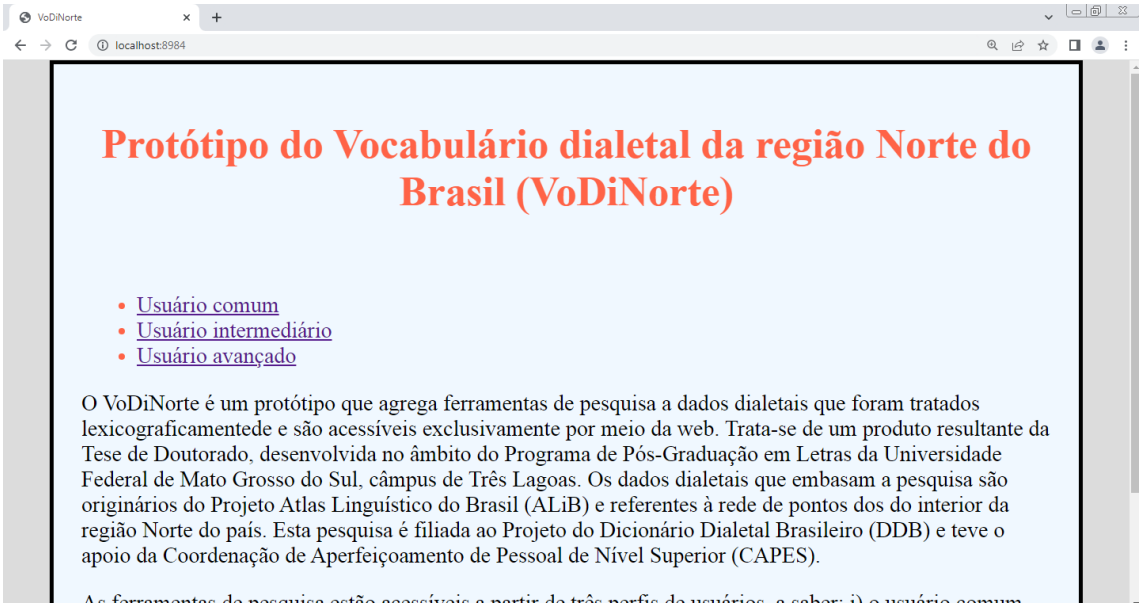
5.1. Orientações gerais

O protótipo do *VoDiNorte* foi desenvolvido por meio de recursos computacionais que funcionam a partir de um navegador de Internet. Essa visão só é possível a partir do computador de origem, tendo em vista que se trata de um acesso local. Porém, para que o protótipo do *VoDiNorte* fique disponível na rede mundial de computadores é preciso hospedá-lo em um servidor de internet. Por se tratar de um protótipo e por carecer de um serviço de hospedagem definitivo para abrigar e

disponibilizar o produto na web, o protótipo do *VoDiNorte* poderá ser acessado, provisoriamente, por meio de um servidor de internet da *Universidad Paris 13*, acessando o link <http://vodinorte.bombadil.fr/>.

Na página inicial do protótipo do *VoDiNorte* o usuário deverá selecionar um perfil que, por sua vez, dará acesso às ferramentas de pesquisa, conforme ilustrado na figura 36. Assim, três tipos de perfis estão disponíveis, a saber: i) usuário comum – leigo e estudantes do Ensino Fundamental; ii) usuário intermediário – leigo, estudantes do Ensino Médio e estudantes do Ensino Superior; iii) usuário avançado – professores universitários e pesquisadores em geral como, por exemplo, especialistas em Lexicografia, Dialetologia e demais áreas relacionadas aos estudos linguísticos.

Figura 36: Página inicial do *VoDiNorte*.



Fonte: Protótipo do *VoDiNorte*.

Desse modo, o usuário encontra dentro de cada perfil uma forma de apresentação dos dados lexicográficos apropriada ao seu nível de especialidade, ou seja, cada perfil busca atender as necessidades de pesquisa levando em conta o conhecimento prévio pressuposto para usuários comuns, intermediários e avançados.

Conforme mencionado no item *4.5. A construção da aplicação web*, inicialmente o projeto tinha em mente desenvolver um vocabulário dialetal polifuncional, isto é, um modelo de verbete que exibe de uma só vez todas as informações armazenadas no banco de dados para cada entrada, que é tipicamente encontrado nos dicionários impressos. Todavia, à medida que o estudo avançou

entendemos que essa forma de apresentação de dados não seria a ideal para o usuário de menor especialidade. Além disso, dispomos de recursos computacionais que são capazes de proporcionar um acesso mais direcionado aos usuários e, pensando nessa dinâmica, criamos os três perfis de usuários. Essa organização também nos proporcionou explorar melhor o tipo de funcionalidade que cada ferramenta deve oferecer a partir das particularidades de cada um dos três tipos de usuários.

Vale destacar que, inicialmente, imaginávamos que o protótipo desta Tese atenderia somente a um tipo de usuário considerado intermediário. Todavia, com o desenvolver da pesquisa enxergamos que o usuário intermediário seria o ponto de partida para adequar o projeto a alcançar também um público leigo e um público especializado.

Desse modo, a ferramenta destinada a executar pesquisas lexicográficas de modo polifuncional foi destinada ao usuário intermediário, pois, consideramos que os estudantes de graduação podem se interessar por dados polifuncionais. Destacamos que mesmo os estudantes do ensino médio, por meio de orientações que expliquem a composição desses verbetes, podem fazer um uso produtivo desse tipo de apresentação de dados.

O usuário comum, normalmente, possui uma demanda específica de consulta, ou seja, deseja pesquisar algo pontual em uma obra lexicográfica e, pensando que se trata de um público leigo, o verbete polifuncional não é ideal por exibir mais informação do que esse usuário precisa. Desse modo, as ferramentas polifuncionais desenvolvidas para o perfil de usuário intermediário foram desmembradas em várias ferramentas monofuncionais atendendo, assim, ao usuário comum com objetividade.

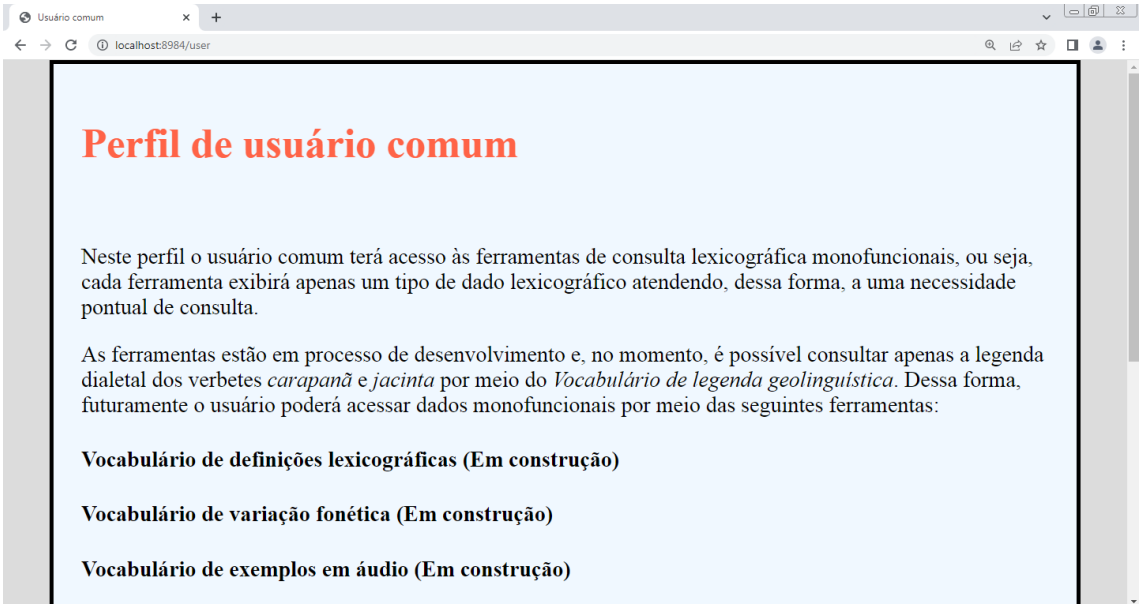
Por sua vez, o usuário avançado dispõe de um motor de busca sofisticado que realiza a recuperação de dados a partir de um conjunto de filtros. Além disso, ao contrário dos perfis anteriores que exibem dados estáticos ao usuário, no perfil avançado as informações são exibidas de maneira dinâmica.

A seguir, descrevemos com mais detalhes o funcionamento das ferramentas de consulta lexicográfica desenvolvidas para cada um dos três perfis de usuário do protótipo do *VoDiNorte*.

5.2. Usuário comum

A partir das informações armazenadas no banco de dados 2 é possível desenvolver ferramentas monofuncionais. Desse modo, neste momento[78], seis grupos de dados foram selecionados para compor seis tipos de vocabulários monofuncionais, a saber: i) vocabulário de definição lexicográfica; ii) vocabulário de variação fonética; iii) vocabulário de exemplos em áudio; iv) vocabulário de exemplos em texto; v) vocabulário de remissivas; vi) vocabulário de legenda geolinguística. A figura a seguir ilustra a página inicial do usuário comum.

Figura 37: Página inicial superior do usuário comum do *VoDiNorte*.



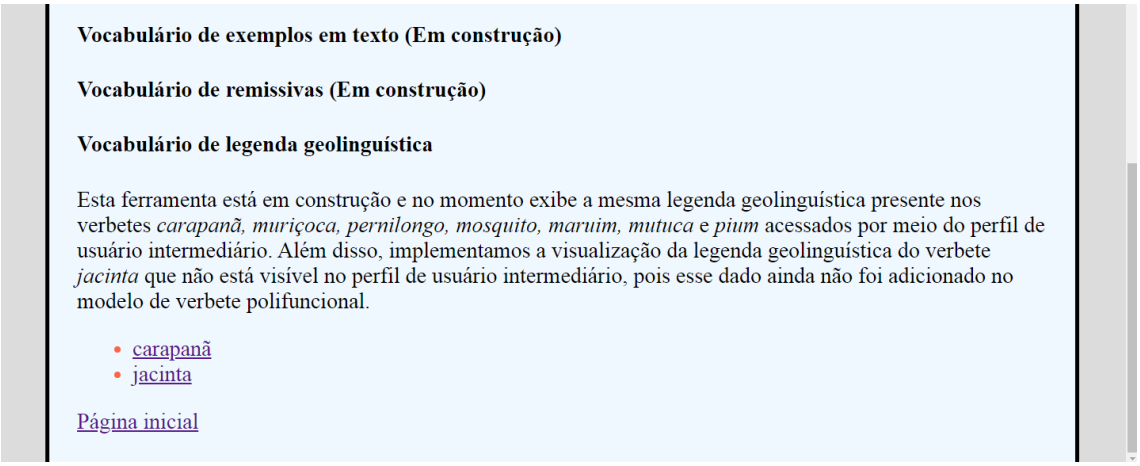
Fonte: Protótipo do *VoDiNorte*.

Na prática os seis tipos de vocabulários monofuncionais podem ser utilizados de maneira independente do conjunto da obra, ou seja, por meio do link direto que acessa cada uma dessas ferramentas o que permite, por exemplo, compartilhar com alunos da educação básica apenas a obra lexicográfica monofuncional desejada.

Destacamos, ainda, que o *Vocabulário de legenda geolinguística* está em desenvolvimento e, atualmente, é possível ter acesso ao mapa interativo relativo aos verbetes *carapanã* e *jacinta*, conforme ilustrado na figura a seguir.

[78] Outras ferramentas monofuncionais podem ser desenvolvidas futuramente.

Figura 38: Vocabulário de legenda geolinguística na página inicial inferior do usuário comum do *VoDiNorte*.

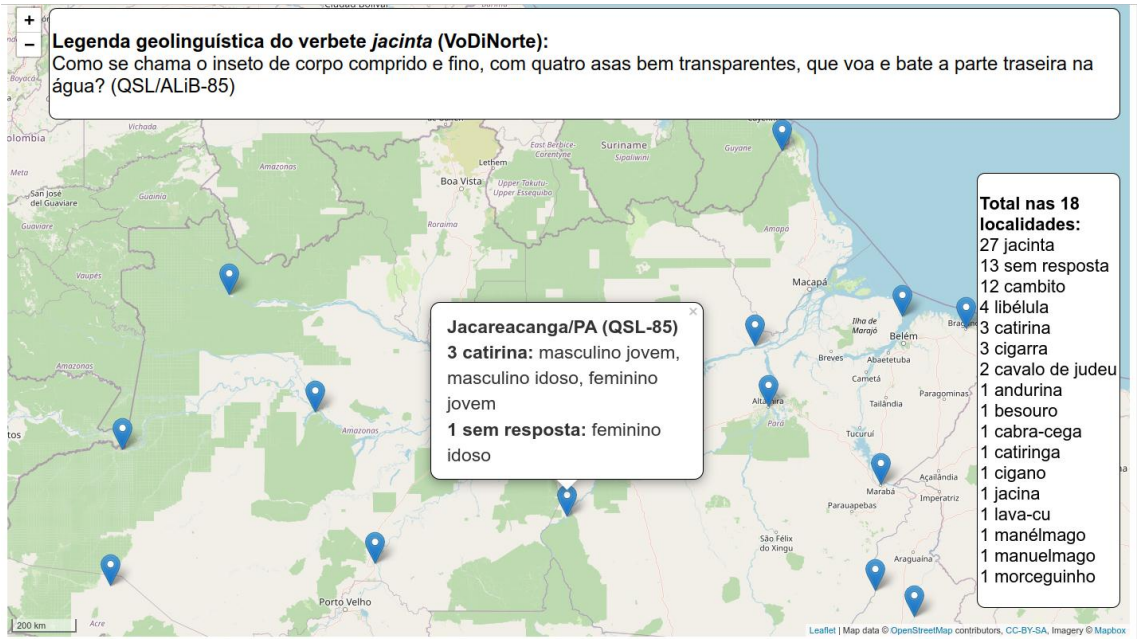


Fonte: Protótipo do *VoDiNorte*.

Como é possível atestar, a figura 38 exibe os verbetes *carapanã* e *jacinta* em formato de link. Desse modo, ao clicar em um deles uma nova aba do navegador se abre e um mapa interativo com os respectivos dados geolinguísticos é exibido. Vale destacar, que o mapa geolinguístico do verbete *carapanã* é o mesmo que está disponível no perfil de usuário intermediário. Dessa forma, não se faz necessário mostrá-lo nesta seção, tendo em vista que uma apresentação deste mapa é feita no item *5.3 Usuário intermediário*.

Todavia, o verbete *jacinta*, bem como os dados relacionados à sua legenda geolinguística ainda não foram implementados no modelo de verbete polifuncional de modo que, na figura a seguir, é possível visualizar a apresentação dessas informações.

Figura 39: Representação cartográfica do verbete *jacinta*.



Fonte: Protótipo do *VoDiNorte*.

É possível observar, por meio da figura 39, que as 18 localidades que formam a rede de pontos do ALiB, no interior da região Norte, estão sinalizadas com um alfinete azul. Ao clicar em um alfinete uma caixa de texto é exibida ao usuário com os dados relacionados às respostas dos quatro informantes para a pergunta 85 do QSL. Assim, é possível identificar que em Jacareacanga/PA três informantes (masculino jovem, masculino idoso, feminino jovem) mencionaram *catirina* como resposta para a pergunta 85 do QSL e uma informante (feminino idoso) não soube responder. Além disso, há uma caixa de texto no canto direto do mapa em que são exibidas todas as respostas que foram produzidas para a pergunta 85 do QSL, juntamente com a quantidade de vezes que foram mencionadas em todo o *corpus*.

Vale destacar que o mapa interativo utilizado para as representações cartográficas é um recurso computacional de código aberto e de distribuição livre[79] desenvolvido pela *Leaflet*[80] com a contribuição da *OpenStreetMap*[81]. A inserção de cada

[79] Softwares em que o código fonte é aberto, ou seja, o usuário pode baixar, estudar e modificar o código e distribuí-lo livremente, desde que se respeite as condições da licença. Para mais informações acesse a *Free Software Foundation*: <https://www.fsf.org/about/>. Acesso em: 16 abr. 2023.
[80] Biblioteca *javaScript* para o desenvolvimento de mapas interativos. Para mais informações acesse: <https://leafletjs.com/>. Acesso em: 16 abr. 2023.
[81] Plataforma colaborativa de dados geográficos livres e abertos, licenciada pela *Open Database Licence (ODbL)*. Para mais informações acesse: <http://openstreetmap.org/about>. Acesso em: 16 abr. 2023.

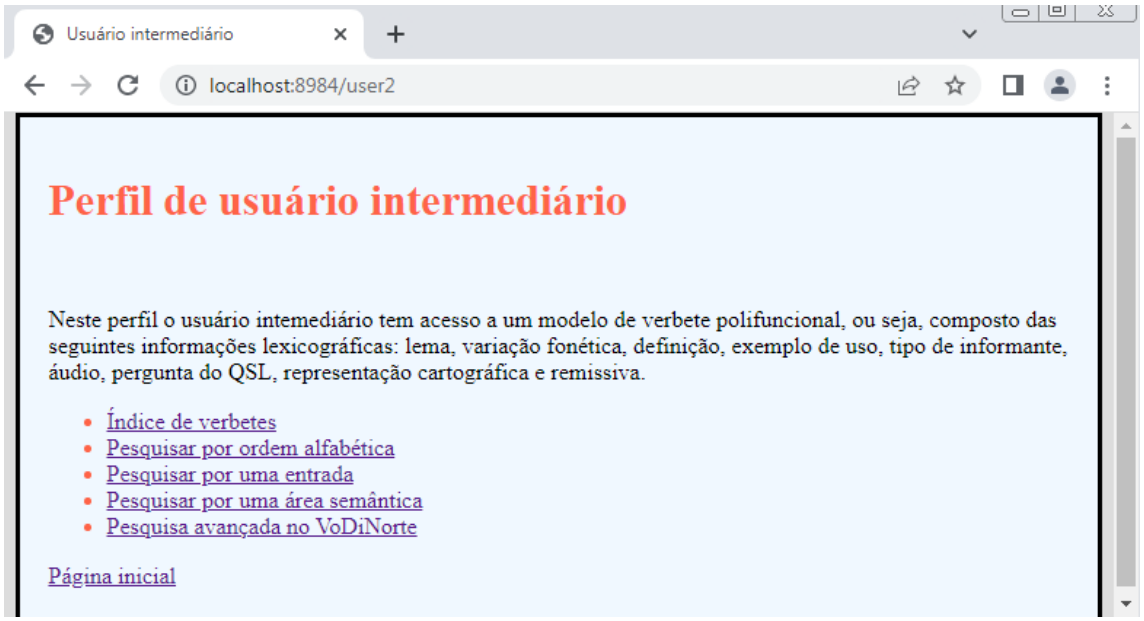
alfinete no mapa é feita por meio de coordenadas do Sistema de Posicionamento Global (*Global Positioning System - GPS*) e uma escala dinâmica em quilômetros está inserida no canto inferior esquerdo do mapa. Frisamos, ainda, que os mapas dinâmicos passarão por melhorias futuras, a fim de que possam atender aos critérios metodológicos do Instituto Brasileiro de Geografia e Estatística (IBGE). Desse modo, a apresentação desse tipo de mapa tem o objetivo de demonstrar a possibilidade que está ao alcance do pesquisador, sobretudo brasileiro, de desenvolver uma solução computacional capaz de representar um conjunto de dados em um mapa dinâmico e disponibilizar tais informações em um site na Internet.

Ressaltamos, ainda, que à exceção do *Vocabulário de legenda geolinguística*, os demais vocabulários monofuncionais, previstos para o usuário comum, estão em fase de desenvolvimento e no momento não estão acessíveis.

5.3. Usuário intermediário

Caracterizada como uma ferramenta de acesso a dados lexicográficos polifuncionais, o usuário intermediário poderá percorrer caminhos diferenciados até chegar a determinado verbete. Para tanto, a tela inicial desse perfil exibe cinco possibilidades de pesquisa: i) Índice de verbetes; ii) Pesquisar por ondem alfabética; iii) Pesquisar por uma entrada; iv) Pesquisar por uma área semântica; v) Pesquisa avançada no *VoDiNorte* conforme apresentado na figura a seguir:

Figura 40: Página inicial do usuário intermediário.



Fonte: Protótipo do *VoDiNorte*.

Dessa forma, ao clicar em *Índice de verbetes* o usuário será direcionado a uma lista, como mostra a Figura 41, que contém todas as entradas do vocabulário, organizadas em ordem alfabética, em que o acesso ao conteúdo do verbete ocorre ao clique do mouse em uma determinada entrada.

Figura 41: *Índice de verbetes* do *VoDiNorte*.

Fonte: Protótipo do *VoDiNorte*.

Semelhantemente ao *Índice de verbetes* a *Pesquisa por ordem alfabética*, ilustrada na Figura 42, exibe um menu com as letras do alfabeto que, por sua vez, relacionam todas as entradas existentes para cada letra. Esse tipo de visualização torna a pesquisa mais rápida em comparação com a ferramenta *Índice de verbetes* que exibe uma longa lista com todas as entradas do vocabulário. Destacamos, ainda, que esse índice poderá ser organizado, futuramente, na horizontal de modo a ocupar um espaço menor na tela do usuário.

Figura 42: *Pesquisa por ordem alfabética* do *VoDiNorte*.

Fonte: Protótipo do *VoDiNorte*.

Outra possibilidade de pesquisa que o usuário dispõe é a digitação da entrada que deseja acessar na caixa de pesquisa, conforme apresentado na Figura 43 e, caso a referida entrada faça parte da nomenclatura do vocabulário, o verbete aparecerá em outra página. Todavia, se a entrada não fizer parte da nomenclatura do protótipo do *VoDiNorte* ou for digitada com erros ortográficos, uma mensagem de erro será exibida e o usuário deverá clicar no botão *Página inicial* ou na seta de voltar do navegador de internet a fim de realizar uma nova pesquisa.

Figura 43: *Pesquisa por uma entrada* do *VoDiNorte*.

Fonte: Protótipo do *VoDiNorte*.

