# Os caminhos do Projeto ALiB na intersecção entre a Linguística e a Computação: Vitalidade, Geolocalização e Mudança Semântico-Lexical

Daniela Barreiro Claro, Laila Santos, Rerisson Cavalcante, Silvana Ribeiro

https://doi.org/10.1590/SciELOPreprints.14167

Submetido em: 2025-11-17
Postado em: 2025-12-01 (versão 1)
(AAAA-MM-DD)

# Os caminhos do Projeto ALiB na intersecção entre a Linguística e a Computação: Vitalidade, Geolocalização e Mudança Semântico-Lexical

## The roadmap of the Project ALiB under the intersection of Linguistics and Computing: Vitality, Geolocation, and Lexical Semantic Change

**Tipo de Contribuição** (Relato de Pesquisa).

**Daniela Barreiro Claro**
ORCID:https://orcid.org/0000-0001-8586-1042
Professora Titular em Ciência da Computação na Universidade Federal da Bahia, Doutora pela Université d'Angers - França (2006), FORMAS (Centro de Pesquisa em Dados e Linguagem Natural) - Instituto de Computação, dclaro@ufba.br

**Laila Pereira Mota Santos**
ORCID: https://orcid.org/0009-0003-1849-0300
Doutoranda em Ciência da Computação na Universidade Federal da Bahia, Mestre em Ciência da Computação pela Universidade Federal da Bahia (2024), FORMAS (Centro de Pesquisa em Dados e Linguagem Natural) - Instituto de Computação, laila.pereira@ufba.br

**Rerisson Cavalcante de Araujo**
ORCID: https://orcid.org/0000-0002-7255-5422
Professor Associado de Linguística da Universidade Federal da Bahia, Doutor em Letras pela Universidade de São Paulo (USP), FORMAS (Centro de Pesquisa em Dados e Linguagem Natural), rerissoncavalcante@gmail.com

**Silvana Soares Costa Ribeiro**
ORCID: https://orcid.org/0000-0002-9397-6314
Professora Titular de Língua portuguesa da Universidade Federal da Bahia, Pesquisadora PQ2 do CNPQ, Doutora em Letras pela Universidade Federal da Bahia (2012), silvanar@ufba.br

**RESUMO**: O Projeto Atlas Linguístico do Brasil (Projeto ALiB) inaugurou novas direções para a Dialetologia brasileira ao adotar uma metodologia pluridimensional e ao estabelecer uma colaboração inédita entre linguistas e cientistas da Computação. Essa parceria resultou na concepção do ALiBWeb, um sistema de terceira geração voltado para a disponibilização integrada de áudios, cartas linguísticas e informações georreferenciadas. A criação dessa infraestrutura tecnológica fortaleceu a interface entre Linguística e Computação, impulsionando a formação de novos pesquisadores e a produção de novos projetos. Apesar desses avanços, muitos dos desenvolvimentos computacionais associados ao Projeto ALiB permanecem pouco conhecidos pela comunidade linguística, evidenciando a necessidade de maior divulgação das contribuições provenientes dessa intersecção interdisciplinar. Assim, o presente artigo visa apresentar um conjunto de trabalhos desenvolvidos sob a ótica da Computação, sendo analisado sob a perspectiva do Projeto ALiB, descrevendo os caminhos percorridos e os casos de interface entre a Linguística e a Computação. Um total de três trabalhos que versam sobre vitalidade, geolocalização e mudança semântico lexical foi analisado e os seus resultados foram apresentados.

**PALAVRAS-CHAVE**: ALiB, Vitalidade, Geolocalização, Mudança Semântica Lexical

**ABSTRACT**: The Linguistic Atlas of Brazil Project (ALiB) opened new directions in Brazilian dialectology by adopting a multidimensional methodology and establishing an unprecedented collaboration between linguists and computer scientists. This partnership led to the creation of ALiBWeb, a third-generation system designed to provide integrated access to audio recordings, linguistic maps, and georeferenced information. The development of this technological infrastructure strengthened the interface between Linguistics and Computing, fostering the training of new researchers and the production of new projects. Despite these advances, many of the computational developments associated with ALiB remain little known within the linguistic community, highlighting the need for broader dissemination of the contributions emerging from this interdisciplinary intersection. Thus, this study aims to present a set of works developed from a Computing perspective and analyzed through the lens of ALiB, describing the paths taken and the points of interface between Linguistics and Computing. A total of three studies, focused on vitality, geolocation, and lexical semantic change, were examined, and their results are presented.

**KEYWORDS**: ALiB, Vitality, Geolocation, Lexical Semantic Change

# 1 Introdução

O Projeto Atlas Linguístico do Brasil (Projeto ALiB) foi inovador na Dialetologia brasileira em muitos aspectos, como, por exemplo, no estabelecimento de uma metodologia pluridimensional na construção de atlas linguísticos. Outro aspecto inovador do ALiB foi na parceria entre dialetólogos e cientistas da Computação, para a criação de um atlas linguístico de terceira geração, que permitisse *"o acesso direto à voz do próprio informante em perfeita sincronização com a indicação do ponto da rede ou de exibição, via*

*Internet, de cartas e localização de pontos de inquérito e respectivas ocorrências registradas"*(CARDOSO, S. A. M. S. et al., 2014). Esse objetivo se manifestou na construção do ALiBWeb, que abriu as portas para diversas outras colaborações entre Dialetologia e Computação.

O processo de informatização no que tange ao Projeto ALiB é caracterizado pelo desenvolvimento de um Banco de Dados e de um Sistema Web, este último denominado ALiBWeb (CLARO; PAIM; JESUS, 2021). O desenvolvimento do Banco de Dados do ALiB envolveu os requisitos de negócios e a geração do MER - Modelo de Entidades e Relacionamentos (CHEN, 1976), sendo mapeado para um Sistema Gerenciador de Banco de Dados Relacional (SGBD-R), o PostgreSQL. Uma vez que o Banco de Dados do ALiB criado, desenvolveu-se o sistema ALiBWeb (CLARO; OLIVEIRA; PAIM, 2022).