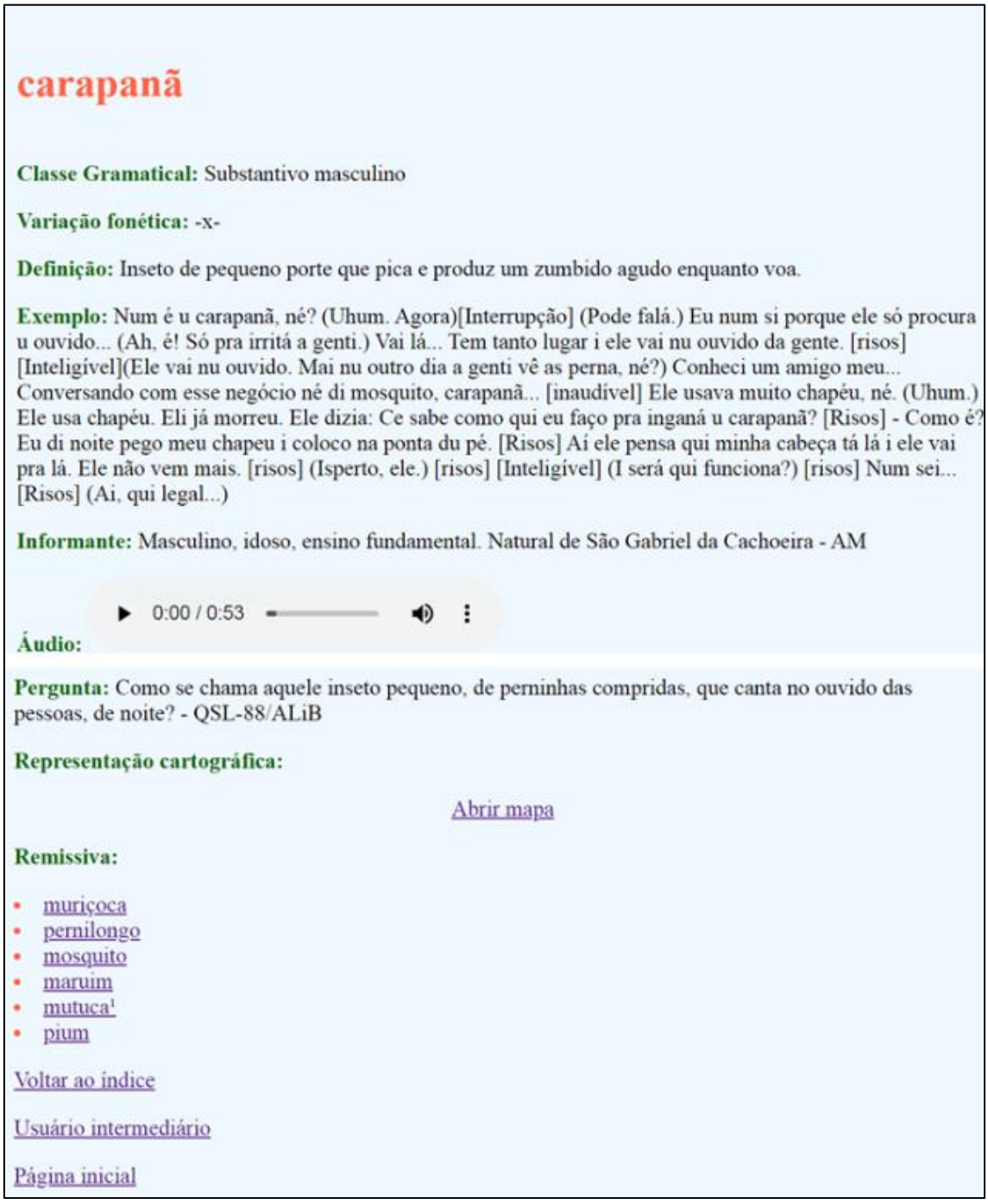
O usuário poderá também realizar buscas por meio da *Pesquisa por uma área semântica*, conforme ilustrado na figura 44. Assim, é possível visualizar as entradas tomando como ponto de partida a organização semântica feita pelo Projeto ALiB para o QSL. Além das 14 classificações propostas pelo ALiB, o usuário também dispõe de uma subclassificação na qual estão agrupadas as entradas do vocabulário. Desse modo, se um estudante quiser, por exemplo, pesquisar especificamente sobre todos os nomes de insetos que estão catalogados no protótipo do *VoDiNorte*, poderá clicar em *Fauna* e posteriormente em *insetos*, que a ferramenta de consulta lexicográfica exibirá a relação das entradas que se enquadram nessa descrição.

Figura 44: *Pesquisa por uma área semântica* do VoDiNorte.

Fonte: Protótipo do *VoDiNorte*.

Como é possível observar na Figura 44, as opções de entrada listadas na subcategoria *insetos* são *carapanã, muriçoca, pium, mosquito, varejeira e jacinta.* Assim, ao clicar em uma delas o verbete será exibido na tela como, por exemplo, no item *carapanã*, ilustrado na figura 45:

Figura 45: Verbete *carapanã* do *VoDiNorte*.



carapanã

Classe Gramatical: Substantivo masculino

Variação fonética: -x-

Definição: Inseto de pequeno porte que pica e produz um zumbido agudo enquanto voa.

Exemplo: Num é u carapanã, né? (Uhum. Agora)[Interrupção] (Pode falá.) Eu num si porque ele só procura u ouvido... (Ah, é! Só pra irritá a genti.) Vai lá... Tem tanto lugar i ele vai nu ouvido da gente. [risos] [Inteligível](Ele vai nu ouvido. Mai nu outro dia a genti vê as perna, né?) Conheci um amigo meu... Conversando com esse negócio né di mosquito, carapanã... [inaudível] Ele usava muito chapéu, né. (Uhum.) Ele usa chapéu. Eli já morreu. Ele dizia: Ce sabe como qui eu faço pra inganá u carapanã? [Risos] - Como é? Eu di noite pego meu chapeu i coloco na ponta du pé. [Risos] Aí ele pensa qui minha cabeça tá lá i ele vai pra lá. Ele não vem mais. [risos] (Isperto, ele.) [risos] [Inteligível] (I será qui funciona?) [risos] Num sei... [Risos] (Ai, qui legal...)

Informante: Masculino, idoso, ensino fundamental. Natural de São Gabriel da Cachoeira - AM

Áudio: 0:00 / 0:53

Pergunta: Como se chama aquele inseto pequeno, de perninhas compridas, que canta no ouvido das pessoas, de noite? - QSL-88/ALiB

Representação cartográfica:

Abrir mapa

Remissiva:

- muriçoca
- pernilongo
- mosquito
- maruim
- mutuca[1]
- pium

Voltar ao índice

Usuário intermediário

Página inicial

Fonte: Protótipo do *VoDiNorte*.

O verbete *carapanã* possui nove campos que exibem informações lexicográficas, a saber: classe gramatical; variação fonética; definição; exemplo; informante; áudio; pergunta; representação cartográfica; remissiva.

Assim, a *classe gramatical* representa um tipo de informação de caráter normativo, pois está alinhada às regras gramaticais da língua portuguesa. No entanto, a *variação fonética* exibe o modo de falar de uma UL que é própria da oralidade. No exemplo ilustrado na figura 45 o símbolo -*x*- indica que o informante mencionou o item léxico *carapanã* sem qualquer traço de oralidade.

A *definição*, por sua vez, é uma paráfrase lexicográfica construída por meio do modelo aristotélico, ou seja, em que o definidor é descrito a partir de um gênero próximo e de uma diferença específica. Ao seu turno, o *exemplo* exibe a fala do informante no momento da entrevista.

No campo destinado aos dados do *informante* o usuário pode identificar o sexo, a idade, a escolaridade e a localidade do entrevistado. Esses dados podem ser de pouca utilidade para os usuários comuns, mas são de grande valia para o estudioso mais especializado que, aliás, dispõe de outros recursos de pesquisa no perfil de usuário avançado. No entanto, os dados do informante estão visíveis ao usuário intermediário por se tratar de um modelo de verbete polifuncional.

O *áudio* é formado por uma ferramenta que executa um arquivo em *mp3* referente à fala do informante. Esse áudio é armazenado em uma pasta específica da aplicação web e é executada pela ferramenta de áudio por meio de uma orientação escrita em linhas de código no arquivo *.xqm*.

A indicação da pergunta do QSL que o verbete em questão se relaciona é exibida no campo *Pergunta* e, em seguida, o campo denominado *Representação cartográfica* exibe um link (*Abrir mapa)* em que o usuário pode consultar dados sobre a distribuição de ocorrências daquele verbete em um mapa interativo. Para dosar a quantidade de informação exibida na tela o mapa, ilustrado na figura 46, foi programado para ser exibido em outra aba do navegador ao clicar no link *Abrir mapa*. Assim, o usuário pode escolher se deseja ou não acessar o mapa, tornando a visualização dos dados da tela mais limpa, além de permitir uma navegação gradual.

Figura 46: Representação cartográfica do verbete *carapanã*.

Fonte: Protótipo do *VoDiNorte*.

Como é possível observar, o mapa interativo da figura 46 exibe o total das respostas dadas à pergunta 88 do QSL nas 18 localidades da rede de pontos do ALiB, do interior da região Norte. Desse modo, *carapanã* aparece como a resposta mais mencionada entre os entrevistados, seguida de *muriçoca*, *pernilongo*, *mosquito*, *marium*, *mutuca* e *pium*. Cada alfinete azul corresponde a uma das 18 localidades e ao clicar em um deles uma caixa de texto se abre com dados que especificam a resposta e o tipo de informante como, por exemplo, na cidade de Itaituba/PA em que três informantes (masculino idoso, feminino jovem, feminino idoso) disseram *carapanã*, dois informantes (masculino jovem, masculino idoso) mencionaram *muriçoca* e dois entrevistados (masculino jovem, feminino jovem) responderam *pernilongo*.

O último elemento da microestrutura do verbete é o sistema de remissiva que elenca as entradas relacionadas ao verbete principal que foram mencionadas com uma frequência menor em comparação com a frequência do verbete principal. Isso significa que a entrada *carapanã* é o verbete principal do vocabulário, pois foi a UL mais mencionada entre os entrevistados e possui seis remissivas (*muriçoca*, *pernilongo*, *mosquito*, *maruim*, *mutuca e pium*) que remetem apenas ao verbete principal.

Destacamos, ainda, que devido à escassez de tempo não foi possível avançar na no desenvolvimento das informações da microestrutura do protótipo e, desse modo, apenas sete verbetes estão com todos os dados relativos à microestrutura, a saber, o verbete completo *carapanã* e seus verbetes remissivos *muriçoca, pernilongo, mosquito, maruim, mutuca e pium*. Assim, esses verbetes possuem a descrição da classe gramatical, da variação fonética, da definição, do exemplo, da indicação do tipo de informante, da fala em áudio do informante (apenas para *carapanã, muriçoca, mosquito* e *pium*), da pergunta do QSL, da representação cartográfica e da relação de remissivas.

Além desses sete verbetes prototípicos é possível acessar outras 241 entradas que reúnem as respostas dos informantes do município de Oiapoque e que foram alocadas no banco de dados 2 para testar as funcionalidades das ferramentas destinadas ao usuário intermediário. Desse modo, esses 241 candidatos a verbete exibem ao usuário os seguintes dados: i) lema; ii) exemplo; iii) informações relacionadas ao tipo de informante e sua localidade; iv) pergunta do QSL; v) remissiva[82].

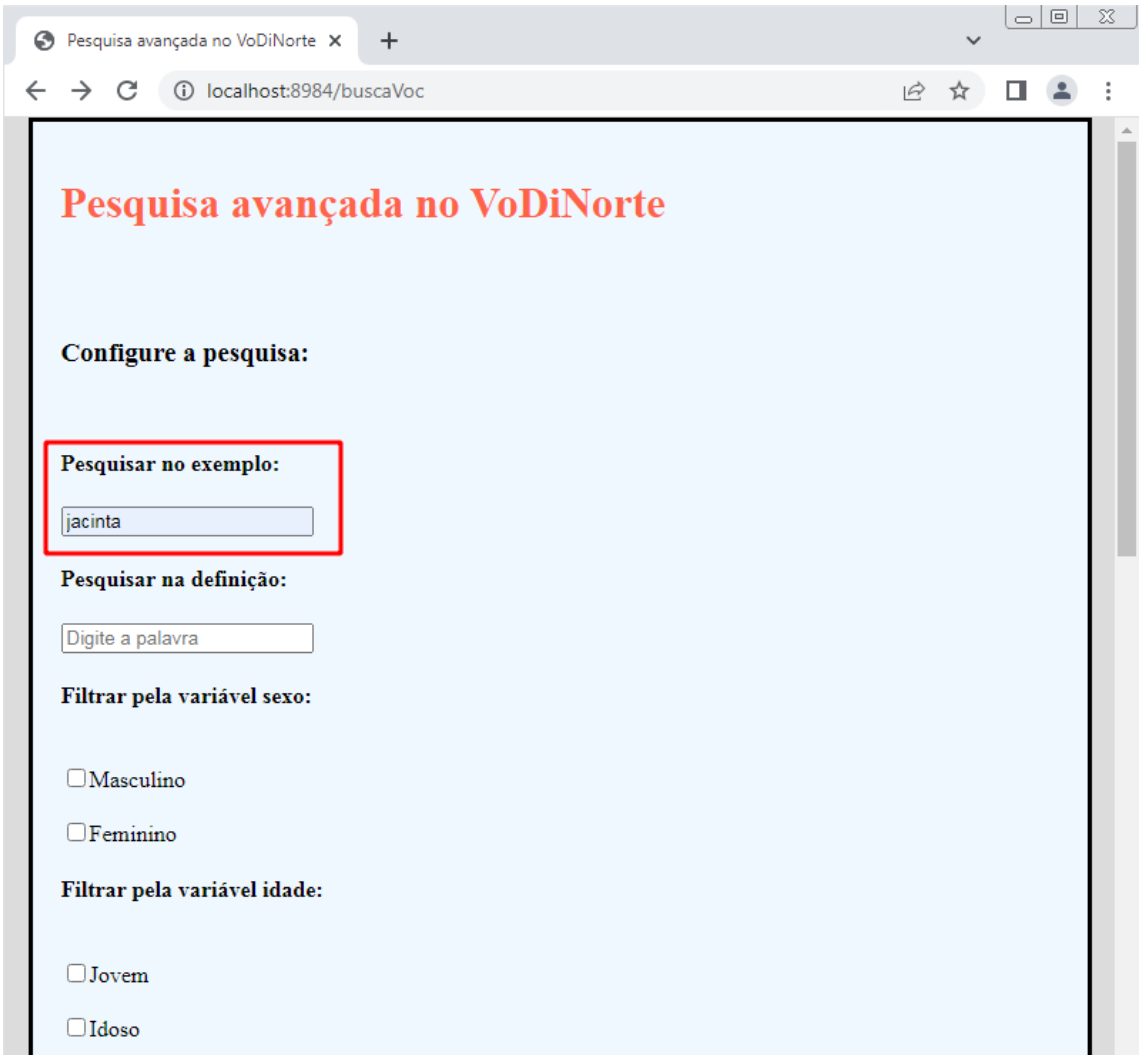
Por sua vez, a última ferramenta de pesquisa desenvolvida para o usuário intermediário se configura em um motor de busca capaz de exibir resultados por meio da configuração de filtros. A base de seu funcionamento é semelhante ao motor de busca existente para o perfil de usuário avançado, que está descrito na próxima seção (*5.4. Usuário avançado*). Todavia, a *Pesquisa avançada no VoDiNorte* realiza buscas apenas no banco de dados 2 o que significa que as informações exibidas se limitam a parcela de dados selecionados para o perfil de usuário intermediário, ou seja, não corresponde a uma busca em todo o corpus da pesquisa (banco de dados 1). Além disso, nesse perfil o usuário não dispõe da caixa de pesquisa que filtra os dados por meio de uma pergunta do QSL, ao passo que essa função existe na ferramenta de *Pesquisa avançada no corpus*.

Os filtros que permitem a busca avançada para o usuário intermediário podem ser classificados de duas formas: i) filtros que funcionam com a inserção de caracteres em uma caixa de pesquisa – destinados a recuperar informações referentes ao exemplo e à definição; ii) filtros que funcionam como botões de seleção – desenvolvidos para selecionar dados relativos ao perfil do informante e à localidade.

[82] Algumas remissivas não correspondem ao verbete em questão porque estão funcionando como teste do mecanismo de remissão e serão corrigidas futuramente.

Para exemplificar seu funcionamento solicitamos à ferramenta que exiba um exemplo que contenha a UL *jacinta*, como é possível observar na figura a seguir:

Figura 47: Pesquisa avançada no *VoDiNorte* para recuperar, no exemplo, a UL *jacinta*.



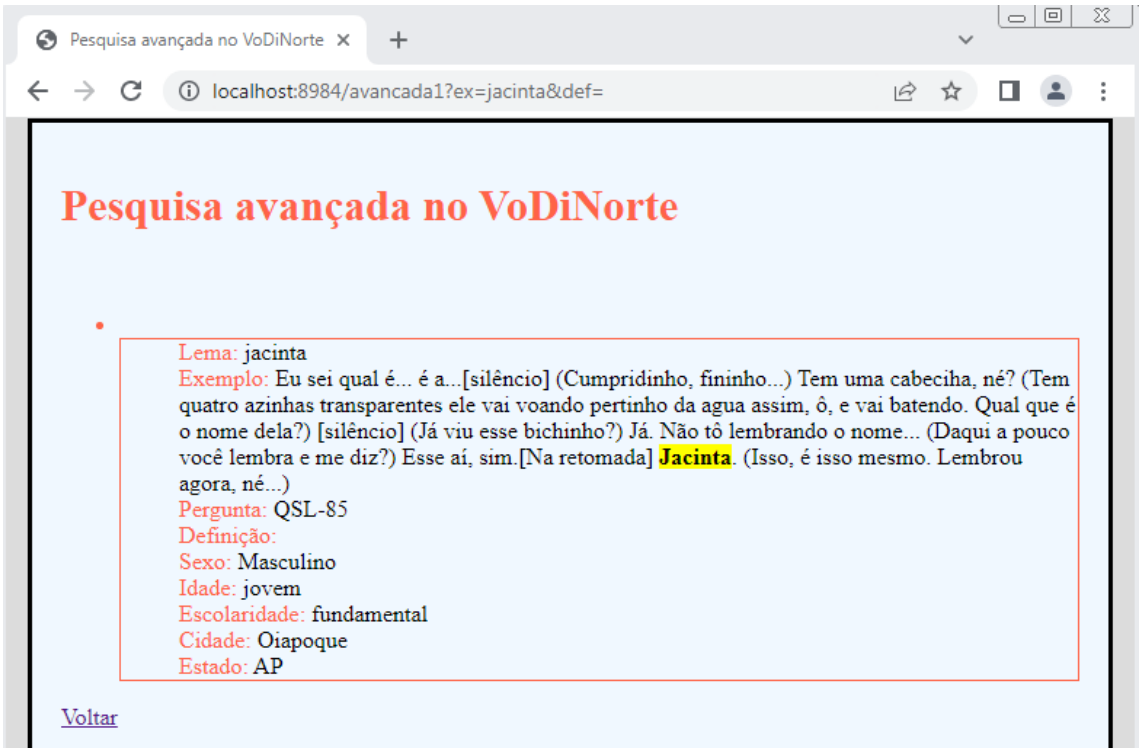
Fonte: Protótipo do *VoDiNorte*.

A partir da figura 47 é possível visualizar os filtros que estão nomeados em negrito, a saber: *Pesquisar no exemplo*; *Pesquisar na definição*; *Filtrar pela variável sexo*; *Filtrar pela variável idade*. O usuário também dispõe de mais um filtro, que não aparece na figura 47 devido à limitação de espaço, denominado *Filtrar pela variável localidade* em que é possível selecionar uma das 14 localidades do interior da região Norte para processar uma busca avançada.

Destacamos, ainda, que os filtros não utilizados pelo usuário faz com que o motor de busca selecione todos os dados, ou seja, se uma variável não for selecionada

significa que o consulente não deseja aplicar aquele filtro. Assim, o resultado para solicitação apresentada na figura 47 pode ser visto na figura a seguir:

Figura 48: Resultado da pesquisa avançada no *VoDiNorte* para recuperar, no exemplo, a UL *jacinta*.



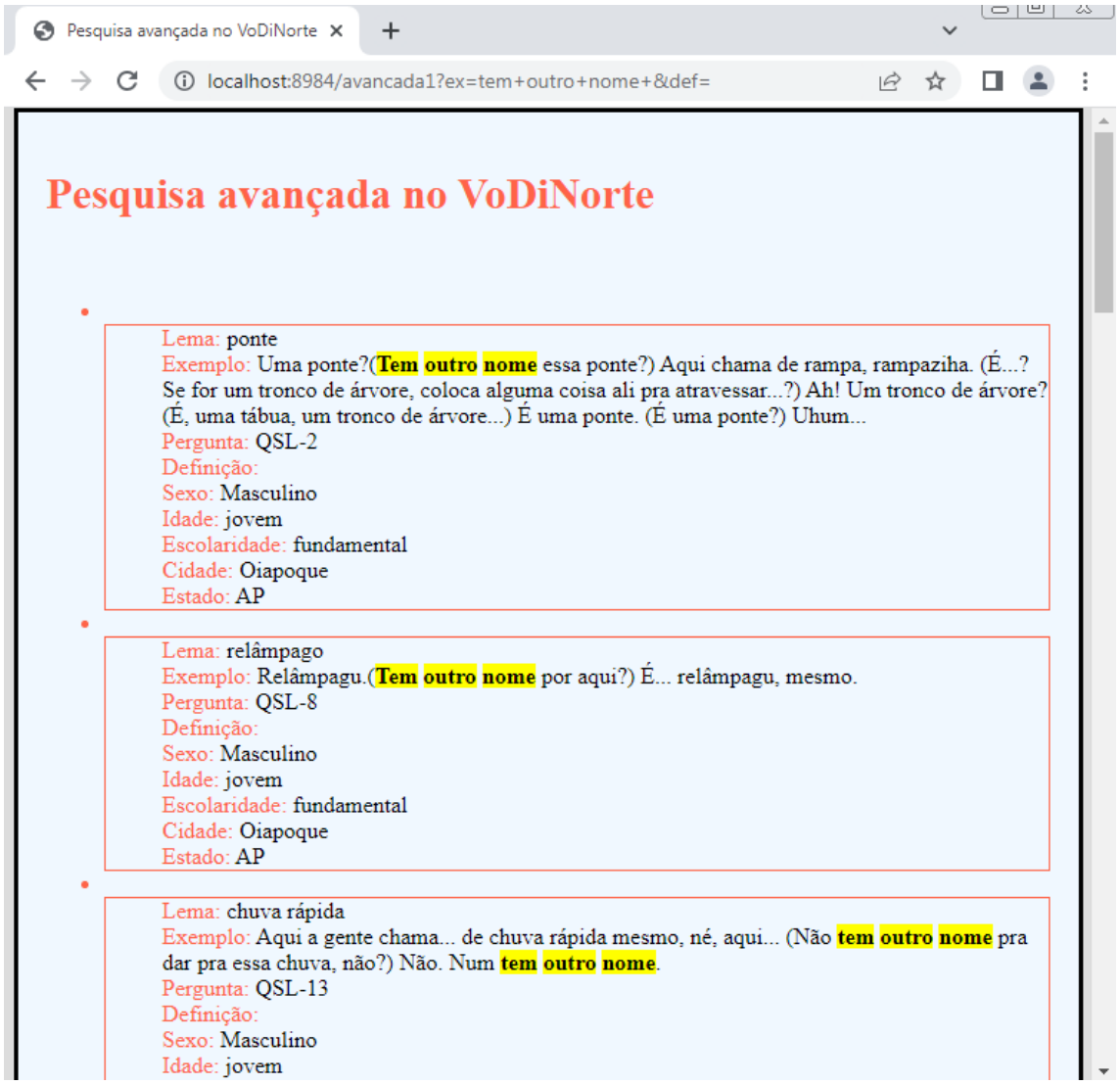
Fonte: Protótipo do *VoDiNorte*.

Como é possível observar na figura 48, a UL pesquisada aparece destacada em amarelo para facilitar a visualização do dado solicitado pelo usuário e foi programada para exibir também as informações referentes à pergunta do QSL, à definição[83], ao sexo, idade e escolaridade, além da cidade e estado. Vale acrescentar que essa ferramenta pode recuperar itens lexicais simples, como é o caso de *jacinta*, e também compostos. Assim, caso o usuário queira, por exemplo, fazer um estudo sobre a interação entre entrevistador e entrevistado para verificar outros nomes dados a um mesmo referente é possível recuperar, no exemplo, as falas em que há o uso da estrutura textual *tem outro nome*. O resultado para esta solicitação será uma lista com todos os verbetes e seus

[83] Como o protótipo do vocabulário ainda carece da escrita das definições lexicográficas, o espaço destinado a esse componente da microestrutura se encontra em branco.

respectivos exemplos que contêm esse conjunto textual, como é possível observar na figura a seguir:

Figura 49: Resultado da pesquisa avançada no *VoDiNorte* para recuperar, no exemplo, o conjunto textual *tem outro nome*.

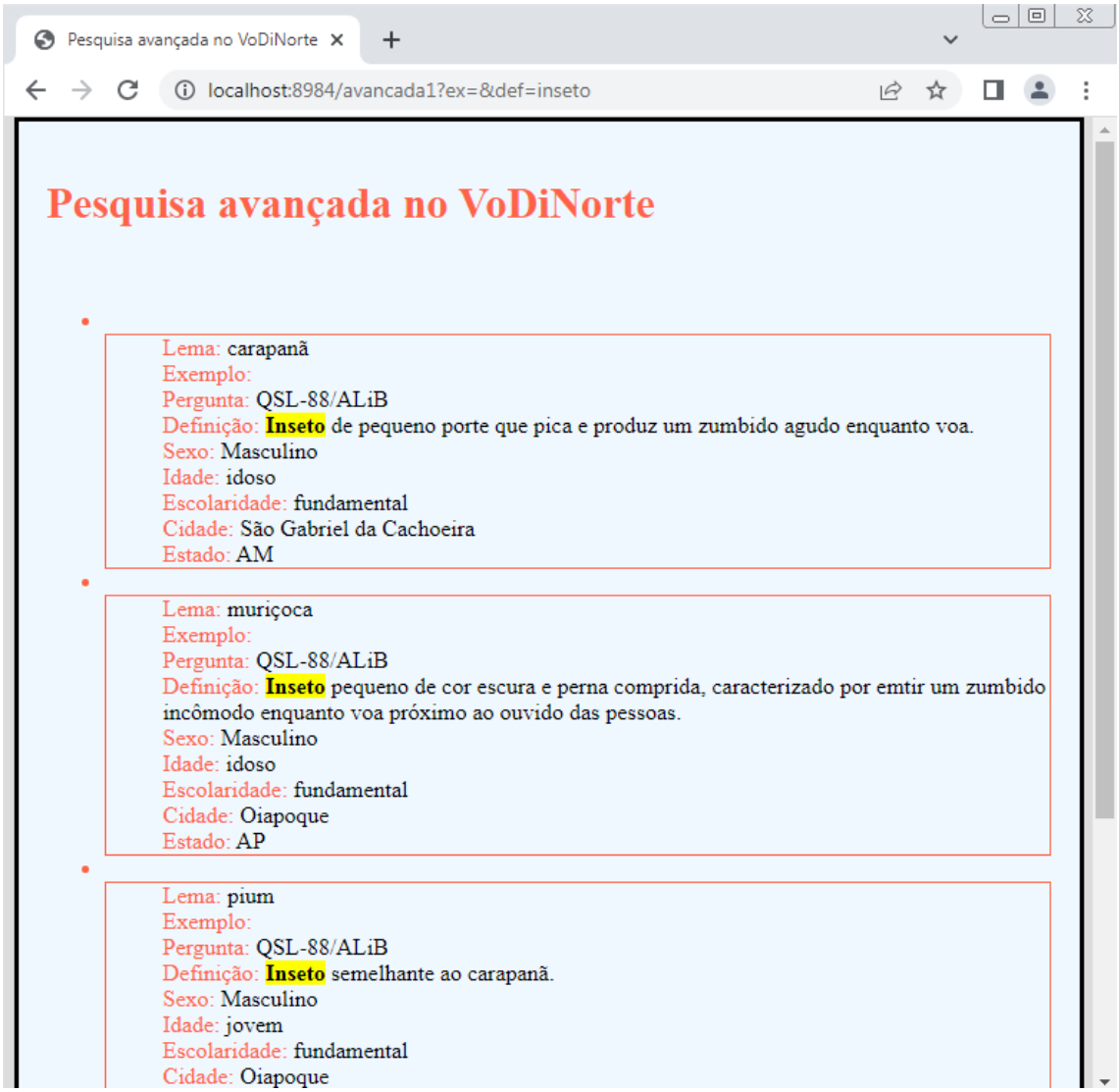


Fonte: Protótipo do *VoDiNorte*.

A segunda caixa de pesquisa disponibilizada ao usuário intermediário permite recuperar informações presentes no texto definitório dos verbetes. Como mencionado anteriormente, escrevemos a definição lexicográfica apenas dos verbetes *carapanã*, *muriçoca*, *pernilongo*, *mosquito*, *maruim*, *mutuca* e *pium* para realizar testes nas ferramentas de busca. Dessa forma, é possível pesquisar por uma UL simples ou composta que esteja presente na definição de um verbete como, por exemplo, no caso

ilustrado na figura a seguir, que representa o resultado de uma busca feita no campo *Pesquisar na definição* em que *a* UL *inseto* foi digitada. O motor de busca, por sua vez, exibiu o seguinte tela:

Figura 50: Resultado da pesquisa avançada no *VoDiNorte* para recuperar, na definição, o item lexical *inseto*.



Fonte: Protótipo do *VoDiNorte*.

Acrescentamos, ainda, que é possível estreitar as buscas selecionando uma das variáveis para sexo (masculino ou feminino), para idade (jovem ou idoso) e para localidade (14 municípios). No entanto, tendo em vista que o banco de dados 2 não está completo, as buscas por meio desses filtros é inviável. Desse modo, esse motor de pesquisa avançada faz parte do conjunto de ferramentas classificadas como prototípicas, pois o banco de dados 2 se encontra, atualmente, em construção.

5.4. Usuário avançado

As ferramentas desenvolvidas para o perfil de usuário avançado estão acessíveis e juntas se configuram como um motor de busca que permite recuperar informações no *corpus* de pesquisa de formas variadas. Além disso, os resultados exibidos são dinâmicos, isto é, mudam de acordo com as configurações dos filtros que são escolhidos pelo usuário a cada pesquisa.

Desse modo, a figura 51 ilustra a página inicial do perfil de usuário avançado em que é possível acessar o motor de pesquisa clicando em *Pesquisa avançada no corpus*. Além disso, é possível realizar buscas por meio do *Concordanceador da Linguateca*, que é uma ferramenta externa, acessada por meio de um link direto até a página web da Linguateca[84], em que o usuário pode conferir o contexto de uso de uma UL em um conjunto de *corpora*. Esse tipo de busca é útil, por exemplo, para compreender o uso de um item lexical de baixa frequência ou que não esteja dicionarizado nas principais obras lexicográficas da língua portuguesa.

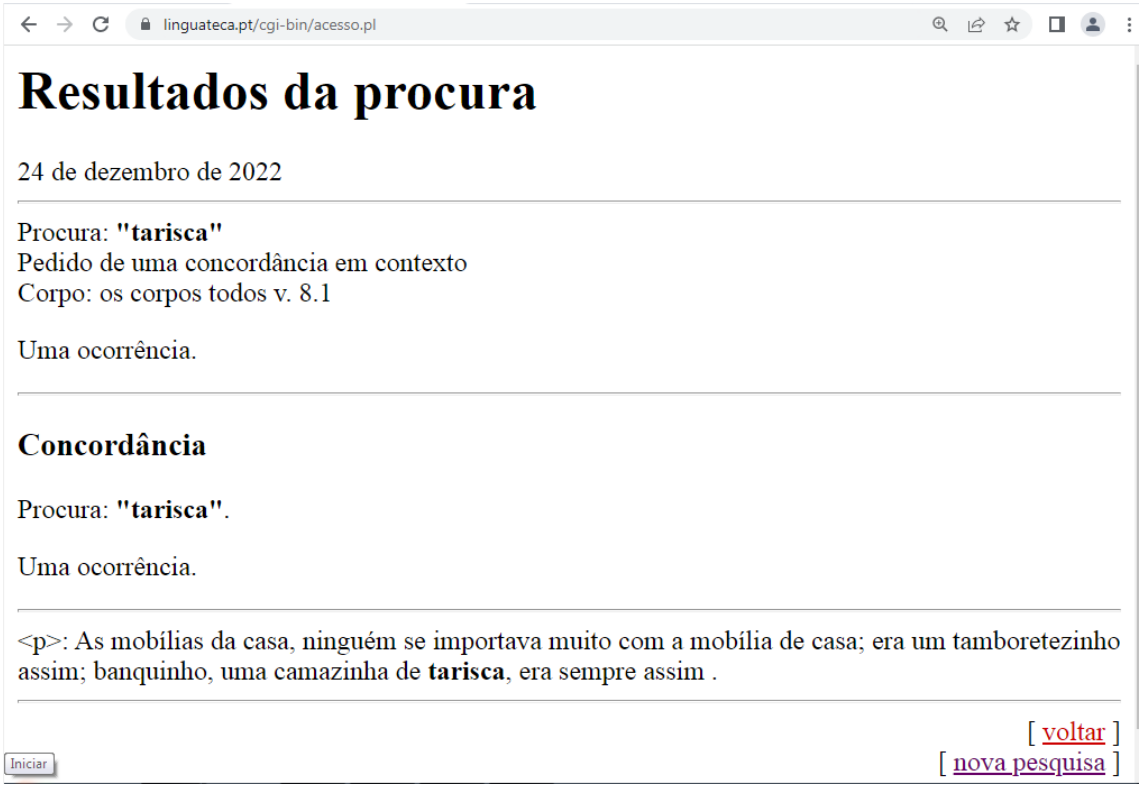
Figura 51: Tela inicial do perfil de usuário avançado.

Fonte: Protótipo do *VoDiNorte*.

[84] Projeto que visa a desenvolver recursos computacionais para o processamento da língua portuguesa de forma livre. Para mais informações acessar: <https://www.linguateca.pt/acesso/corpus.php?corpus=TODOS>. Acesso em 24 dez. 2022.

Antes de apresentarmos a *Pesquisa avançada no corpus* vamos ilustrar, por meio de uma experiência que tivemos durante a transcrição dos dados no arquivo *XML*, um exemplo de uso do *Concordanceador da Linguateta*. Para tanto, a UL *tarisca* foi documentado nas entrevistas do Projeto ALiB com o significado de lâminas do moinho que trituram a mandioca brava. Esse uso nos despertou o interesse de verificar como esse item lexical figurava nos dicionários de língua portuguesa e, par nossa surpresa, a UL *tarisca* não foi encontrada em nenhuma obra lexicográfica. Desse modo, recorremos ao concordanceador da Linguateca e encontramos um uso relacionado às tábuas de madeira, conforme é possível observar na figura a seguir:

Figura 52: Resultado para a UL *tarisca* no concordanceador da Linguateca.



Fonte: Linguateca. Acesso em: 24 dez. 2022.

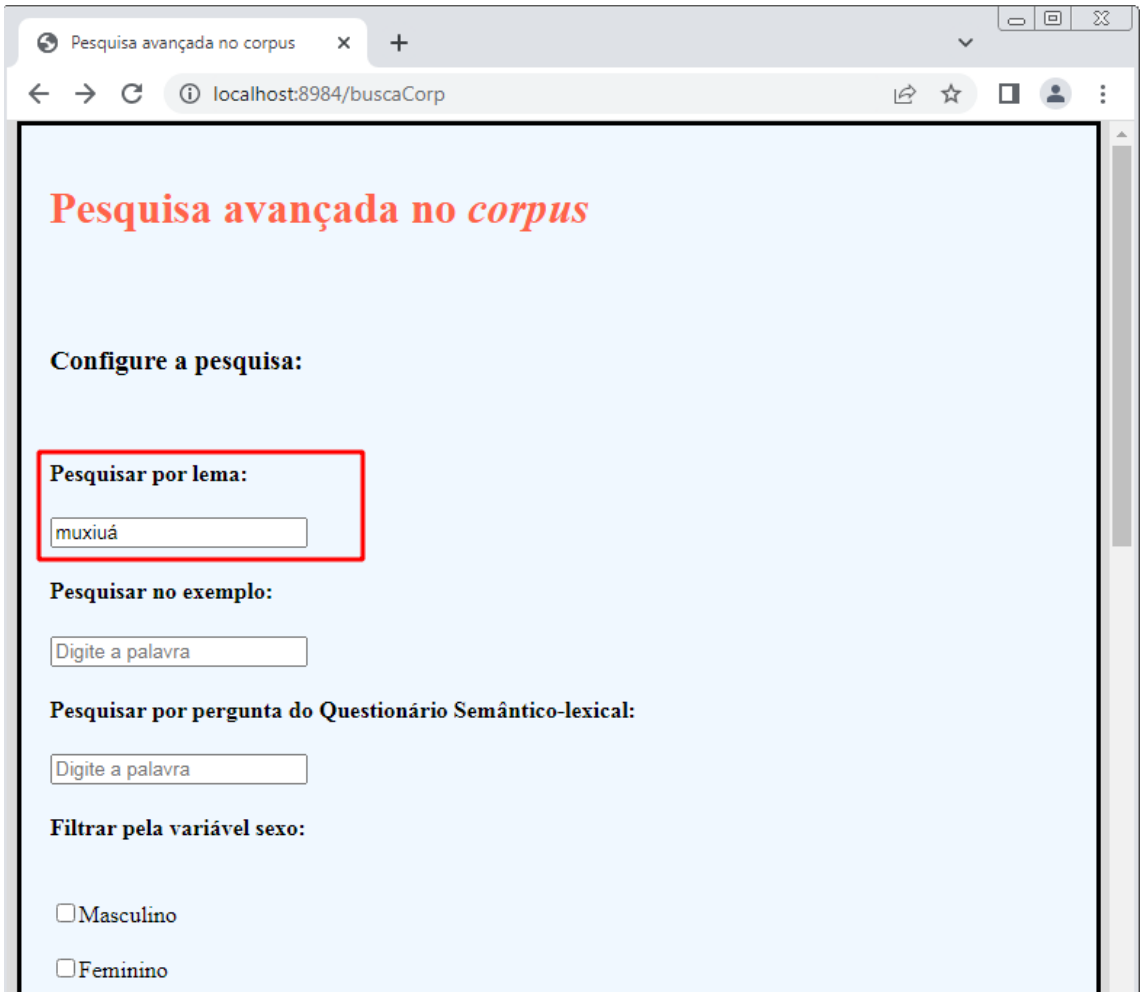
O resultado exibido na figura 52 não se assemelha ao contexto de uso documentado pelo projeto ALiB o que nos leva a classificar *tarisca* como um item lexical de uso peculiar, regional e pertencente a uma linguagem de especialidade, ou seja, veiculada na manufatura da farinha de mandioca. Nesse sentido, fica claro que

realizar buscas no *corpus* da Linguateca é uma tarefa importante para o linguista que está buscando, sobretudo, itens lexicais de baixa frequência e/ou não dicionarizados.

Ressaltamos, ainda, que o *Concordanceador da Linguateca* não é um produto desenvolvido no âmbito desta Tese e está associada ao perfil do usuário avançado por meio de um link externo por ser uma ferramenta útil às investigações lexicais. Desse modo, links de outros websites que sejam importantes para auxiliar o usuário avançado a processar o léxico de maneiras diversificadas podem ser agregados ao protótipo do *VoDiNorte* futuramente, pois é importante que uma plataforma on-line de processamento de dados lexicais ofereça um leque ampliado de ferramentas ao usuário.

Feita essa primeira apresentação do *Concordanceador da Linguateca* focaremos, na sequência, a ferramenta de *Pesquisa avançada no corpus*. Assim, nesse motor de busca o usuário avançado dispõe de um conjunto de filtros que recuperam informações do banco de dados 1, ou seja, do *corpus* completo da pesquisa que contêm todas as transcrições e não apenas os dados selecionadas para serem exibidos, de forma estática, como ocorre nas ferramentas de busca lexicográfica dos perfis de usuário comum e intermediário. Dessa forma, no perfil de usuário avançado a pesquisa é realizada com dados dinâmicos, ou seja, mudam de acordo com o tipo de informação requerida pelo usuário na página do motor de busca. Os filtros dessa ferramenta de pesquisa são caracterizados de duas maneiras, a saber: i) caixas de pesquisa – destinadas a recuperar informações referentes ao lema, ao exemplo e à pergunta do QSL; ii) botões de seleção – desenvolvidos para selecionar dados relativos ao perfil de informante e à localidade. Embora haja vários filtros que podem ser utilizados de maneira combinada, é possível realizar uma busca preenchendo apenas uma das caixas de pesquisa, conforme ilustrado na figura a seguir:

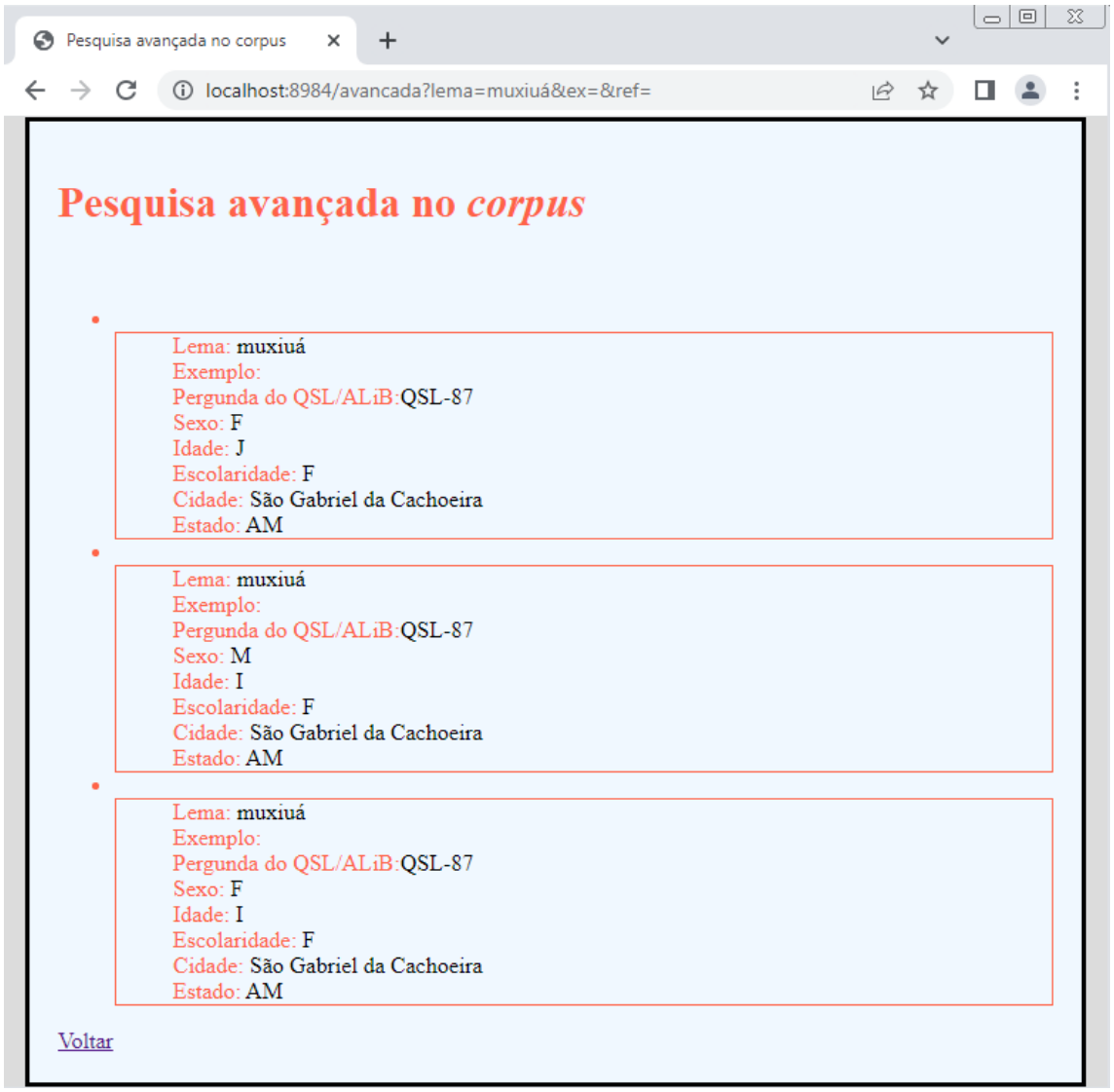
Figura 53: Utilizando a *Pesquisa por lema* para buscar a UL *muxiuá*.



Fonte: Protótipo do *VoDiNorte*.

Ao processar a solicitação da figura 53 o motor de busca irá recuperar todos os candidatos a lema do *corpus*, referente ao item lexical *muxiuá*, armazenados no banco de dados 1 como se pode observar na figura a seguir:

Figura 54: Resultados da *Pesquisa por lema* para a UL *muxiuá*.



Fonte: Protótipo do *VoDiNorte*.

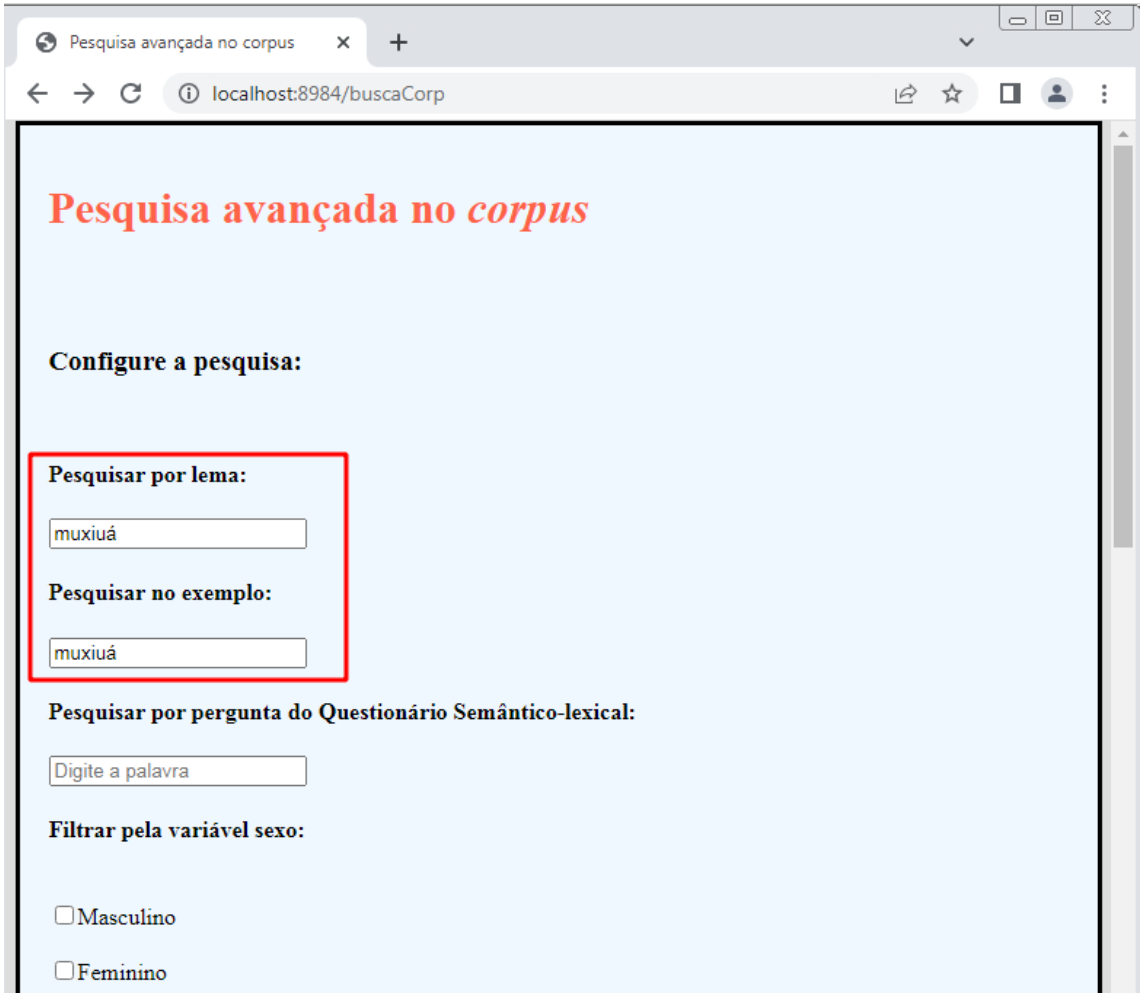
Como é possível observar, a figura 54 exibe informações estruturadas em formato de coluna e segue a seguinte ordem: i) lema; ii) exemplo; iii) pergunta do QSL/ALiB; iv) sexo; v) idade; vi) escolaridade; vii) cidade; viii) estado. Destacamos que o campo *exemplo* aparece vazio, pois não foi solicitado esse tipo de informação na busca ilustrada na figura 53. Esclarecemos, também, que os dados dos campos *sexo*, *idade* e *escolaridade* estão abreviados, porque essa estrutura foi adotada na arquitetura

do *XML*[85] no início da pesquisa e, nesse primeiro momento, não tínhamos pensado que no âmbito da Lexicografia Eletrônica as abreviações não são necessárias. Mas, como alterar esse tipo de dado demandaria muito tempo, já que deve ser feito manualmente, tais abreviações permanecem nos resultados da ferramenta de busca avançada no *corpus*. Dessa forma, no campo *sexo* poderão surgir abreviações como *M* (masculino) e *F* (feminino), e no campo *idade* poderão aparecer abreviações como *I* (idoso) e *J* (jovem). No que tange à escolaridade, a abreviação *F* faz menção ao Ensino Fundamental, que representa todos os informantes do interior entrevistados pelo Projeto ALiB.

Como mencionado anteriormente, é possível fazer uso dos filtros em conjunto. Dessa forma, para recuperar informações relacionadas ao *lema* e ao *exemplo* é preciso escrever algo na caixa *Pesquisar no exemplo* como ilustrado na figura a seguir:

[85] No banco de dados 2 as abreviações não foram utilizadas, pois esse arquivo *XML* foi construído muitos meses após a criação do banco de dados 1.

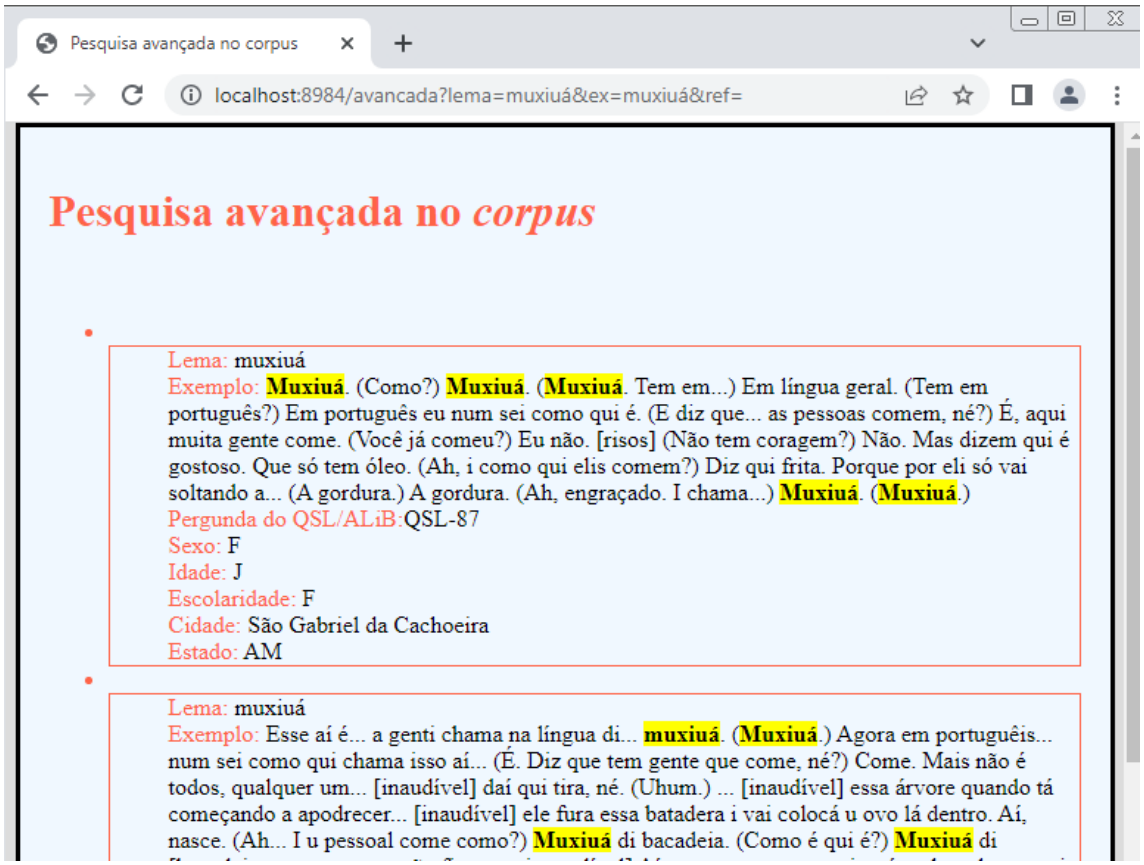
Figura 55: Busca avançada no *corpus* da UL *muxiuá* por meio dos filtros *lema* e *exemplo*.



Fonte: Protótipo do *VoDiNorte*.

O resultado dessa solicitação pode ser visto na figura 56 em que os dados relacionados ao exemplo são exibidos e a UL pesquisada aparece destacada em amarelo, facilitando a visualização da informação requerida pelo usuário. Acrescentamos, ainda, que a figura 56 mostra apenas um recorte dos resultados devido à limitação de espaço da figura. No entanto, o usuário durante seu acesso ao protótipo do *VoDiNorte* tem a visão de todos os resultados devendo, para tanto, rolar a roda do mouse para baixo.

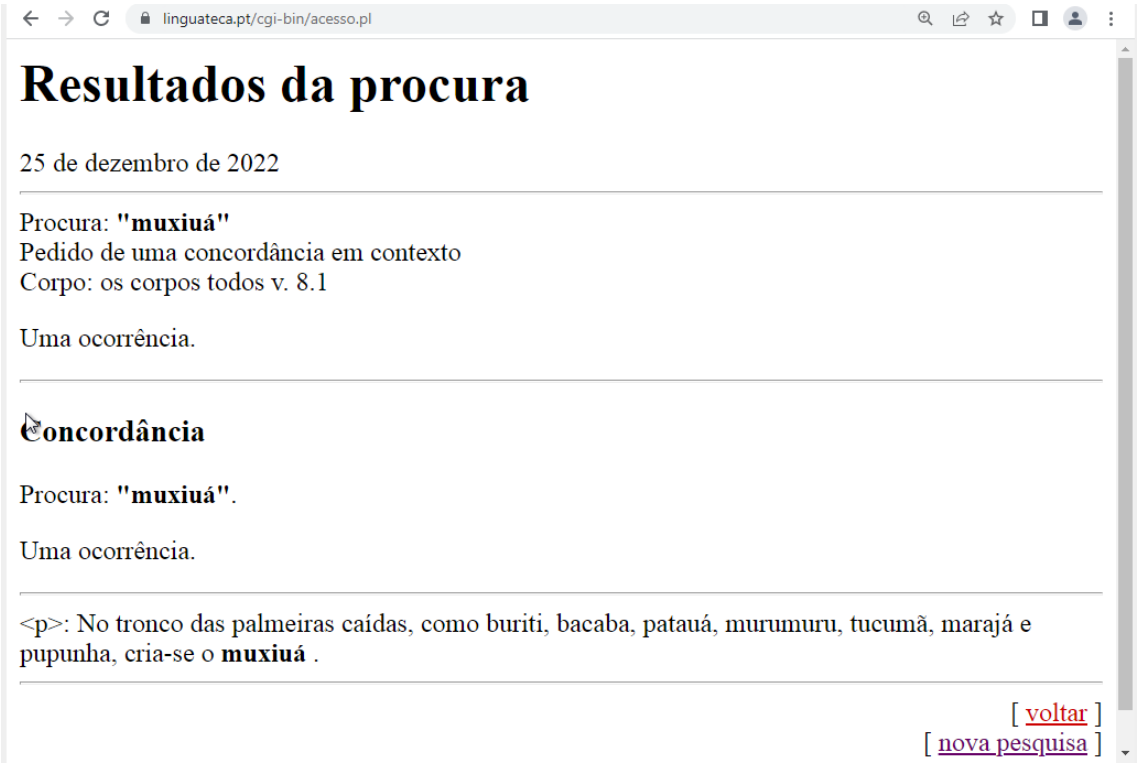
Figura 56: Resultados para a busca avançada no corpus da UL *muxiuá* por meio dos filtros *lema* e *exemplo*.



Fonte: Protótipo do *VoDiNorte*.

Vale destacar que a UL *muxiuá* não está dicionarizada e tão pouco foi encontrada pelo buscador Google, demonstrando um uso bastante restrito. Porém, ao digitá-la na ferramenta de pesquisa da Linguateca, uma única ocorrência foi exibida e com um uso convergente em relação ao emprego registrado pelos entrevistadores do Projeto ALiB, ou seja, um tipo de larva encontrada em paus podres, como é possível observar na figura a seguir:

Figura 57: Resultado para a UL *muxiuá* no concordanceador da Linguateca.



Fonte: Linguateca. Acesso em: 24 dez. 2022.

Desse modo, a pesquisa relacionada à UL *muxiuá* representa mais um exemplo que justifica a importância do *Concordanceador da Linguateca* para os estudos lexicais e, por isso, escolhemos torná-lo acessível de maneira rápida ao usuário a partir de um link direto na página inicial do protótipo do *VoDiNorte*, na área destinada ao perfil de usuário avançado.

Outro ponto importante a ser destacado diz respeito à caixa de busca *Pesquisar por lema* que pode recuperar, além de UL simples ou compostas, itens lexicais a partir de letras que formam um lema como, por exemplo, um prefixo, um radical ou um sufixo. Para ilustrar essa funcionalidade, digitamos os caracteres *ss* na caixa de pesquisa por lema e, ao clicar em *Pesquisar*, o resultado foi um conjunto de candidatos à entrada que levam *ss* em sua grafia como, por exemplo, girassol, sanguessuga, pão massa grossa, travessa etc. Desse modo, essa capacidade de recuperação de caracteres torna possível a busca de elementos mínimos formadores de uma UL, que podem ser úteis em estudos ordem morfológica.

Vale destacar que a implementação dessa funcionalidade representa um uso das tecnologias disruptivas aplicado na elaboração de obras lexicográficas eletrônicas. Nesse sentido, reiteramos que esse tipo de possibilidade não existe no Houaiss

eletrônico (2009), como apontado na seção *2.2. A versão eletrônica do Dicionário Houaiss da Língua Portuguesa (2009)*, e é mencionada por Fuertes-Olivera; Bergenholtz (2015) como uma inovação do *Diccionario de posibilidades de español* como se pode averiguar na seguinte citação:

> Finalmente habrá un DICCIONARIO DE "POSIBILIDADES" DE ESPAÑOL "UNIVERSIDAD DE VALLADOLID" que será un diccionario completamente novedoso. Con este diccionario un usuario puede llevar a cabo diferentes tipos de búsquedas booleanas con el objetivo de encontrar una palabra o un concepto que recuerda de forma tan vaga que no sabe exactamente qué es lo que tiene que buscar[86] (FUERTES-OLIVERA; BERGENHOLTZ, 2015, p. 92-93).

Nesse ponto, uma importante reflexão se faz necessária no sentido de colocar em evidência uma diferença que distingue as ferramentas que formam o motor de busca lexicográfica do perfil de usuário avançado, das ferramentas desenvolvidas para os demais perfis de usuário do protótipo do *VoDiNorte*. Essa diferença pode ser considerada como um divisor de águas e se resume, basicamente, na forma como as aplicações computacionais foram pensadas que, por sua vez, estão associadas a dois tipos de saberes essenciais: i) conhecimentos técnicos relacionados à construção de ferramentas computacionais; ii) conhecimentos relacionados à Lexicografia Eletrônica aplicada ao desenvolvimento de obras lexicográficas inovadoras.

Esses dois tipos de saberes são perceptíveis no decorrer desta Tese de modo que é possível classificar as ferramentas computacionais produzidas para o protótipo do *VoDiNorte* em dois grupos[87], a saber: i) ferramentas do tipo *Faster Horses* – desenvolvidas durante os estudos iniciais e que foram sendo melhoradas à medida em que a pesquisa avançou; ii) ferramentas do tipo *Model T Fords* – que começaram a ser desenvolvidas em um momento de maior consolidação da pesquisa do ponto de vista teórico e metodológico, porém, não havia mais tempo para grandes mudanças.

Em meio a esse cenário, escolhas tiveram que ser feitas de modo que o protótipo desta Tese pudesse ser visto não apenas como um vocabulário dialetal, mas como uma plataforma on-line prototípica que reúne ferramentas diversificadas de pesquisa a dados

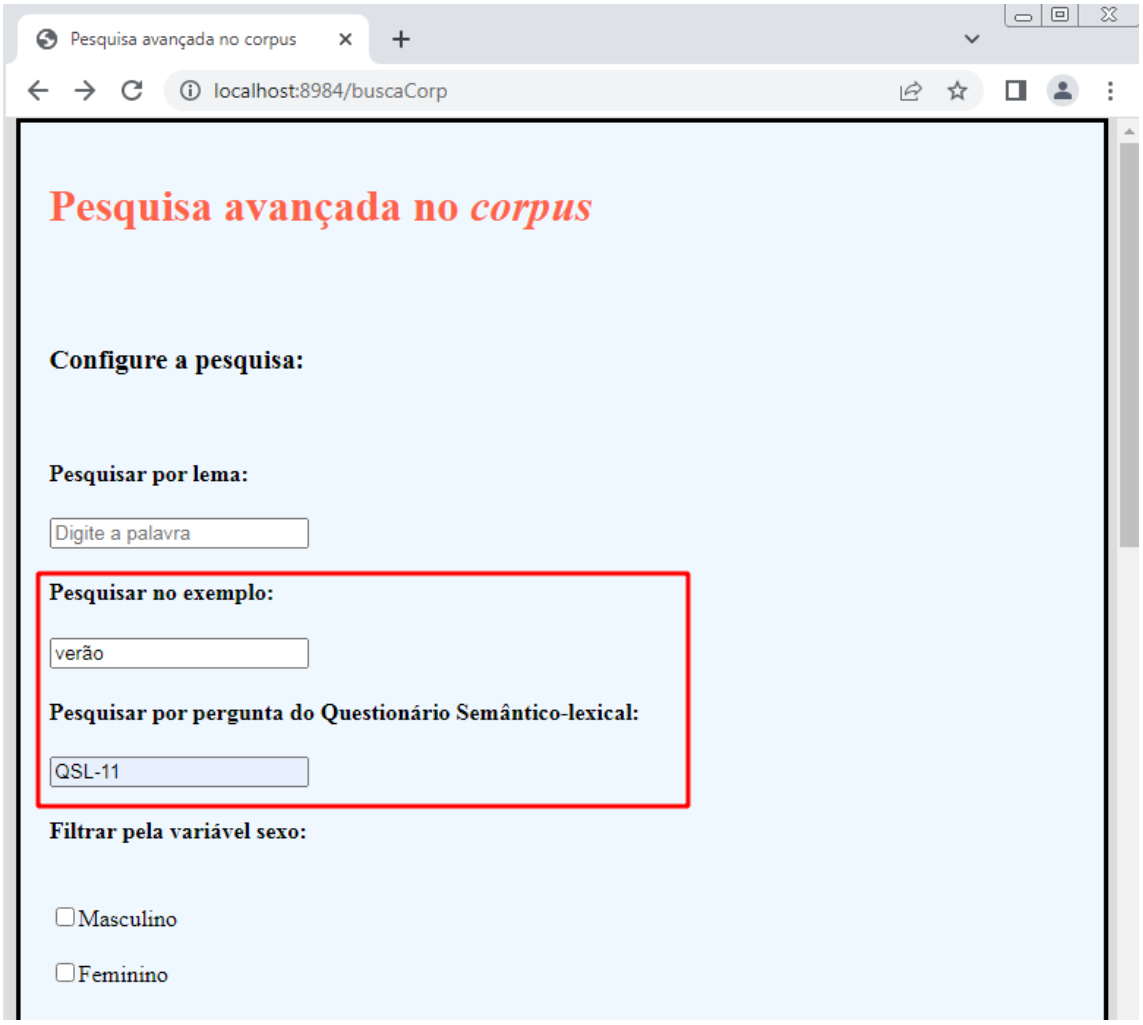
[86] "Finalmente haverá um DICCIONARIO DE "POSIBILIDADES" DE ESPAÑOL "UNIVERSIDAD DE VALLADOLID" que será um dicionário completamente novo. Com este dicionário um usuário pode realizar diferentes tipos de buscas boleanas com o objetivo de encontrar uma palavra ou um conceito que do qual se lembra tão vagamente que não se lembra exatamente o que procurar." (T.N.).

[87] Conforme a classificação de dicionários eletrônicos apresentada na seção *2.6. Critérios de classificação de dicionários eletrônicos.*

tratados lexicograficamente. Tais ferramentas, atualmente, se apresentam por meio de três situações distintas de uso que correspondem aos três perfis de usuário do protótipo do *VoDiNorte*.

Convém, ainda, explicitar outras maneiras de realizar uma busca no motor de pesquisa do perfil de usuário avançado. Assim, além da pesquisa por meio das caixas que recuperam informações relacionadas ao lema e ao exemplo, o motor de busca avançada pode exibir resultados filtrados pela pergunta do QSL. Esse filtro pode ser usado isoladamente, combinado com outras variáveis ou até mesmo ser utilizado juntamente com uma UL específica do lema ou do exemplo, como ilustrado na figura a seguir:

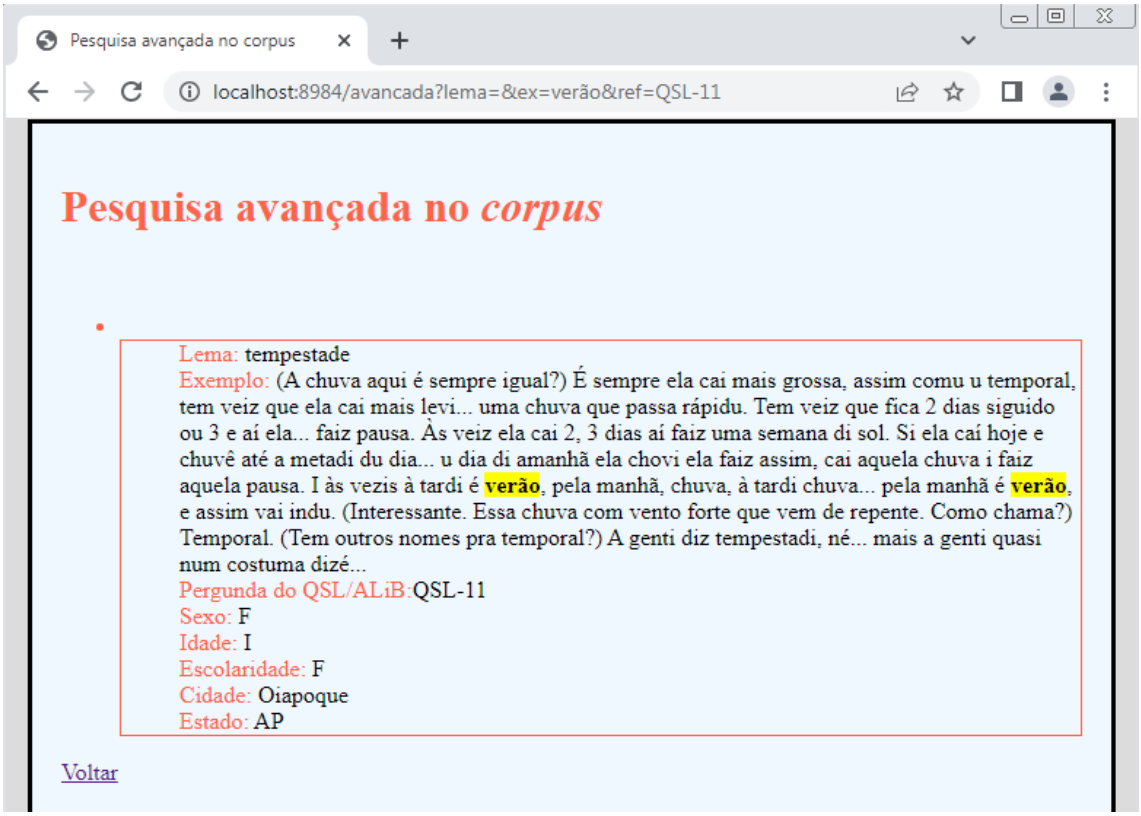
Figura 58: Pesquisa avançada no *corpus* utilizando filtros relacionados ao exemplo e à pergunta do QSL.



Fonte: Protótipo do *VoDiNorte*.

Como é possível observar, por meio da figura 58, solicitamos ao motor de pesquisa que recupere verbetes em que no exemplo ocorra o uso da UL *verão* e, ainda, restringimos o resultado para exibir apenas as respostas dadas para a pergunta 11 do QSL, no qual o informante é indagado sobre o nome que se dá para uma chuva com vento forte que vem de repente. O resultado exibido está apresentado na figura a seguir:

Figura 59: Resultado para a pesquisa avançada no *corpus* para a UL *verão* no filtro do exemplo e *QSL-11* no filtro da pergunta do QSL.



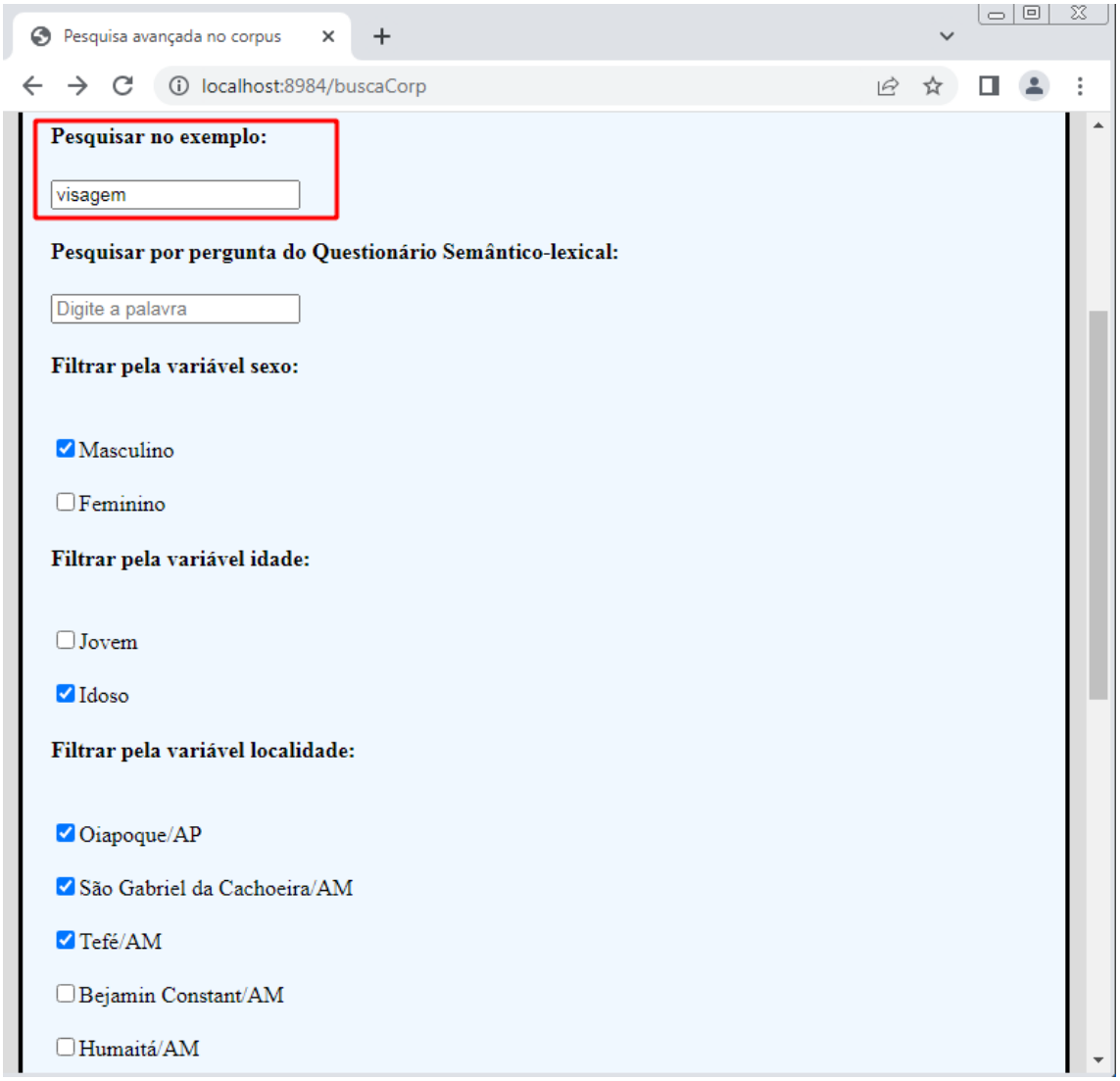
Fonte: Protótipo do *VoDiNorte*.

Por ser uma pesquisa bastante restrita, a ferramenta de busca trouxe apenas um resultado em que a UL *verão* foi mencionada duas vezes por uma informante idosa, do sexo feminino e moradora da cidade de Oiapoque/AP para representar o tempo sem chuva, isto é, a UL *verão* é utilizada para denominar o tempo seco, sem chuva.

No intuito de explorar uma pesquisa com uso combinado de outros filtros solicitamos, em uma nova busca, a recuperação de dados na barra de pesquisa do exemplo referente ao item lexical *visagem* e, além disso, indicamos para que se mostrem apenas os resultados em que o informante seja idoso e do sexo masculino.

Também limitamos a busca aos municípios de Oiapoque/AP, São Gabriel da Cachoeira/AM e Tefé/AM, como pode ser observado na figura a seguir:

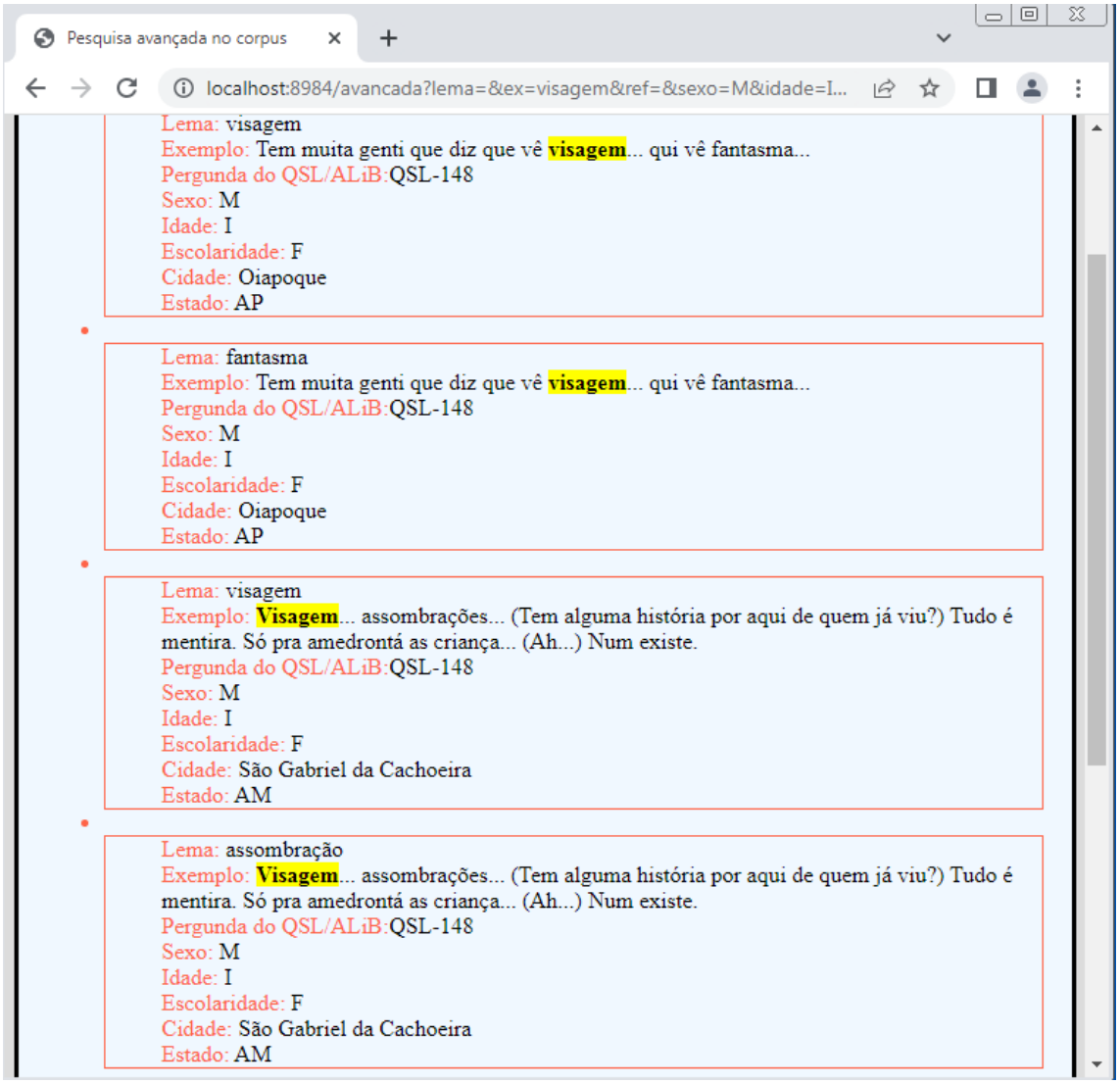
Figura 60: Pesquisa avançada no *corpus* com uma combinação de filtros para a UL *visagem* no exemplo.



Fonte: Protótipo do *VoDiNorte*.

Acrescentamos, ainda, que a solicitação ilustrada na figura 60 poderia ser mais afunilada se tivéssemos indicado a exibição apenas de respostas em que o lema também fosse a UL *visagem* fornecida por alguns informantes como resposta para a pergunta 148 do QSL-ALiB: "O que algumas pessoas dizem já ter visto, à noite, em cemitérios ou em casas, que se diz que é do outro mundo?" (COMITÊ NACIONAL..., 2001, p. 33). Desse modo, o resultado para a solicitação da figura 60 pode ser vista na figura a seguir:

Figura 61: Resultado para a pesquisa avançada no *corpus* com uma combinação de filtros para a UL *visagem* no exemplo.



Fonte: Protótipo do *VoDiNorte*.

Como é possível constatar, a figura 61 reúne as respostas dadas por informantes idosos, do sexo masculino e morador da cidade de Oiapoque/AP, São Gabriel da Cachoeira/AM e Tefé/AM que mencionou *visagem* ao responderem a pergunta 148 do QSL. Vale destacar que os resultados para o município de Tefé/AM não aparecem na figura 61 devido ao limite de espaço da captura de tela.

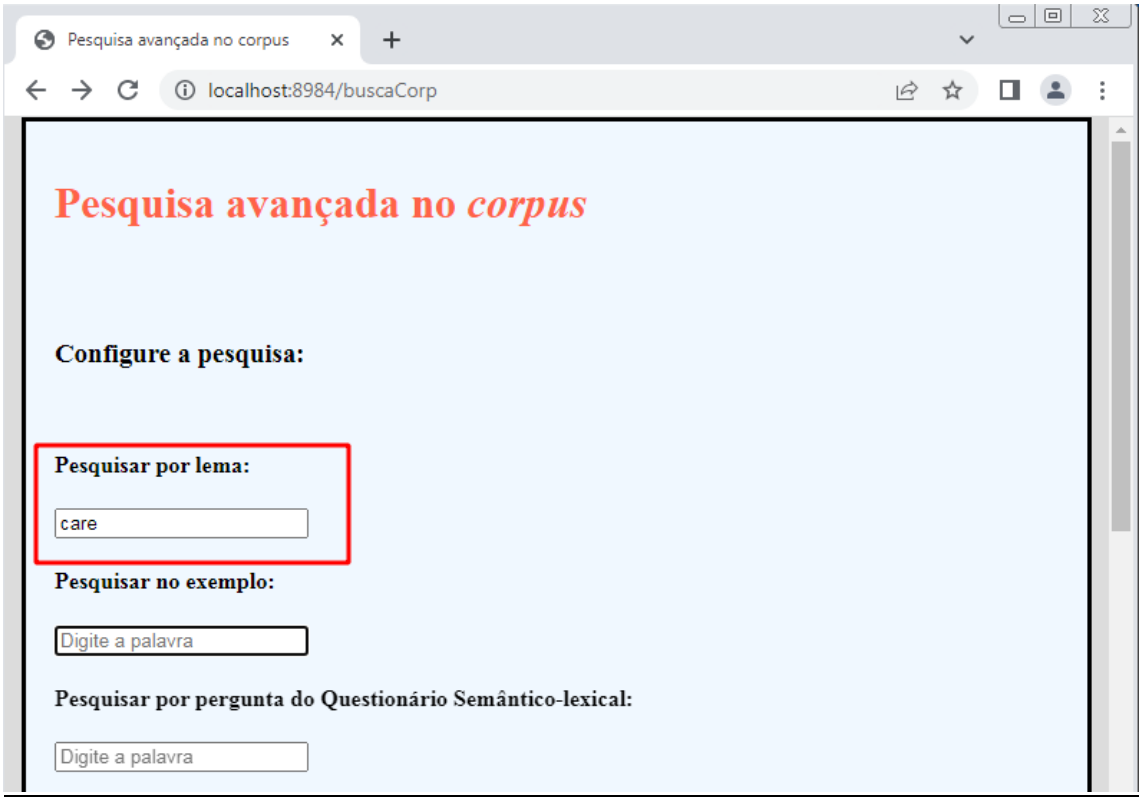
Com a finalidade de ilustrar outra possibilidade de uso do protótipo do *VoDiNorte*, supomos que um usuário queira identificar se houve algum informante que mencionou a UL *careca* ou *carequinha* para se referir ao *pão francês* nos inquéritos do

Projeto ALiB no interior da região Norte do Brasil. A pergunta que aborda esse tipo de variação lexical é a de número 186 do QSL, a saber: “Como se chama isto (mostrar um pão francês)”. Vamos, ainda, considerar que esse usuário não conheça o QSL e, portanto, não poderá recorrer à barra de busca por uma pergunta do QSL.

Diante desse contexto, o usuário poderá fazer dois tipos de pesquisa, a saber: i) digitar *pão careca* na barra de busca por *lema* e na barra de busca por *exemplo* e visualizar os resultados. Porém, essa busca não contempla possíveis variações sufixais; ii) realizar a busca por lema a partir do termo *pão.* Todavia, os resultados exibidos serão abrangentes, pois adicionarão também as menções a outras UL como *pão duro*, referente à pergunta 138 “Como se chama a pessoa que não gosta de gastar o seu dinheiro e, às vezes, até passa dificuldades para não gastar?” e *pão bengala* referente à pergunta 187 “Como se chama isto? (mostrar um pão bengala)” do QSL.

Assim, para estreitar a pesquisa e ser mais assertivo naquilo que se quer identificar é possível realizar a busca por meio da raiz do item léxico principal, ou seja, digitar na barra *Pesquisar por lema* o radical *care*, como ilustrado na figura a seguir:

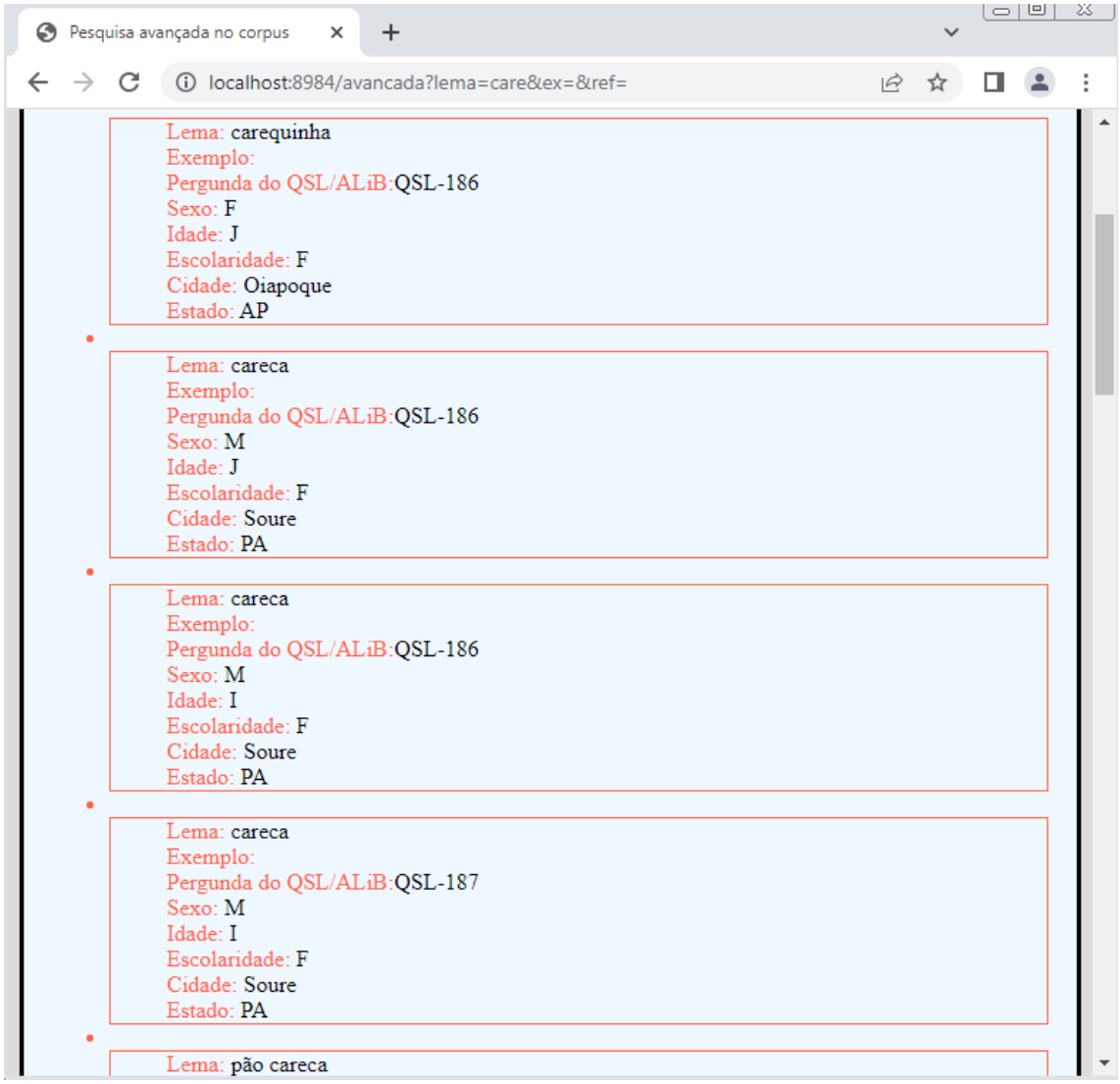
Figura 62: Pesquisa avançada no *corpus* a partir do radical *care*.



Fonte: Protótipo do *VoDiNorte*.

Ressaltamos que a busca por radical só funciona na barra de pesquisa destinada aos lemas. Isso significa que a barra de pesquisa das abonações só recupera UL separadas por um branco, ou seja, palavras completas. Isso significa que, ao tentar buscar por um radical, prefixo ou sufixo na barra *Pesquisar por exemplo* nada é mostrado ao usuário, pois esse tipo de funcionalidade está operacional apenas na busca por lema. Os resultados para a solicitação de pesquisa da figura 62 podem ser vistos a seguir:

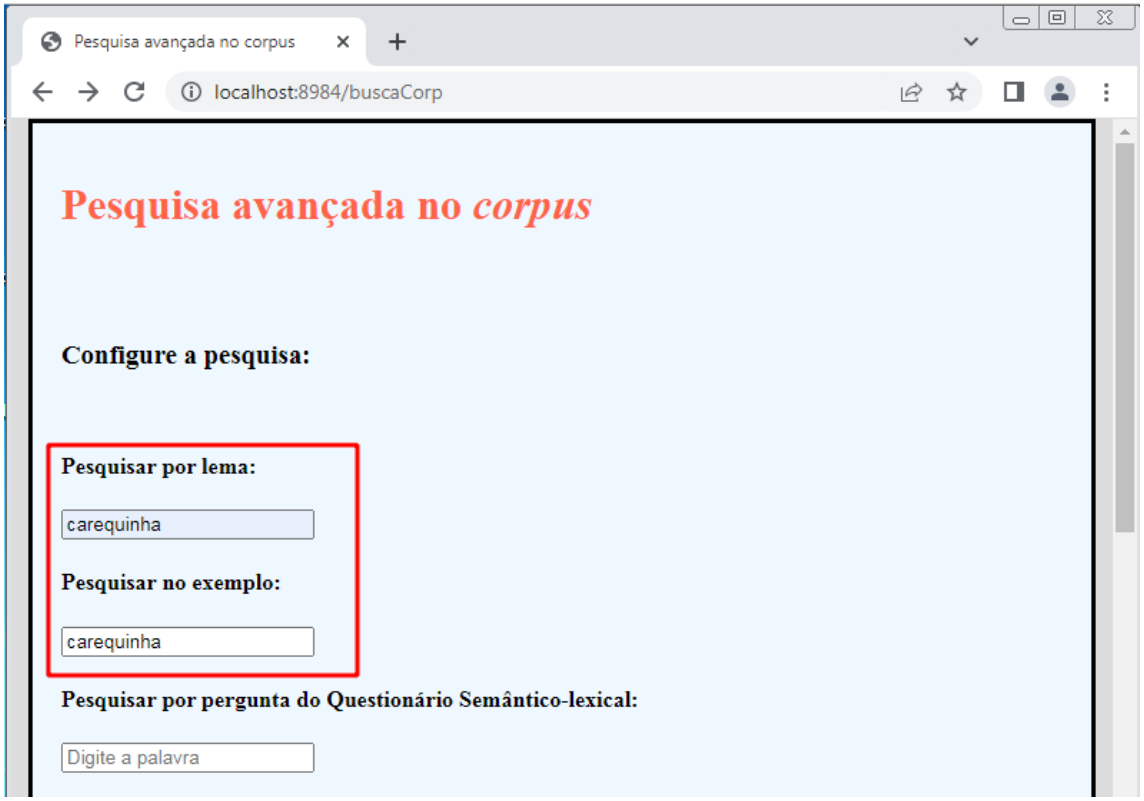
Figura 63: Resultado para a pesquisa avançada no *corpus* a partir do radical *care*.



Fonte: Protótipo do *VoDiNorte*.

Dentre os resultados gerados pelo motor de busca é possível observar, na figura 63, apenas as UL *carequinha*, *careca* e *pão careca* por conta da limitação de espaço da captura de tela. Desse modo, o usuário deve rolar a página para baixo e, caso queira visualizar o texto do exemplo de uma UL, terá que digitar na página de *Pesquisa avançada no corpus* o item lexical *carequinha* nos campos *Pesquisa por lema* e *Pesquisa por exemplo*, conforme a figura a seguir:

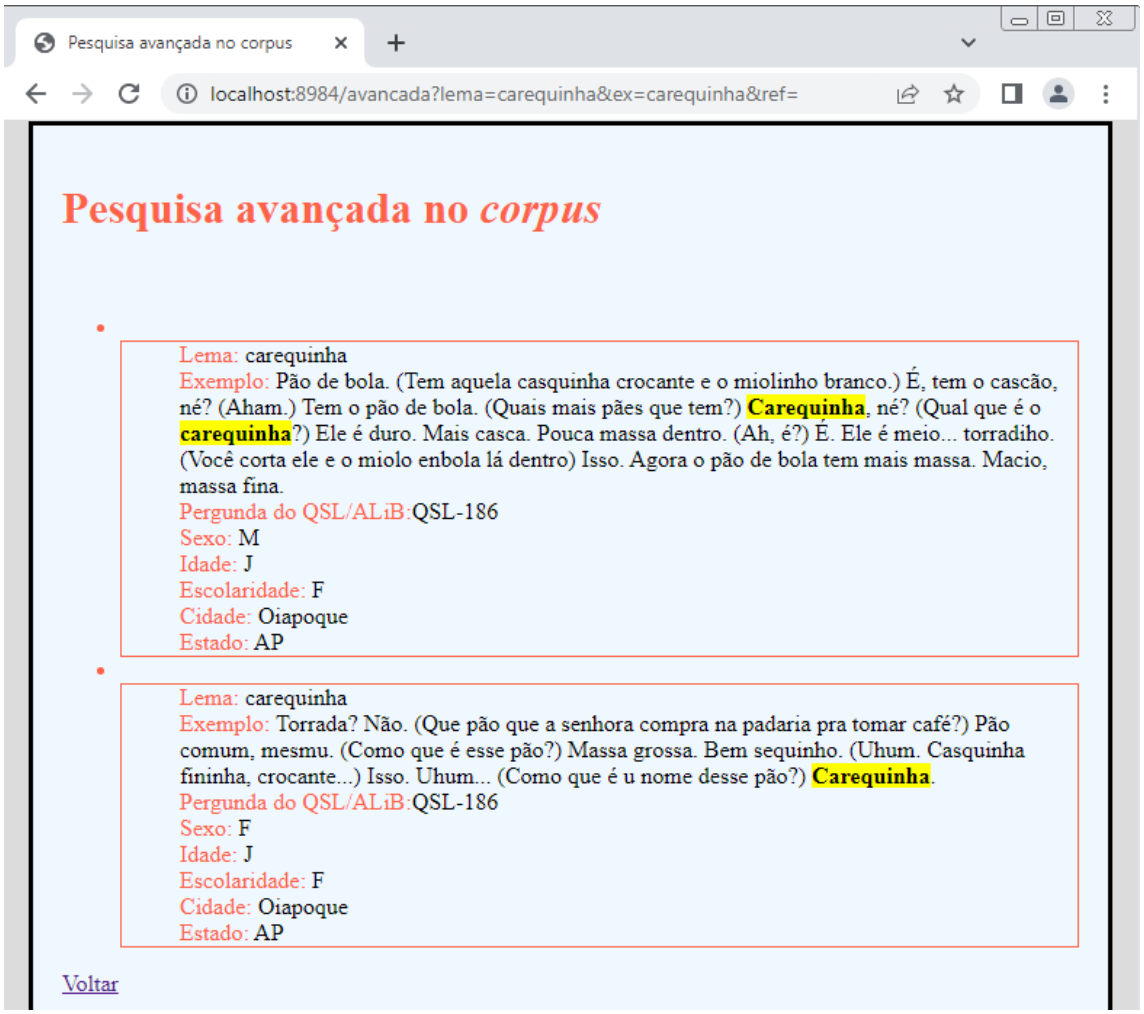
Figura 64: Pesquisa avançada no *corpus* para recuperar o *lema* e o *exemplo* da UL *carequinha*.



Fonte: Protótipo do *VoDiNorte*.

A importância de se digitar a mesma UL na pesquisa por *lema* e na pesquisa por *exemplo*, nesse caso, é para garantir a precisão dos resultados, pois se o usuário utilizar apenas a caixa de pesquisa por *exemplo* a busca se ampliará para todos os possíveis verbetes que tenham a UL *carequinha* registrada na *exemplo* e isso acarretaria possíveis resultados em desconformidade com o sentido que se deseja, isto é, *carequinha* como sinônimo de *pão francês*. Dessa forma, o resultado exibido pelo motor de busca para a solicitação ilustrada na figura 64 pode ser visualizada a seguir:

Figura 65: Resultado da pesquisa avançada no *corpus* para recuperar o *lema* e o *exemplo* da UL *carequinha*.

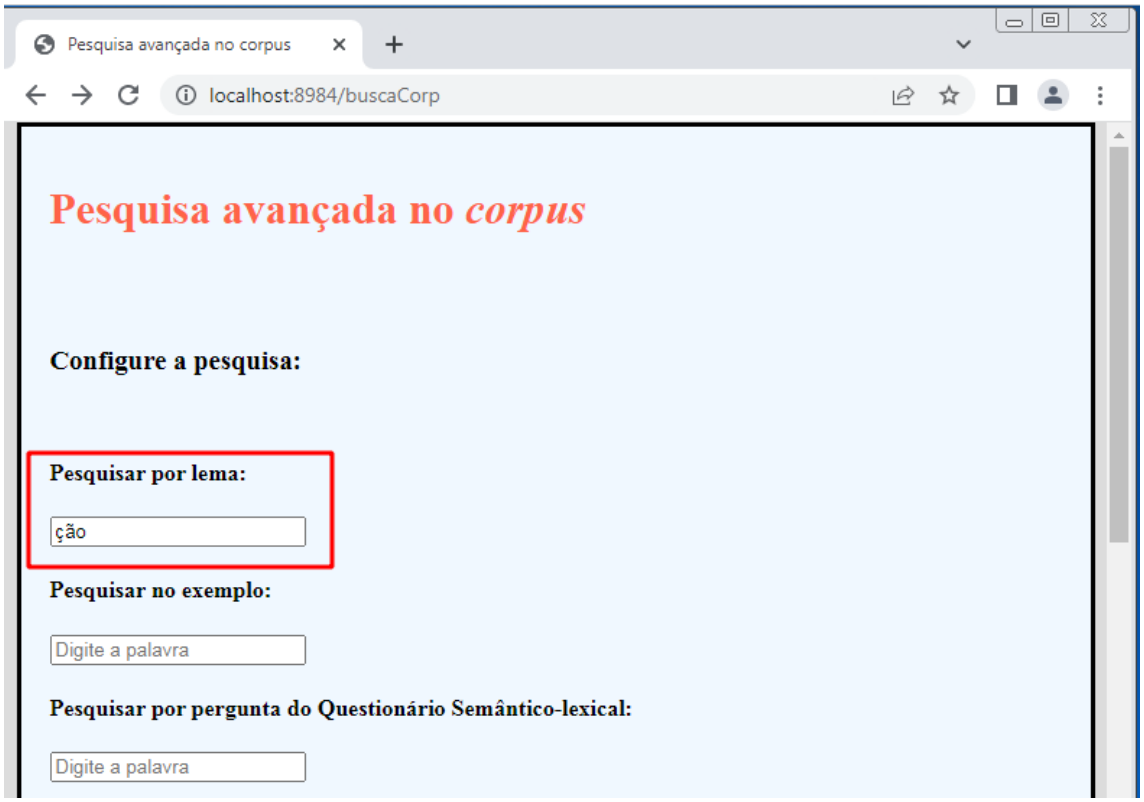


Fonte: Protótipo do *VoDiNorte*.

Como é possível observar nas falas dos informantes, ilustradas na figura 65, a UL *carequinha* foi utilizada para se referir a um tipo de pão que possui uma casca mais grossa, podendo ser equivalente ao pão francês. Vale ressaltar que o termo *carequinha* também remete a dois outros nomes mencionados pelos informantes, ou seja, pão *massa grossa* e *pão de bola*.

Destacamos, ainda, que semelhantemente à pesquisa por meio de um radical é possível realizar buscas a partir de outros caracteres/letras que formam uma UL. Assim, se um usuário desejar, por exemplo, identificar lemas que possuam a sufixação *ção* em sua estrutura morfológica deverá digitar esse sufixo na barra *Pesquisar por lema*, conforme a figura a seguir:

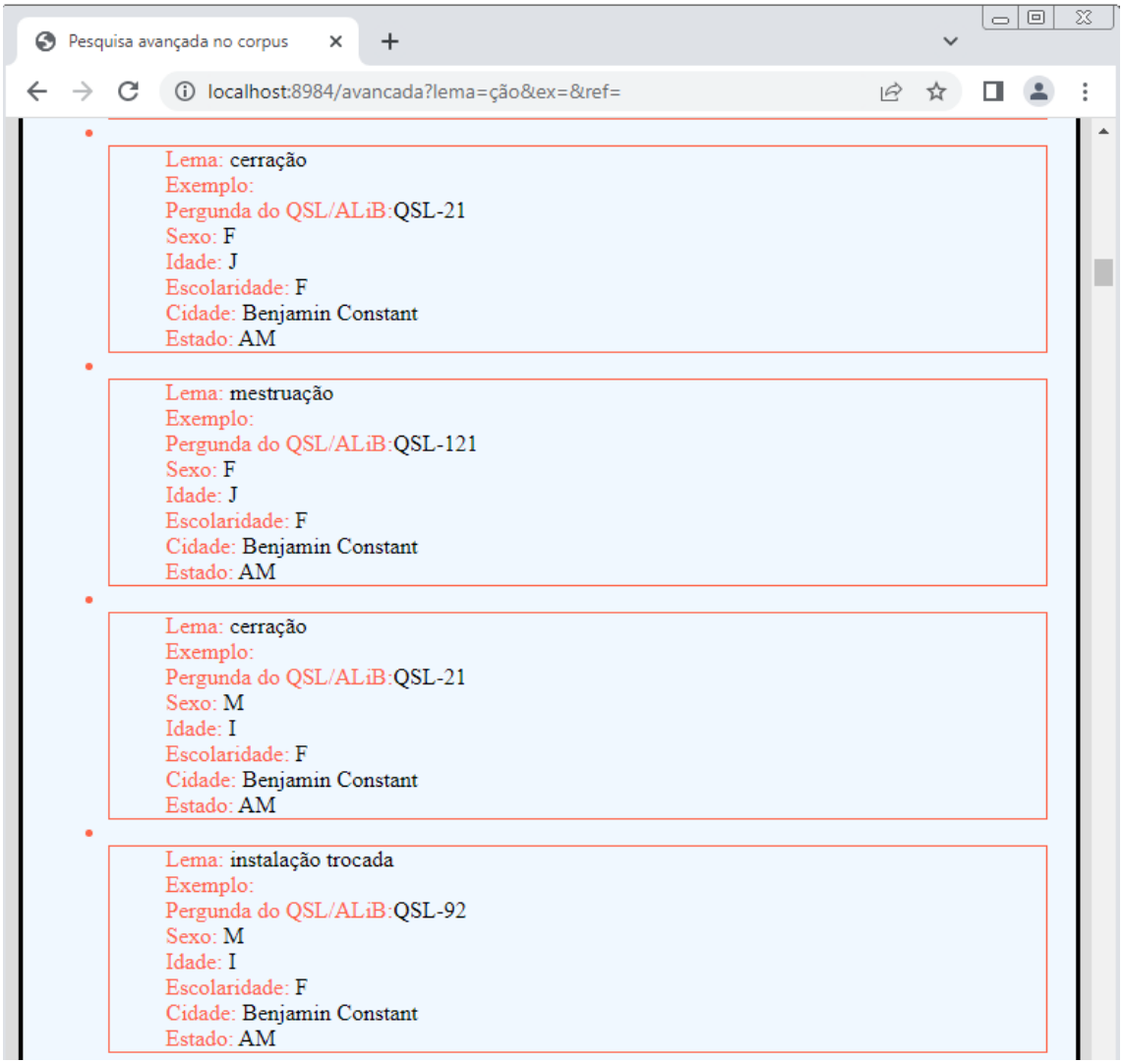
Figura 66: Pesquisa avançada no *corpus* para recuperar lemas com o sufixo *ção*.



Fonte: Protótipo do *VoDiNorte*.

Ao processar a busca os resultados serão exibidos ao usuário que poderá percorrer a lista rolando a página para baixo, a fim de verificar todas as respostas dadas pelos informantes que apresentem o sufixo *ção* que pode estar presente em UL simples ou compostas. Algumas dessas UL podem ser observadas na figura a seguir:

Figura 67: Resultado da pesquisa avançada no *corpus* para recuperar lemas com o sufixo *ção*.



Fonte: Protótipo do *VoDiNorte*.

Em síntese, os exemplos apresentados nessa seção têm como objetivo demonstrar que o motor de pesquisa avançada no *corpus* está operante e pode proporcionar resultados que são de interesse de usuários avançados como, por exemplo, professores universitários e pesquisadores que atuam no ramo da Linguística em geral.

Reiteramos que essa ferramenta representa os últimos avanços alcançados no âmbito desta Tese no que diz respeito ao uso das tecnologias disruptivas para desenvolver obras lexicográficas inovadoras. Certamente há, ainda, um largo caminho a ser trilhado para finalizar as ferramentas propostas para os perfis de usuário comum e intermediário, além de explorar as possibilidades de melhorias que podem ser implementadas no protótipo do *VoDiNorte*.

Dessa forma, apresentamos, na sequência, nas conclusões da pesquisa, reflexões sobre os resultados alcançados, bem como tecemos uma breve discussão sobre a pergunta e a hipótese de pesquisa.

CONCLUSÕES

Dentre os resultados obtidos com o trabalho desenvolvido nesta Tese gostaríamos de destacar um, em especial, que foi o aprendizado relacionado ao uso da tecnologia computacional aplicada à Lexicografia e à Dialetologia, pois a partir dessa experiência será possível trilhar novos caminhos que busquem utilizar as tecnologias disruptivas para o desenvolvimento de produtos inovadores no ramo da Linguística.

Além do conhecimento adquirido também alcançamos resultados satisfatórios como o desenvolvimento das ferramentas de pesquisa que fazem parte do protótipo do *VoDiNorte*. Destacamos, ainda, que durante o percurso metodológico foi possível responder à pergunta de pesquisa, bem como confirmar a hipótese apresentada na Introdução desta Tese.

Como mencionado anteriormente, durante o primeiro ano da pesquisa tínhamos o conhecimento de alguns softwares[88] que poderiam ser utilizados para produzir uma obra lexicográfica e, posteriormente, publicá-la em formato on-line. Porém, além de exigir uma apropriação de conhecimento técnico relacionado ao funcionamento e à integração desses programas, essa forma de trabalho impede o pesquisador de realizar ajustes na execução das tarefas realizadas pelo computador, de modo que o estudioso fica à mercê dos recursos que o software oferece.

Desse modo, nesse momento a pergunta de pesquisa – como dar tratamento informatizado aos dados orais do Projeto ALiB de modo a recuperar informações de maneira automática e desenvolver um produto lexicográfico on-line? – ainda não podia ser satisfatoriamente respondida. Todavia, durante o segundo ano dos estudos, respostas para essa pergunta surgiram nas aulas ministradas pelo professor Dr. Fabrice Issac, da *Universidad Paris Nord.* Trata-se de duas disciplinas relacionadas ao uso da programação aplicada aos estudos linguísticos que foram ofertadas pelo Programa de Pós-Graduação em Estudos de Linguagens/FAALC da UFMS em Campo Grande que, definitivamente, reorientaram a trajetória metodológica desta Tese.

No início ficamos céticos em relação à proposta que o professor Fabrice oferecia aos alunos, que era ensiná-los a desenvolverem suas próprias ferramentas computacionais que seriam fundamentais no curso da pesquisa de cada um. Porém, com

[88] Como, por exemplo, os programas mencionados nas seções *3.3. Linguística de Corpus* e *3.4. Linguística Computacional.*

o passar dos meses, o ceticismo se transformou em oportunidade de adquirir um conhecimento muito útil aos estudos no ramo da Linguística. No entanto, infelizmente, poucos foram os pós-graduandos que souberam aproveitar essa oportunidade.

Assim, foi por intermédio dessas aulas que tivemos acesso ao *XML* e às demais linguagens de programação que fazem parte do protótipo do *VoDiNorte*. Desse modo, o tratamento eletrônico e lexicográfico dos dados do Projeto ALiB começou a ser feito por meio de tentativas de estruturar dados orais em *tags* de um arquivo *XML*. Para tanto, foi preciso elencar o que gostaríamos que o computador fizesse para, posteriormente, estabelecer um modelo de banco de dados que se adequasse aos objetivos do projeto.

À medida que o *XML* foi tomando forma e o aprendizado sobre seu funcionamento foi se consolidando, novos desafios iam surgindo e outros conhecimentos eram construídos até o momento em que iniciamos os testes com as expressões *X-Query*, no editor do software *BaseX*. É neste estágio metodológico que pudemos enxergar o potencial das ferramentas que estávamos construindo, o que nos animou a continuar enfrentando o desafio, pois os percalços encontrados durante o desenvolvimento do protótipo do *VoDiNorte* foram muitos.

Em síntese, com o arquivo *XML* formado e com uma amostra de dados armazenados em suas *tags* foi possível escrever pequenos programas que realizam a filtragem de informações específicas. Assim, a partir desse momento a pergunta de pesquisa havia sido respondida. Além disso, também era possível vislumbrar o funcionamento da aplicação web que, aos poucos, foi sendo lapidada por reformulações em relação ao desenvolvimento das ferramentas de filtragem e apresentação dos dados dialetais tratados lexicograficamente.

Destacamos, ainda, que a partir desse momento a hipótese construída no início desta Tese se confirmou e ficou cada vez mais evidente à medida que os estudos avançavam, pois, de fato, pudemos vivenciar com a construção do protótipo do *VoDiNorte* que investir na aquisição de conhecimentos pontuais relacionados à Ciência da Computação permite ao linguista desenvolver ferramentas computacionais de baixa complexidade dispensando, dessa forma, a contratação de programadores para realizar a automação de determinadas tarefas, além de abrir horizontes para uma metodologia de trabalho interdisciplinar.

É sabido que a busca por conhecimentos específicos sobre programação é um grande desafio principalmente para aqueles que possuem formação na área das Ciências Humanas. No entanto, investir no aprendizado de conteúdos pontuais dentro do campo

da Ciência da Computação é uma atitude que pode gerar resultados satisfatórios aos linguistas em geral.

Destacamos, nesse interim, a importância de iniciativas voltadas para uma maior integração entre os departamentos de Letas e da Ciência da Computação, a fim de aumentar a troca de conhecimentos entre alunos de graduação e de pós-graduação dentro das universidades brasileiras. Um trabalho desse tipo potencializa a consolidação de conhecimentos teóricos e práticos no âmbito acadêmico, que pode resultar em produtos computacionais planejados e desenvolvidos por alunos e pesquisadores em um contexto interdisciplinar entre diferentes áreas do conhecimento.

Um exemplo desse tipo de iniciativa pode ser vista na oferta de disciplinas sobre o uso aplicado de linguagens de programação na construção de soluções computacionais úteis aos estudos linguísticos como, por exemplo, ocorreu em 2020 quando o convênio que a UFMS assinou com a *Univeridad Paris Nord* resultou na oferta de duas disciplinas de programação voltadas para as Ciências Humanas aos alunos de pós-graduação da UFMS.

Essa integração de áreas do conhecimento resulta em bons frutos e esta Tese é prova disso. Desse modo, se queremos que mais graduandos e pós-graduandos façam uso de linguagens de programação no desenvolvimento de seus estudos é preciso investir na implantação de disciplinas relacionadas à Ciência da Computação com interface na Linguística, como, por exemplo, a Linguística de *Corpus,* a Linguística Computacional, o Processamento de Linguagem Natural entre outras, a fim de possibilitar que estudiosos do ramo da Linguística desenvolvam, ao menos em nível elementar, suas próprias ferramentas computacionais em suas pesquisas.

Porém, tendo em vista a baixa oferta de disciplinas relacionadas ao uso da programação aplicada aos estudos linguísticos nos cursos de graduação e pós-graduação em Letras nas Universidades públicas brasileiras, o linguista que desejar apropriar-se de conhecimentos específicos do campo da Ciência da Computação para desenvolver aplicações úteis em seus estudos deverá agir por conta própria, já que a oferta desse tipo de disciplina depende de políticas públicas de educação. Mas, há uma alternativa e, conforme apresentado na confirmação da hipótese desta pesquisa, o esforço da jornada empreendida pelo linguista é compensador, como o demonstrado ao longo deste trabalho.

Em síntese, é possível elencar alguns argumentos que sustentam a hipótese de que investir na aquisição de conhecimentos pontuais relacionados à Ciência da

Computação permite ao linguista desenvolver ferramentas computacionais de baixa complexidade dispensando, dessa forma, a contratação de programadores para realizar a automação de determinadas tarefas, além de abrir horizontes para uma metodologia de trabalho interdisciplinar. Além disso, essa lista poderá auxiliar na tomada de decisão de alguns estudiosos que possam estar interessados em construir um roteiro de estudos relacionados à Informática. Assim, elencamos: i) a possibilidade de editar e melhorar as próprias ferramentas computacionais que, aliás, podem originar novas ferramentas/produtos; ii) a liberdade para testar a compatibilidade dos produtos desenvolvidos com outros sistemas; iii) a economia financeira; iv) a solução para o problema de ruído em relação à comunicação entre programador e cliente que comumente ocorre no desenvolvimento de programas sob encomenda e, v) a possibilidade de transitar em uma área do conhecimento bastante abrangente e em constante desenvolvimento.

Além desses benefícios iniciais, pudemos identificar que ampliar conhecimentos relacionados à Ciência da Computação, a fim de aplicá-los em pesquisas no ramo da Linguística muda a perspectiva do pesquisador definitivamente. Isso significa que o linguista passa a enxergar seu objeto de estudo por ângulos deferentes que antes não eram possíveis de serem focalizados. Assim, a elaboração de metodologias de estudos linguísticos pautadas no uso de ferramentas computacionais próprias resulta em um novo leque de oportunidades para se estudar a língua.

Realizadas essas considerações pontuais acerca da hipótese da pesquisa, na sequência são focalizados mais objetivamente os resultados gerais do estudo. Desta forma, ao analisar a trajetória empreendida nesses quatro anos é possível concluir que o maior amadurecimento teórico-metodológico relacionado à construção de ferramentas computacionais e à elaboração de dicionários genuinamente eletrônicos se deu no último ano do trabalho. Isso porque o acesso a tais conhecimentos ocorreu de forma gradual, impactando esta Tese de duas maneiras.

O primeiro impacto foi o da mudança de paradigmas, ou seja, o entendimento de que no ofício de se elaborar um dicionário eletrônico é preciso desvincular-se do modelo tradicional, usado durante séculos na Lexicografia Impressa para poder enxergar os métodos da Lexicografia Eletrônica que não desqualificam o modelo tradicional, mas estabelecem rigor teórico-metodológico para que os dicionários eletrônicos tenham qualidade e não sejam meros pretextos para se ganhar dinheiro na Internet. Ao seu tempo, o segundo impacto diz respeito ao uso de conhecimentos

pontuais relacionados ao universo da Ciência da Computação para desenvolver as ferramentas específicas para o protótipo do *VoDiNorte*.

Assim, durante os estudos empreendidos no decorrer do Curso de Doutorado, mais especificamente na fase de produção desta Tese, muitos desafios e percalços fizeram parte de uma intensa rotina que resultou em frutos e sementes aqui julgados importantes. Dentre os frutos destacamos: i) o tratamento lexicográfico e eletrônico de dados orais; ii) a criação de uma base de dados em *XML*; iii) o desenvolvimento de ferramentas computacionais personalizadas para a manipulação de dados lexicais; iv) o desenvolvimento de uma aplicação web; v) a criação de diferentes motores de busca a dados tratados lexicograficamente a partir de um perfil de usuário. Dentre as sementes, apresentamos possibilidades de estudos futuros.

No que tange os resultados obtidos, destacamos a *Ferramenta de Pesquisa avançada no corpus* que é acessada por meio do perfil avançado de usuário que, além de estar finalizada e funcional, representa uma inovação do ponto de vista da Lexicografia Eletrônica podendo ser classificada como uma obra lexicográfica do tipo *Model T Fords* (FUERTES-OLIVERA; TARP, 2014, p. 13-16). Como apresentado na seção *5.4 Usuário avançado*, esse motor de busca foi construído pensando no pesquisador que deseja recuperar informações específicas no *corpus* por meio de um conjunto de filtros. Desse modo, os dados dialetais referentes às localidades do interior da região Norte do país, coletados pelo Projeto ALiB, podem ser consultados remotamente e visualizados a partir das variáveis sexo, idade e localidade. Além disso, é possível selecionar uma pergunta do QSL ou fazer buscas nos textos que representam os candidatos à entrada e/ou nos exemplos.

Assim, no que se refere ao protótipo do *VoDiNorte*, a *Ferramenta de pesquisa avançada no corpus* é uma ferramenta pronta para o uso, ao passo que as demais soluções computacionais voltadas para o usuário comum e intermediário estão em construção ou carecem da seleção e organização de dados do *corpus* para que sejam implementadas, pois não foi possível desenvolver todas as funcionalidades planejadas em decorrência da necessidade de cumprimento do prazo regulamentar para o fechamento da Tese.

Destacamos, também, que os frutos alcançados por meio da metodologia empregada no protótipo do *VoDiNorte* podem servir de base para o desenvolvimento de outros projetos no campo da Linguística. Dessa forma, é possível criar arquivos em *XML* estruturando as *tags* de acordo com os objetivos de cada projeto e, posteriormente,

realizar a recuperação automática de informações por meio da construção de expressões *X-Query* no editor do *BaseX*. Em suma, o pesquisador que desejar criar suas próprias ferramentas computacionais de processamento de dados lexicais pode iniciar sua jornada investindo, em um primeiro momento, em três assuntos basilares pertencentes ao universo da programação, a saber:

i) conhecimento da linguagem *XML*;

ii) funcionamento do software *BaseX*;

iii) utilização das expressões *X-Query*.

Esses três ingredientes possibilitam ao pesquisador dar os primeiros passos para o desenvolvimento de soluções informatizadas e inovadoras no âmbito das pesquisas no campo da Linguística.

Acrescentamos, ainda, que o uso de dados em formato *XML* confere compatibilidade com outros sistemas, ou seja, é possível utilizar um mesmo arquivo *XML* para fornecer dados para vários tipos de dicionários eletrônicos.

Sendo assim, a base de dados em *XML* desenvolvida nesta Tese pode contribuir com o Projeto do *Dicionário Dialetal Brasileiro* de forma simples, já que não há impedimentos de ordem técnica que dificultem o acesso às informações armazenadas em cada *tag* do arquivo em *XML* do protótipo do *VoDiNorte*.

Em síntese, além dos frutos explícitos há, também, outro tipo de resultado que não pode ser visto de imediato, pois se trata da experiência que vivenciamos durante esses quatro anos de pesquisa. Desse modo, essa classe de resultado pode ser identificada como um conjunto de sementes que, em tempo oportuno, poderão ser utilizadas em projetos futuros conforme o explicitado a seguir.

Caminhando para a finalização das conclusões propiciadas pela execução da pesquisa pontuamos que a empreitada desta Tese nos mostrou que o *XML* é uma linguagem de marcação versátil e pode ser utilizada como base em projetos diversificados. Assim, quando aprendemos a manipular dados por meio dessa linguagem vemos o potencial dessa ferramenta e ficamos a imaginar sua aplicação na construção de *corpus* com dados terminológicos, fraseológicos, literários, de notícias, entre outros, com a finalidade de desenvolver produtos e realizar estudos de diversas ordens.

Destacamos, também, que uma consultoria com um especialista da Ciência da Computação não deve ser descartada, pois esse profissional pode apontar caminhos a serem trilhados, ou seja, conteúdos específicos a serem desbravados com a finalidade de

melhor desenvolver o projeto, tendo em vista que há muitas possibilidades de se executar uma tarefa usando soluções computacionais.

Em suma, esperamos que a metodologia construída para esta Tese que, conforme o demonstrado, alcançou resultados significativos no que diz respeito à construção de ferramentas computacionais para o processamento de dados lexicais, possa encorajar outros estudiosos que não têm formação na área da Ciência da Computação a traçarem um planejamento de estudos objetivando a criação de suas próprias soluções computacionais no ramo da Linguística.

Outro ponto a ser destacado são os perfis de usuário comum e intermediário que necessitam de mais trabalho. Dessa forma, esse labor representa uma empreitada que se enquadra nas perspectivas de ações futuras, tendo em vista que não foi possível concluir todas as ferramentas de pesquisa lexicográfica do protótipo do *VoDiNorte* pelas razões expostas.

Além disso, outra ação futura que poderá ser incrementada ao protótipo do *VoDiNorte* é a integração de funcionalidades de grandes empresas de tecnologia como, por exemplo, a pesquisa por imagens da *Google* e o *chatGPT*[89] da *OpenAi*. Essas integrações podem ser feitas mediante a configuração de uma API[90] que é disponibilizada por essas empresas, permitindo o funcionamento dessas ferramentas dentro de uma plataforma online. Desse modo, o usuário poderá, por exemplo, acessar imagens geradas pelo *Google Imagens* para determinado verbete ou, ainda, interagir com o *chatGPT* por meio de perguntas relacionadas ao verbete que está sendo pesquisado no protótipo do *VoDiNorte,* a fim de buscar e/ou comparar definições lexicográficas, exemplos de uso, regiões onde uma expressão é utilizada ou qualquer outro tipo de pergunta, tendo em vista que o software foi "treinado" para fornecer respostas para uma grande gama de temas e/ou áreas do conhecimento.

Em síntese, podemos concluir que esta Tese não encerra o assunto, mas dá início a uma nova jornada que pode ser comparada ao ato de subir uma escadaria. O primeiro degrau foi vencido. Agora, faltam os demais!

---

[89] *Chatbot* que interage com o usuário em linguagem natural e pode executar variadas tarefas. Seu uso pode ser feito diretamente pelo site da empresa desenvolvedora ou, ainda, por meio da integração de suas funcionalidades em outros sistemas. Mais informações em: <https://openai.com/>. Acesso em: 10 mar. 2023.

[90] Do inglês *Application Programming Interface* e traduzido para o português como *Interface de Programação de Aplicativo*. Uma *API*, em linhas gerais, é um conjunto de especificações técnicas que conecta um website a um serviço de terceiros, podendo ser gratuito para um tráfego limitado de dados e pago para um fluxo maior de pesquisa.

REFERÊNCIAS

AGUILERA, Vanderci de Andrade. *Atlas Lingüístico do Paraná.* Curitiba: Imprensa Oficial do Estado, 1994.

ALMEIDA, Edilene Maria de Oliveira. *Atlas Linguístico da Mata Sul de Pernambuco-Almaspe*. 2009. 149f. Dissertação (Mestrado em Linguagens e Cultura) – Universidade Federal da Paraíba, João Pessoa, 2009.

AMARAL, Amadeu. *O dialeto caipira*. 2. ed. São Paulo: HUICITEC/Secretaria da Cultura, Ciência e Tecnologia, 1976.

ANTHONY, Laurence. *AntConc.* Versão 3.2.4w. Software de computador. Tokyo, Japan: Waseda University, 2011.

ARAGÃO, Maria do Socorro Silva de; BEZERRA DE MENEZES, Cleusa P. *Atlas Lingüístico da Paraíba*. Brasília: UFPB/CNPq, Coordenação Editorial, 1984; v. 1, 2.

BESSA, José Rogério Fontenele (coordenador). *Atlas Linguístico do Ceará.* Vol.I – Introdução, Vol.II – Cartogramas. Universidade Federal do Ceará. Fortaleza: Edições UFC, 2010.

BAJO PÉREZ, Elena. *Los Diccionarios: Introducción a la lexicografía del español*. Gijón: Ediciones Trea, S.L, 2000, p. 35-47.

BERBER SARDINHA, TONY. *Linguística de Corpus.* Barueri, SP: Manole, 2004.

BÍBLIA SAGRADA. Traduzida em português por João Ferreira de Almeida. Revista eatualizada, 2ª ed. São Paulo: Sociedade Bíblica do Brasil, 1993.

BIDERMAN, Maria Tereza Camargo. *Teoria linguística* Teoria lexical e linguística computacional. 2ª ed. São Paulo: Martins Fontes, 2001.

BIDERMAN, Maria Tereza Camargo. O conhecimento, a terminologia e o dicionário. *Ciência e Cultura.* [online]. 2006, vol. 58, n. 2, p. 35-37. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?pid=S0009-67252006000200014&script=sci_arttext>. Acesso em: 20 dez. 2021.

BRITO, Roseanny Melo de. *Atlas dos falares do baixo Amazonas - AFBAM*. 2011. 297 f. Dissertação (Mestrado em Sociedade e Cultura na Amazônia) – Universidade Federal do Amazonas, Manaus/AM, 2011. Disponível em: <https://tede.ufam.edu.br/handle/tede/2355>. Acesso em: 20 nov. 2021.

BOSQUE, Ignacio. Sobre la teoria de la definición lexicográfica. *Verba*. Anuario Galego de Filoloxía, vol. 9, 1982, p. 105-123.

CARDOSO, Leticia Pinto. *Atlas Linguístico dos Falares de Manaus* – ALFAMA. 2018. 119 f. Dissertação (Mestrado em Letras) – Universidade Federal do Amazonas, Manaus,

2018. Disponível em: <https://tede.ufam.edu.br/handle/tede/6724>. Acesso em: 10 nov. 2021.

CARDOSO, Suzana Alice. *Geolinguística* - tradição e modernidade**.** São Paulo: São Paulo, 2010.

CARDOSO, Suzana Alice Marcelino da Silva et al. *Atlas Linguístico do Brasil*. Vol. 1. Londrina: EDUEL, 2014a.

CARDOSO, Suzana Alice Marcelino da Silva et al. *Atlas Linguístico do Brasil.* Vol. 2. Londrina: EDUEL, 2014b.

CARDOSO, Suzana Alice Marcelino da Silva. *Atlas Lingüístico de Sergipe II*. Rio de Janeiro: S. A. M. da S. Cardoso, 2002. 2v.

CHAMBERS, Jack. y TRUDGILL, Peter. *La dialectología*. Madrid: Visor Libros, S. L., 1994.

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

CORREIA DE SOUZA, Cemary. *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, Salvador/BA, 2019.

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

COSTA, Daniela de Souza Silva. *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.

CRISTIANINI, Adriana Cristina. *Atlas Semântico-Lexical da Região do Grande ABC*. 2007. 772f. Tese (Doutorado – Programa de Pós-Graduação em Lingüística. Área de concentração: Semiótica e Lingüística Geral) – Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo, 2007.

CRUZ, Maria Luiza de Carvalho. *Atlas Lingüístico do Amazonas*. 2004. Tese (Doutorado em letras vernáculas) – Universidade Federal do Rio de Janeiro, Rio de Janeiro/RJ, 2004.

CUBA, M. A. *Atlas Linguístico da Mesorregião Sudeste de Mato Grosso.* Dissertação (Mestrado em Estudo de Linguagens) – Universidade Federal de Mato Grosso do Sul, Campo Grande/MS, 2009.

ENCARNAÇÃO, Márcia Regina Teixeira da. *Atlas semântico-lexical de Caraguatatuba, Ilhabela, São Sebastião e Ubatuba* - municípios do Litoral Norte de São Paulo. 2010. 741f. Tese (Doutorado) – Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo, 2010.

EZQUERRA, Manuel Alvar. 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>. Acesso em: 23 nov. 2020.

FERREIRA, Aurélio Buarque de Holanda. *Dicionário Aurélio*. 5.ª ed., Curitiba: Melhoramentos, 2010.

FERREIRA, Carlota et al. *Atlas Lingüístico de Sergipe.* Salvador: UFBA - Instituto de Letras/Fundação Estadual de Cultura de Sergipe, 1987.

FIGUEIREDO JUNIOR, Selmo Ribeiro. *Atlas linguístico pluridimensional do português paulista*: níveis semântico-lexical e fonético-fonológico do vernáculo da região do Médio Tietê. 2018. 6 t. Tese (Doutorado Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2019.

FUERTES-OLIVERA, Pedro Antonio; TARP, Sven. *Theory and Pratice os Specialised Online Dictionaries. Lexicography versus Terminography*. Berlim/Boston: Gruyer, 2014.

FUERTES-OLIVERA, Pedro Antonio; BERGENHOLTZ, Henning. Introduction: the construction of internet dictionaries. In: FUERTES-OLIVERA, Pedro Antonio; BERGENHOLTZ, Henning. *e-Lexicography:* The Internet, Digital Initiative and Lexicography. London/New York: Continuum, 2011, p. 3-16.

GRANGER, Sylviane. Introduction: Electronic lexicography – from challenge to opportunity. In: GRANDER, Silviane; PAQUOT, Magali. *Eletronic Lexicography.* United Kindom: Oxford University Press, 2012, p. 1-11.

GRANJA, María Álvarez de la; SEOANE, Ernesto González. El tratamiento lexicográfico del léxico dialectal. In: GRANJA, María Álvarez de la; SEOANE, Ernesto González. (eds.) *Léxico dialectal lexicografía en la Iberorromania*. Madrid/Frankfurt am Maim: Iberoamericana/Vervuert, 2018, p. 9-21.

GRÜN, Chistian. *BaseX*. Versão 8.5.1. Software de computador. Earlangen, Alemanha: BaseX Team, 2016. Disponível em : <https://files.basex.org/releases/8.5.1/>. Acesso em: 16 abr. 2023.

HABERT, Benoît. Outiller la linguistique: de l'emprunt de techniques aux rencontres de savoirs. *Revue Française de Linguistique Appliquée*, [*s.l.*], Vol. IX, p. 5-24, 2004. Disponível em: <https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2004- 1-page-5.htm?contenu=article> . Acesso em: 14 dez. 2020.

HABERT, Benoît. Portrait de linguiste(s) à I'instrument. *Texto!* [en ligne], décembre 2005, vol. X, nº 4. Disponível em: <http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html#simondon05>.Acesso em: 15 jan 2021.

HAENSCH, Günther. *Los diccionarios del español en el umbral del siglo XXI***.** Salamanca: Ediciones Universidad de Salamanca, 1997.

HARTMANN, Reinhard Rudolf Karl. *Teaching and Researching Lexicography*. London, New York: Routledge, 2001. 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: 23 set. 2021.

HARTMANN, Reinhard Rudolf Karl; JAMES, Gregory. *Dictionary of Lexicography.* London, New York: Routledge, 1998.

HOUAISS, Antônio, *Dicionário Eletrônico da Língua Portuguesa.* Rio de Janeiro: Objetiva, 2009. CD-ROM.

KARLBERG, Luísa Galvão Lessa. *Atlas etnolinguístico do Acre*. Rio Branco: Edufac, 2018. Disponível em: <http://www2.ufac.br/editora/livros/atlasetnolinguisticodoacre.pdf>. Acesso em: 22 nov. 2021.

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.

KOCH, Walter; Klassmann, Mário Silfredo; ALTENHOFEN, Cléo. *Atlas Lingüístico-etnográfico da Região Sul do Brasil*. Porto Alegre/Florianópolis/Curitiba: Ed. UFRGS/Ed. UFSC/ Ed. UFPR, 2002. v. 1, v. 2.

MAIA, Edson Galvão. *Atlas Linguístico do Sul Amazonense (ALSAM)*. 2018, 845f. Tese (Doutorado em Estudos da Linguagem), Londrina, Universidade Estadual de Londrina – UEL, 2018.

MARROQUIM, Mário. *A língua do Nordeste*. 3. ed. Curitiba: HD Livros, 1996.

LAGORIO, Consuelo Alfaro; FREIRE, José Ribamar Bessa. Aryon Rodrigues e as Línguas Gerais na historiografia linguística. *DELTA*, São Paulo, v. 30 especial, p. 571-589, 2014. Disponível em: <https://www.scielo.br/j/delta/a/WnSPssPLq4TMX59swnwmJbs/?lang=pt> Acesso em: 05 mar. 2023. <https://doi.org/10.1590/0102-445008157345493422>.

LEROYER, Patrick. Change of paradigm: from Linguistics to Information Science and from dictionaries to lexicography information tolls. In: FUERTES-OLIVERA, Pedro Antonio; BERGENHOLTZ, Henning. *e-Lexicography:* The Internet, Digital Initiative and Lexicography. London/New York: Continuum, 2011. p. 121-140.

LEW, Robert; DE SCHRYVER, Gilles-Maurice. Dictionary users in the digital revolution. Oxford: *International Journal of Lexicography*, v. 27, Issue 4, December 2014, p. 341–359. Disponível em: <https://academic.oup.com/ijl/article-abstract/27/4/341/932743#no-access-message>. Acesso em: 26 set.2022.

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* (UFBA), v. 41, p. 49-70, 2010.

MAIA, Edson Galvão. *Atlas Linguístico do Sul Amazonense (ALSAM)*. 2018, 845f. Tese (Doutorado em Estudos da Linguagem) – Universidade Estadual de Londrina, UEL, Londrina/PR, 2018.

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

McENERY, Tony; HARDIE, Andrew. *Corpus Linguistics: Method, Theory and Practice*. United Kingdom: Cambridge University Press, 2012.

MEDINA GUERRA, Antonia María. La microestructura del diccionario: la definición. In: _____. (coord.). *Lexicografía española.* Barcelona: Ariel Lingüística, 2003, p. 127-146.

MICHAELIS. *Dicionário Escolar Língua Portuguesa*. São Paulo: Melhoramentos, 2015.

NASCENTES, Antenor. *O linguajar carioca*. 2. ed. Rio de Janeiro: Organização Simões, 1953.

NAVARRO CARRASCO, Ana Isabel. 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.

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

NIELSEN, Sandro; Pedro Antonio. FUERTES-OLIVERA. Development in Lexicography: From Polyfunctional to Monofunctional Accounting Dictionaries. *Lexikos*. [*S.l.*], v. 23 n. 1, p. 323-347, dez., 2013. Disponível em: <https://lexikos.journals.ac.za/pub/article/view/1218>. Acesso em: 31 ago. 2022.

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.

OLIVEIRA, Ana Pinto Pires de. Regionalismos Brasileiros: a questão da distribuição geográfica. In: OLIVEIRA, Ana Pinto Pires de; ISQUERDO, Aparecida Negri. (Orgs) *As Ciências do Léxico.* Lexicologia, Lexicografia, Terminologia. 2ª ed., Campo Grande: Editora UFMS, 2001, p. 109-115.

OLIVEIRA, Aniele Souza de. *Léxico brasileiro em dicionários monolíngues e bilíngues*: estudo metalexicográfico da variação em perspectiva dialetal e histórica. 2017, 354 f. Tese (Doutorado em Língua e Cultura) – Universidade Federal da Bahia, Salvador, BA, 2017.

OLIVEIRA, Dercir. Pedro de (Org.). *ALMS - Atlas Lingüístico de Mato Grosso do Sul*.

1. ed. Campo Grande: Editora UFMS, 2007.

PEREIRA, Maria das Neves. *Atlas geolinguístico do litoral potiguar.* Tese (Doutorado em Letras Vernáculas) – Universidade Federal do Rio de Janeiro, Faculdade de Letras, 2007. 2v. Vol I: 123 p., Vol II: 189 p.

PEREIRA, Renato Rodrigues. Estrutura Lexicográfica. In: PEREIRA, Renato Rodrigues. *O dicionário pedagógico e a homonímia:* em busca de parâmetros didáticos. Tese (Doutorado em Linguística e Língua Portuguesa) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Araraquara, 2018, p. 39-50.

PORTO DAPENA, José Álvaro. *Manual de técnica lexicográfica.* Madrid: ARCO/LIBROS, S.A., 2002.

RAZKY, Abdelhak. (Org.) *Atlas lingüístico sonoro do Pará*. Belém: PA/CAPES/UTM, 2004. CD-Room.

RAZKY, Abdelhak; RIBEIRO, Celeste Maria da Rocha; SANCHES, Romário Duarte. *Atlas Linguístico do Amapá.* São Paulo: Labrador, 2017.

REY DEBOVE, Josette. Du dictionnaire au dictionnaire de langue. In: *Etute linguistique et sémiotique des dictionnaires français contemporains.* Paris: Monton, 1971, p. 19-37. Disponívelem: <https://books.google.com.br/books?id=o89aimry0ToC&pg=PA19&hl=pt-BR&source=gbs_toc_r&cad=4#v=onepage&q&f=false >. Acesso em: 20 set. 2021.

RIBEIRO, José et. al. *Esboço de um Atlas Lingüístico de Minas Gerais.* v. 1. Rio de Janeiro: Fundação Casa de Rui Barbosa; Universidade Federal de Juiz de Fora, 1977.

RODRÍGUEZ BARCIA, Suzana. *Introducción a la lexicografia.* Madrid: Síntesis, 2016.

ROMANO, Valter Pereira. *Atlas Geossolinguístico de Londrina:* um estudo em tempo real e tempo aparente. 2012. 366f. Dissertação (Mestrado em Estudos da Linguagem) – Universidade Estadual de Londrina, Londrina, 2012.

ROSSI, Nelson. *Atlas Prévio dos Falares Baianos*. Rio de Janeiro: INL, 1963.

SÁ, Edmilson José 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.

SÁ, Talita Rodrigues de. *Pelos caminhos da cartografia linguística paraense*: um estudo semântico-lexical do Distrito Mosqueiro numa perspectiva socioeducacional. 2013. 282 f. Dissertação (Mestrado em Educação) – Universidade do Estado do Pará, Belém, 2013. Disponível em: <https://ccse.uepa.br/ppged/wp-content/uploads/dissertacoes/07/talita_rodrigeus_de_sa.pdf>. Acesso em: 22 nov. 2021.

SAGER, Juan C. *Essays on definition:* Terminology and lexicography research and practice. Vol 4, John Benjamins Publishing Company: Amsterdam, Philadelphia, 2000.

SANTOS JUNIOR, Jorge Luiz Nunes dos. *Glossário de termos da Agricultura:* um estudo terminológico sobre o manejo do solo. 2015. 191 f. Dissertação (Mestrado em Estudos de Linguagens) – Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, 2015.

SANTOS JUNIOR, Jorge Luiz Nunes; ISQUERDO, Aparecida Negri. Você sabe fazer farinha? [Sei.] Me ensina? O léxico da mandioca na região Norte do Brasil. *Sociodialeto*, [*S.l.*]: NUPESD/UEMS, v. 13, n. 38, 2022, no prelo.

SAPIR, Edward. Língua e ambiente. In: SAPIR, Edward. *Lingüística como ciência*. (Textos organizados por J. Mattoso Câmara). Rio de Janeiro: Livraria Acadêmica, 1969, p. 43-62.

SEABRA, Maria Cândida Trindade Costa de. Língua, Cultura, Léxico. In: SOBRAL, Gilberto Nazareno Telles; LOPES, Norma da Silva; RAMOS, Jânia Martins. *Linguagem, Sociedade e Discurso*. São Paulo: Blucher, 2015, p. 65-84.

SECO, Manuel. *Estudios de Lexicografía Española.* Madri: Editorial Credos/ Biblioteca Románica Hispánica, 2ª ed, 2003.

SEPULVEDA, Susana Serra. Gramática y diccionario. El problema del *contorno* en lexicografía española**.** *Boletín de filología*, [*S.l.: s.n.*], tomo XLI, p. 197-240, 2006. Disponível em: <https://core.ac.uk/download/pdf/46545294.pdf>. Acesso em: 10 out. 2019.

SILVA, Greize Alves da. *Atlas linguístico topodinâmico e topoestático do estado do Tocantins (ALITTETO).* 2018. 394 f. Tese (Doutorado em Estudos da Linguagem) – Universidade Estadual de Londrina, Londrina, PR, 2018. Disponível em: <http://www.bibliotecadigital.uel.br/document/?code=vtls000218332>. Acesso em: 20 nov. 2020.

SINCLAIR, John. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991.

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. [*S.l.*]: *Lexikos,* v. 29, 2019, p. 224-249. Disponível em: <http://lexikos.journals.ac.za>. Acesso em: 20 ago. 2022.

TARP, Sven. The concept of dictionary. In: FUERTES-OLIVERA, Pedro Antonio. *The Routledge Handbook of Lexicography*. London, New York: Routledge, 2018, p. 237-249.

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

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

TARP, Sven. *Lexicography in the Borderland between Knowledge and Non-knowledge*: General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Niemeyer, 2008.

ANEXOS

Anexo 1 – Questionário Semântico-lexical (COMITÊ NACIONAL ..., 2001, p. 21-38)

ACIDENTES GEOGRÁFICOS

1. CÓRREGO/ RIACHO

... um rio pequeno, de uns dois metros de largura?

2. PINGUELA

... tronco, pedaço de pau ou tábua que serve para passar por cima de um _____ (cf. item 1)?

3. FOZ

... o lugar onde o rio termina ou encontra com outro rio?

4. REDEMOINHO (DE ÁGUA)

Muitas vezes, num rio, a água começa a girar, formando um buraco, na água, que puxa para baixo. Como se chama isto?

5. ONDA DE MAR

... o movimento da água do mar? Imitar o balanço das águas.

6. ONDA DE RIO

... o movimento da água do rio? Idem item 5.

FENÔMENOS ATMOSFÉRICOS

7. REDEMOINHO (DO VENTO)

...o vento que vai virando em roda e levanta poeira, folhas e outras coisas leves?

8. RELÂMPAGO

... um clarão que surge no céu em dias de chuva?

9. RAIO

...uma luz forte e rápida que sai das nuvens, podendo queimar uma árvore, matar pessoas e animais, em dias de mau tempo?

10. TROVÃO

... o barulho forte que se escuta logo depois de um _____ (cf. item 9)?

11. TEMPORAL / TEMPESTADE / VENDAVAL

... uma chuva com vento forte que vem de repente?

12. NOMES ESPECÍFICOS PARA TEMPORAL

Existem outros nomes para _____ (cf. item 11)?

13. TROMBA D'ÁGUA

... uma chuva de pouca duração, muito forte e pesada?

14. CHUVA FORTE

... urna chuva forte e contínua?

15. CHUVA DE PEDRA

Durante urna chuva, podem cair bolinhas de gelo. Como chamam essa chuva?

16. ESTIAR / COMPOR O TEMPO

Como dizem aqui quando termina a chuva e o sol começa a aparecer?

17. ARCO-ÍRIS

"Quase sempre, depois de uma chuva, aparece no céu uma faixa com listras coloridas e curvas (mímica). Que nomes dão a essa faixa?"

18. GAROA

... uma chuva bem fininha?

19. TERRA UMEDECIDA PELA CHUVA

"Depois de uma chuva bem fininha, quando a terra não fica nem seca, nem molhada, corno é que se diz que a terra fica?"

20. ORVALHO / SERENO

De manhã cedo, a grama geralmente está molhada. Como chamam aquilo que molha a grama?

21. VOEIRO / CERRAÇÃO / NEBLINA

Muitas vezes, principalmente de manhã cedo, quase não se pode enxergar por causa de uma coisa parecida com fumaça, que cobre tudo. Corno chamam isso?

ASTROS E TEMPO

22. AMANHECER

... a parte do dia quando começa a clarear?

23. NASCER (DO SOL)

O que é que acontece no céu de manhã cedo quando começa a clarear?

24. ALVORADA

...a claridade avermelhada do céu antes de _____ (cf. item 23)?

25. PÔR (DO SOL)

E o que acontece no céu no final da tarde?

26. CREPÚSCULO

... a claridade avermelhada que fica no céu depois do _____ (cf. item 25)?

27. ENTARDECER

E quando o sol se põe?

28. ANOITECER

... o começo da noite?

29. ESTRELA MATUTINA / VÊNUS / ESTRELA DA MANHÃ / ESTRELA-D'ALVA

"De manhã cedo, urna estrela brilha mais e é a última a desaparecer. Como chamam esta estrela?"

30. ESTRELA VESPERTINA / VÊNUS / ESTRELA DA TARDE

"De tardezinha, urna estrela aparece antes das outras, perto do horizonte, e brilha mais." Como chamam esta estrela?

31. ESTRELA CADENTE / ESTRELA FILANTE / METEORO / ZELAÇÃO

De noite, muitas vezes pode-se observar urna estrela que se desloca no céu, assim, (mímica) e faz um risco de luz. Como chamam isso?

32. MUDAR / CORRER UMA ESTRELA

"E quando se vê uma _____ (cf. item 31), corno é que se diz?"

IDENTIFICAR OS VERBOS USADOS PARA EX PRESSAR O MOVIM ENTO DA ESTRELA CADENTE.

33. VIA LÁCTEA / CAMINHO DE SANTIAGO

Numa noite bem estrelada, aparece uma banda ou faixa que fica no céu de fora a fora, onde tem muitas estrelas muito perto umas das outras. Como chamam esta banda ou faixa?

34. MESES DO ANO

Quais são os meses do ano?

35. MESES COM NOMES ESPECIAIS

Alguns desses meses têm outro nome, por exemplo, junho, julho, etc.?

36. ONTEM

Hoje é segunda-feira. E domingo, que dia foi?

37. ANTEONTEM

... o dia que foi antes desse dia? [E um dia para trás?]

38. TRASANTEONTEM
... o dia que foi antes de _____ (cf. item 37)? [E mais um dia para trás?]

ATIVIDADES AGROPASTORIS

39. TANGERINA / MEXERICA
... as frutas menores que a laranja, que se descascam com a mão, e, normalmente, deixam um cheiro na mão? Como elas são?
PEDIR PARA DESCREVER, PARA APURAR AS DIFERENÇAS ENTRE AS DESIGNAÇÕES CITADAS PELO INFORMANTE.

40. AMENDOIM
... o grão coberto por uma casquinha dura, que se come assado, cozido, torrado ou moído?

41. CAMOMILA
... urnas florezinhas brancas com miolo amarelinho, ou florezinhas secas que se compram na farmácia ou no supermercado e servem para fazer um chá amarelinho, cheiroso, bom para dor de barriga de nenê/bebê e até de adulto e também para acalmar? Mostrar.

42. PENCA
... cada parte que se corta do cacho da bananeira para pôr para madurar / amadurecer?

43. BANANA DUPLA / FELIPE / GÊMEAS
... duas bananas que nascem grudadas?

44. PARTE TERMINAL DA INFLORESCÊNCIA DA BANANEIRA / UMBIGO / CORAÇÃO
... a ponta roxa no cacho da banana?

45. ESPIGA

Quando se vai colher o milho, o que é que se tira do pé? [Quando se vai à feira comprar milho, compra-se o quê?]

46. SABUGO

Quando se tira da _____ (cf. item 45) todos os grãos do milho, o que sobra?

47. SOCA / TOUCEIRA

Depois que se corta o pé de arroz ou de fumo, ainda fica uma pequena parte no chão. Como se chama essa parte?

48. GIRASSOL

... flor grande, amarela, redonda, com uma rõdela de sementes no meio?

49. VAGEM DO FEIJÃO/BAINHA

Onde é que ficam os grãos do feijão, no pé, antes de serem colhidos?

50. MANDIOCA/AIPIM

... aquela raiz branca por dentro, coberta por uma casca marrom, que se cozinha para comer?

51. MANDIOCA

...uma raiz parecida com _____ (cf. item 50) que não serve para comer e se rala para fazer farinha (polvilho, goma)?

52. CARRINHO DE MÃO / CARRIOLA

... um veículo de uma roda, empurrado por uma pessoa, para pequenas cargas em trechos curtos?

53. HASTES DO CARRINHO DE MÃO

... as duas partes em que a pessoa segura para empurrar o (a) _____ (cf . item 52)?

54. CANGALHA / FORQUILHA

"... a armação de madeira, que se coloca no pescoço de animais (porco, terneiro / bezerro, carneiro, vaca), para não atravessarem a cerca?"

55. CANGALHA

... a armação de madeira que se coloca no lombo do cavalo ou do burro para levar cestos ou cargas? *Mostrar gravura*.

56. CANGA

... a peça de madeira que vai no pescoço do boi, para puxar o carro ou o arado? *Mostrar gravura*.

57. JACÁ / BALAIO

... aqueles objetos de vime, de taquara, de cipós trançado(s), para levar batatas (mandioca, macaxeira, aipim, etc.), no lombo do cavalo ou do burro?

58. BOLSA / BRUACA

E quando se usam objetos de couro, com tampa, para levar farinha, no lombo do cavalo ou do burro? *Mostrar gravura*.

59. BORREGO (DO NASCER ATÉ...)

... a cria da ovelha logo que nasce? E até que idade se dá esse nome?

60. PERDA DA CRIA

Como se diz quando a fêmea de um animal perde a cria?

61. TRABALHADOR DE ENXADA EM ROÇA ALHEIA

... o homem que é contratado para trabalhar na roça de outro, que recebe por dia de trabalho?

62. PICADA / ATALHO ESTREITO

O que é que se abre com o facão, a foice para passar por um mato fechado?

63. TRILHO / CAMINHO / VEREDA / TRILHA

... o caminho no pasto, onde não cresce mais grama, de tanto o animal ou o homem passarem por ali?

FAUNA

64. URUBU
... a ave preta que come animal morto, podre?

65. COLIBRI / BEIJA-FLOR
... o passarinho bem pequeno, que bate muito rápido as asas, tem o bico comprido e fica parado no ar?

66. JOÃO-DE-BARRO
... a ave que faz a casa com terra, nos postes, nas árvores e até nos cantos da casa?

67. GALINHA-D'ANGOLA / GUINÉ / COCAR
...a ave de criação parecida com a galinha, de penas pretas com pintinhas brancas?

68. PAPAGAIO
... a ave de penas coloridas que, quando presa, pode aprender a falar?

69. SURA
... uma galinha sem rabo?

70. COTÓ
... um cachorro de rabo cortado?

71. GAMBÁ
... o bicho que solta um cheiro ruim quando se sente ameaçado?

72. PATAS DIANTEIRAS DO CAVALO
... as patas dianteiras do cavalo?

73. CRINA DO PESCOÇO

... o cabelo em cima do pescoço do cavalo?

74. CRINA DA CAUDA

... o cabelo comprido na traseira do cavalo?

75. LOMBO

... a parte do cavalo onde vai a sela?

76. ANCA / GARUPA / CADEIRA

... a parte larga atrás do _____ (cf. item 75)?

77. CHIFRE

O que o boi tem na cabeça?

78. BOI SEM CHIFRE

... o boi sem _____ (cf. item 77)?

79. CABRA SEM CHIFRE

... a cabra que não tem _____ (cf. item 77)?

80. ÚBERE

Em que parte da vaca fica o leite?

81. RABO

... a parte com que o boi espanta as moscas?

82. MANCO

... o animal que tem urna perna mais curta e que puxa de urna perna?

83. MOSCA VAREJEIRA

... um tipo de mosca grande, esverdeada, que faz um barulhão quando voa?

84. SANGUESSUGA

... um bichinho que se gruda nas pernas das pessoas quando elas entram num córrego ou banhado (cf. item 1)?

85. LIBÉLULA

...o inseto de corpo comprido e fino, com quatro asas bem transparentes, que voa e bate a parte traseira na água?

86. BICHO DE FRUTA

... aquele bichinho branco, enrugadinho, que dá em goiaba, em coco?

87. CORÓ

... aquele bicho que dá em esterco, em pau podre?

88. PERNILONGO

...aquele inseto pequeno, de peninhas compridas, que canta no ouvido das pessoas, de noite? *Imitar o zumbido*.

CORPO HUMANO

89. PÁLPEBRAS / CAPELA DOS OLHOS

... esta parte que cobre o olho? *Apontar*.

90. CISCO

... alguma coisinha que cai no olho e fica incomodando?

91. CEGO DE UM OLHO

... a pessoa que só enxerga com um olho?

92. VESGO

... a pessoa que tem os olhos voltados para direções diferentes? *Completar com um gesto dos dedos*.

93. MÍOPE

... a pessoa que não enxerga longe, e tem que usar óculos?

94. TERÇOL / VIÚVA

...a bolinha que nasce na _____ (cf. item 89), fica vermelha e incha?

95. CONJUNTIVITE / DOR D'OLHOS

...a inflamação no olho que faz com que o olho fique vermelho e amanheça grudado?

96. CATARATA

...aquela pele branca no olho que dá em pessoas mais idosas?

97. DENTES CANINOS / PRESAS

...esses dois dentes pontudos? *Apontar*.

98. DENTES DO SISO / DO JUÍZO

... os últimos dentes, que nascem depois de todos os outros, em geral quando a pessoa já é adulta?

99. DENTES MOLARES / DENTE QUEIRO

... esses dentes grandes no fundo da boca, vizinhos dos _____ (cf. item 98)? *Apontar*.

100. DESDENTADO / BANGUELA

... a pessoa que não tem dentes?

101. FANHOSO / FANHO

... a pessoa que parece falar pelo nariz? *Imitar*.

102. MELECA / TATU

... a sujeirinha dura que se tira do nariz com o dedo?

103. SOLUÇO

... este barulhinho que se faz? Soluçar.

104. NUCA

... isto? *Apontar* .

105. POMO-DE-ADÃO / GOGÓ

... esta parte alta do pescoço do homem? *Apontar*.

106. CLAVÍCULA

... o osso que vai do pescoço até o ombro? *Apontar*

107. CORCUNDA

... a pessoa que tem um calombo grande nas costas e fica assim (*mímica*)?

108. AXILA

... esta parte aqui? *Apontar*.

109. CHEIRO NAS AXILAS

... o mau cheiro embaixo dos braços?

110. CANHOTO

...a pessoa que come com a mão esquerda, faz tudo com essa mão? *Completar com o gesto*.

111. SEIOS / PEITO

... a parte do corpo da mulher com que ela amamenta os filhos?

112. VOMITAR

Se uma pessoa come muito e sente que vai pôr /botar para fora o que comeu, se diz que vai o quê?

113. ÚTERO

... a parte do corpo da mãe onde fica o nenê / bebê antes de nascer?

114. PERNETA

... a pessoa que não tem uma perna?

115. MANCO

... a pessoa que puxa de uma perna?

116. PESSOA DE PERNAS ARQUEADAS

... a pessoa de pernas curvas? *Mímica*.

117. RÓTULA / PATACA

... o osso redondo que fica na frente do joelho?

118. TORNOZELO

... isto? Apontar.

119. CALCANHAR

... isto? *Apontar*.

120. CÓCEGAS

Que sente uma criança quando se passa o dedo na sola do pé? *Mímica*.

CICLOS DA VIDA

·121. MENSTRUAÇÃO

As mulheres perdem sangue todos os meses. Como se chama isso?

122. ENTRAR NA MENOPAUSA

Numa certa idade acaba a/o _____ (cf. item 121). Quando isso acontece, se diz que a mulher

123. PARTEIRA

... a mulher que ajuda a criança a nascer?

124. DAR À LUZ

Chama-se a _____ (cf. item 123) quando a mulher está para _____.

125. GÊMEOS

... duas crianças que nasceram no mesmo parto?

126. ABORTO
Quando a mulher grávida perde o filho, se diz que ela teve _____.

127. ABORTAR
Quando a mulher fica grávida e, por algum motivo, não chega a ter a criança, se diz que ela _____.

128. AMA-DE-LEITE
Quando a mãe não tem leite e outra mulher amamenta a criança, como chamam essa mulher?

129. IRMÃO DE LEITE
O próprio filho da _____ (cf. item 128) e a criança que ela amamenta são o quê um do outro?

130. FILHO ADOTIVO
... a criança que não é filho verdadeiro do casal, mas que é criada por ele como se fosse?

131. FILHO MAIS MOÇO / CAÇULA
... o filho que nasceu por último?

132. MENINO / GURI / PIÁ
Criança pequenininha, a gente diz que é bebê. E quando ela tem de 5 a 10 anos, do sexo masculino?

133. MENINA
E se for do sexo feminino, como se chama?

134. MADRASTA
Quando um homem fica viúvo e casa de novo, o que a segunda mulher é dos filhos que ele já tinha?

135. FINADO / FALECIDO

Numa conversa, para falar de uma pessoa que já morreu, geralmente as pessoas não a tratam pelo nome que tinha em vida. Como é que se referem a ela?

CONVÍVIO E COMPORTAMENTO SOCIAL

136. PESSOA TAGARELA

... a pessoa que fala demais?

137. PESSOA POUCO INTELIGENTE

... a pessoa que tem dificuldade de aprender as coisas?

138. PESSOA SOVINA

... a pessoa que não gosta de gastar seu dinheiro e, às vezes, até passa dificuldades para não gastar?

139. MAU PAGADOR

... a pessoa que deixa suas contas penduradas?

140. ASSASSINO PAGO

... a pessoa que é paga para matar alguém?

141. MARIDO ENGANADO

... o marido que a mulher passa para trás com outro homem?

142. PROSTITUTA

... a mulher que se vende para qualquer homem?

143. XARÁ

... a pessoa que tem o mesmo nome da gente?

144. BÊBADO (DESIGNAÇÕES)

Que nomes dão a uma pessoa que bebeu demais?

145. CIGARRO DE PALHA

Que nomes dão ao cigarro que as pessoas faziam antigamente, enrolado à mão?

146. TOCO DE CIGARRO

...o resto do cigarro que se joga fora?

RELIGIÃO E CRENÇAS

147. DIABO

Deus está no céu e no inferno está _____.

148. FANTASMA

O que algumas pessoas dizem já ter visto, à noite, em cemitérios ou em casas, que se diz que é do outro mundo?

149. FEITIÇO

O que certas pessoas fazem para prejudicar alguém e botam, por exemplo, nas encruzilhadas?

150. AMULETO

... o objeto que algumas pessoas usam para dar sorte ou afastar males?

151. BENZEDEIRA

... uma mulher que tira o mau-olhado com rezas, geralmente com galho de planta?

152. CURANDEIRO

... a pessoa que trata de doenças através de ervas e plantas?

153. MEDALHA

... a chapinha de metal com um desenho de santo que as pessoas usam, geralmente no pescoço, presa numa corrente?

154. PRESÉPIO

No Natal, monta-se um grupo de figuras representando o nascimento do Menino Jesus. Como chamam isso?"

JOGOS E DIVERSÕES INFANTIS

· 155. CAMBALHOTA

... a brincadeira em que se gira o corpo sobre a cabeça e acaba sentado? *Mímica.*

156. BOLINHA DE GUDE

... as coisinhas redondas de vidro com que os meninos gostam de brincar?

157. ESTILINGUE / SETRA / BODOQUE

... o brinquedo feito de uma forquilha e duas tiras de borracha (*mímica*), que os meninos usa m para matar passarinho?

158. PAPAGAIO DE PAPEL / PIPA

... o brinquedo feito de varetas cobertas de papel que se empina no vento por meio de uma linha?

159. PIPA / ARRAIA

E um brinquedo parecido com o (a) _____ (cf. item 158), também feito de papel, mas sem varetas, que se empina ao vento por meio de uma linha?

160. ESCONDE-ESCONDE

... a brincadeira em que uma criança fecha os olhos, enquanto as outras correm para um lugar onde não são vistas e depois essa criança que fechou os olhos vai procurar as outras?

161. CABRA CEGA

...a brincadeira em que uma criança, com os olhos vendados, tenta pegar as outras?

162. PEGA-PEGA

... urna brincadeira em que uma criança corre atrás das outras para tocar numa delas, antes que alcance um ponto combinado?

163. FERROLHO / SALVA / PICULA / PIQUE
... esse ponto combinado?

164. CHICOTE QUEIMADO / LENÇO ATRÁS
... uma brincadeira em que as crianças ficam em círculo, enquanto urna outra vai passando com uma pedrinha, uma varinha, um lenço que deixa cair atrás de uma delas e esta pega a pedrinha, a varinha, o lenço e sai correndo para alcançar aquela que deixou cair?"

165. GANGORRA
... uma tábua apoiada no meio, em cujas pontas sentam duas crianças e quando uma sobe, a outra desce? *Mímica*.

166. BALANÇO
...uma tábua, pendurada por meio de cordas, onde urna criança se senta e se move para frente e para trás? *Mímica*.

167. AMARELINHA
...a brincadeira em que as crianças riscam uma figura no chão, formada por quadrados numerados, jogam uma pedrinha (*mímica*) e vão pulando com urna perna só?
SOLICITAR DESCRIÇÃO DETALHADA

HABITAÇÃO

168. TRAMELA
... aquela pecinha de madeira, que gira ao redor de um prego, para fechar porta, janela...?

169. VENEZIANA

Quando uma janela tem duas partes, como se chama a parte de fora que é formada de tirinhas horizontais que permitem a ventilação e a claridade? *Mostrar gravura*.

170. VASO SANITÁRIO / PATENTE
Quando se vai ao banheiro, onde é que a pessoa se senta para fazer as necessidades?

171. FULIGEM
... aquilo, preto, que se forma na chaminé, na parede ou no teto da cozinha, acima do fogão a lenha?

172. BORRALHO
... a cinza quente que fica dentro do fogão a lenha?

173. ISQUEIRO / BINGA
Para acender um cigarro, se usa fósforo ou _____?

174. LANTERNA
... aquele objeto que se usa para clarear no escuro e se leva na mão assim (*mímica*)?

175. INTERRUPTOR DE LUZ
Corno se chama o objeto que fica nas paredes e serve para acender a lâmpada?

ALIMENTAÇÃO E COZINHA

176. CAFÉ DA MANHÃ
... a primeira refeição do dia, feita pela manhã?

177. GELÉIA
... a pasta feita de frutas para passar no pão, biscoito?

178. CARNE MOÍDA
... a carne depois de triturada na máquina?

179. CURAU / CANJICA

... uma papa cremosa feita com coco e milho verde ralado, polvilhada com canela?

180. CURAU

E essa mesma papa, com milho verde ralado, sem coco, como é que chama?

PEDIR PARA DESCREVER COMO SE FAZ.

181. MUNGUNZÁ / CANJICA

... aquele alimento feito com grãos de milho branco, coco e canela?

182. AGUARDENTE

... a bebida alcoólica feita de cana de açúcar?

183. EMPANTURRADO

Quando uma pessoa acha que comeu demais, ela diz: Comi tanto que estou _____.

184. GLUTÃO

... uma pessoa que normalmente come demais?

185. BALA / CONFEITO / BOMBOM

... aquilo embrulhado em papel colorido que se chupa? *Mostrar*.

PEDIR PARA DESCREVER.

186. PÃO FRANCÊS

... isto? *Mostrar*.

187. PÃO BENGALA

... isto? *Mostrar*.

VESTUÁRIO E ACESSÓRIOS

188. SUTIÃ

... a peça do vestuário que serve para segurar os seios?

189. CUECA

... roupa que o homem usa debaixo da calça?

190. CALCINHA
... a roupa que a mulher usa debaixo da saia?

191. ROUGE
... aquilo que as mulheres passam no rosto, nas bochechas, para ficarem mais rosadas?

192. GRAMPO (COM PRESSÃO) / RAMONA / MISSE
... um objeto fino de metal, para prender o cabelo? Mostrar.

193. DIADEMA / ARCO / TIARA
... o objeto de metal ou plástico que pega de um lado a outro da cabeça e serve para prender os cabelos? *Mímica*.

VIDA URBANA

194. SINALEIRO / SEMÁFORO / SINAL
Na cidade, o que costuma ter em cruzamentos movimentados, com luz vermelha, verde e amarela?

193. LOBADA / QUEBRA MOLAS
...aquele morrinho atravessado no asfalto para os carros diminuírem a velocidade?

196. CALCADA / PASSEIO
Na cidade, os automóveis andam no meio da rua e as pessoas nos dois lados, num caminho revestido de lajes ou ladrilhos. Corno se chama este caminho?

197. MEIO FIO
... o que separa o _____ (cf . item 196) da rua?

198. ROTATÓRIA / RÓTULA
... aquele trecho da rua ou da estrada que é circular, que os carros têm que contornar para evitar o cruzamento direto?

199. LOTE / TERRENO / DATA

... a área que é preciso ter ou comprar para se fazer uma casa na cidade?

200. ÔNIBUS URBANO

...a condução que leva mais ou menos quarenta passageiros e faz o percurso dentro da cidade?

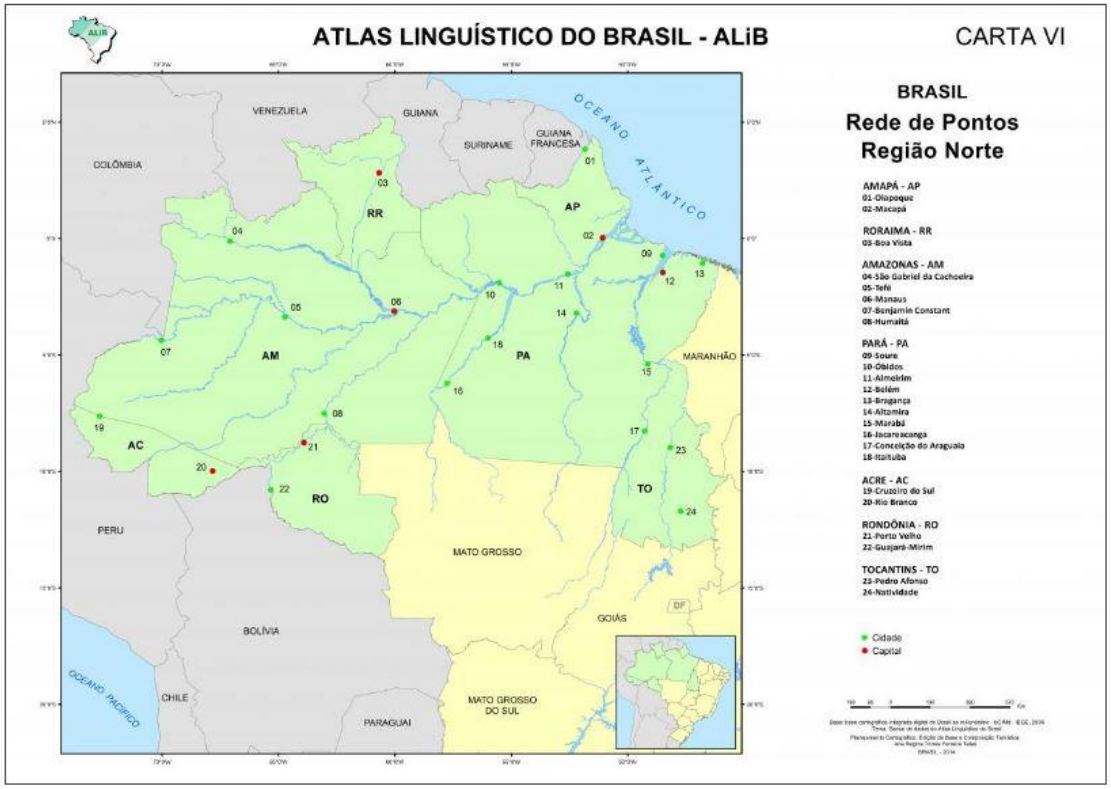
201. ÔNIBUS INTERURBANO

... a condução que leva mais ou menos quarenta passageiros de uma cidade para outra?

202. BODEGA / BAR / BOTECO

... um lugar pequeno, com um balcão, onde os homens costumam ir beber _____ (cf. item 182) e onde também se pode comprar alguma outra coisa?

Anexo 2 – Mapa da rede de pontos da região Norte



FONTE: CARDOSO, S. et alii. Atlas linguístico do Brasil. Cartas linguísticas 1, vol. 2. Londrina EDUEL, 2014.

Anexo 3 – Termo de autorização de uso dos dados do Projeto ALiB

Projeto Atlas Lingüístico do Brasil

D E C LA R A Ç Ã O

Ao utilizar dados do *corpus* do Projeto Atlas Linguístico do Brasil (Projeto ALiB), como referencial empírico do trabalho de Doutorado, intitulado "**Vocabulário dialetal: um produto lexicográfico a partir de dados geolinguísticos da região Norte do Brasil**", que desenvolvo sob a orientação de Aparecida Negri Isquerdo membro da equipe Regional Mato Grosso do Sul e do Comitê Nacional do Projeto ALiB,

DECLARO:

1. Estar ciente de que os materiais do Banco de Dados do **Projeto ALiB** a mim facultados não podem ser repassados, enquanto conjunto de dados, a outro(s) pesquisador(es) e/ou interessado(s) na matéria.
2. Ter pleno conhecimento de que a divulgação parcial ou final do trabalho deve ser sempre acompanhada da indicação da fonte (Banco de Dados do Projeto ALiB) e da citação do nome do orientador.
3. Autorizar que os resultados da análise por mim efetuada sejam utilizados nas publicações do Atlas Linguístico do Brasil, em quaisquer dos volumes que venham a integrar a coleção, mediante a indicação da fonte e a citação do meu nome.
4. Oferecer a minha contrapartida ao Atlas Lingüístico do Brasil colaborando, quando solicitado, na transcrição de dados, catalogação e cópia de materiais e em outras atividades que não impliquem a pesquisa de campo.
5. Disponibilizar os dados transcritos (em Word ou em Excel), codificados e/ou tabulados (no programa de análise utilizado) ao Comitê Nacional do ALiB.

E por estar de acordo, firmo a presente DECLARAÇÃO que tem, também, o CIENTE do Orientador e de um membro do Comitê Nacional do Projeto ALiB, que será enviada ao Arquivo Nacional, na UFBA.

Salvador, 20 de fevereiro de 2021.

Jorge Luiz Nunes dos Santos Junior
Orientando

Aparecida Negri Isquerdo
Orientador

P/ Comitê Nacional do Projeto ALiB

REGISTRADO no Projeto ALiB sob n° **121**