O ALiBWeb é um sistema inovador em âmbito internacional, visto que se utiliza de tecnologias da área de Ciência da Computação com o intuito de melhor desenvolver e publicar os dados do Projeto ALiB. Esse sistema Web, especificado em módulos, permite gerenciar as transcrições dos inquéritos, assim como os informantes e as suas respostas. Além dessas funcionalidades, possui módulos de auditoria, autorização e autenticação de usuários. Esse sistema foi desenvolvido em *Ruby on Rails* (SMART, 2016). O ALiBWeb foi concebido para geração de cartas linguísticas automatizadas, desenvolvidas por linguistas com o intuito de validar os fenômenos linguísticos que ocorrem em cada localidade (CLARO; PAIM; JESUS, 2021; CLARO; OLIVEIRA; PAIM, 2022). A atual fase do ALiBWeb concentra a inserção dos dados através de processos de segmentação textual e segmentação de áudio, com o intuito de popular os dados com os inquéritos transcritos.

A partir do desenvolvimento do ALiBWeb, a interface entre a Linguística e a Computação se estreitou, permitindo que novos caminhos fossem trilhados e, assim, alavancou-se o desenvolvimento de novos projetos mentorados e desenvolvidos sob a ótica da Computação.

Os projetos desenvolvidos sob a perspectiva da Computação envolveram pesquisadores em diferentes níveis, desde estudantes de Graduação, assim como Mestrandos e Doutorandos da Computação, através da utilização dos dados do Projeto ALiB. Embora esses trabalhos com os dados do Projeto ALiB tenham sido publicados em eventos da área de Ciência da Computação, poucos linguistas têm acesso e conhecimento destes caminhos trilhados através desta intersecção entre o Projeto ALiB e o Centro de Pesquisa

FORMAS[1].

O presente trabalho se propõe a apresentar um conjunto destes trabalhos que foram desenvolvidos a partir do desenvolvimento do ALiBWeb sob o direcionamento da Computação, com o intuito de delinear os caminhos percorridos na intersecção da Linguística e da Computação.

Por meio deste trabalho, pretende-se responder às seguintes questões de pesquisa:

1. *Quais são os caminhos percorridos, frutos do ALiB, sob a ótica da Computação?*
2. *Como se observa a interface entre a Linguística e a Computação?*

Especificamente, objetiva-se descrever um conjunto de três trabalhos publicados após a imersão no desenvolvimento do ALiBWeb, que contribuiu para um conhecimento mais aprofundado em Linguística, especificamente em Dialetologia, por parte dos pesquisadores envolvidos com a Computação. Os três trabalhos selecionados abordam os temas da vitalidade dos termos do ALiB em redes sociais, a geolocalização dos termos e as possíveis mudanças semântico-lexicais observadas ao longo dos anos. Ao final deste trabalho, pretende-se descrever as principais contribuições dessas obras baseadas em critérios estabelecidos.

Este artigo está organizado em seções, como segue. A seção 2 apresenta a Metodologia proposta; a seção 3 descreve os resultados, sendo subdividida em três trabalhos principais deste conjunto, descrevendo sobre a vitalidade dos termos em redes sociais, a geolocalização dos termos e as mudanças semânticas lexicais. E, por fim, discute alguns pontos ainda em aberto e apresenta as considerações finais na Seção 4.

# 2 Metodologia

Esta seção descreve os procedimentos metodológicos adotados para a seleção deste conjunto dos três estudos apresentados neste trabalho. A metodologia foi estruturada de modo a permitir uma visão integrada das iniciativas desenvolvidas sob a ótica da Computação, a partir dos dados do Projeto ALiB, ressaltando as interfaces entre a Linguística e a Ciência da Computação.

[1]Centro de Pesquisa em Dados e Linguagem Natural: http://formas.ufba.br

Inicialmente, foi realizado um levantamento dos trabalhos desenvolvidos no contexto do Projeto ALiB e de suas derivações em projetos vinculados à área de Computação. Esse levantamento contemplou artigos publicados em anais de eventos da área e trabalhos de pós-graduação associados ao grupo de pesquisa FORMAS (Centro de Pesquisa em Dados e Linguagem Natural). A seleção dos trabalhos considerou três critérios principais: (i) o uso direto dos dados do Projeto ALiB; (ii) a proposição de métodos computacionais aplicados à análise linguística; e (iii) a representatividade temática na interface entre Linguística e Computação.

Após a identificação e seleção dos trabalhos, procedeu-se à sistematização das abordagens metodológicas de cada um, a fim de estabelecer um panorama comparativo. Cada estudo foi descrito segundo quatro dimensões principais: (a) o objetivo específico, (b) os dados e ferramentas utilizadas, (c) os métodos de análise empregados e (d) os principais resultados alcançados.

A análise foi conduzida de forma descritiva e interpretativa, buscando evidenciar os diferentes modos de integração entre os métodos da Linguística e da Computação. Os três estudos selecionados, sobre vitalidade dos termos, geolocalização e variação semântico-lexical, são apresentados na Seção 3, cada um com a respectiva caracterização metodológica, resultados e discussão.

Por fim, esta metodologia permitiu identificar padrões, convergências e desafios no uso de técnicas computacionais aplicadas à análise dos dados linguísticos do Projeto ALiB, contribuindo para delinear os caminhos percorridos nessa interseção disciplinar.

# 3 Resultados

Essa seção descreve os resultados obtidos por meio da leitura e síntese dos três trabalhos selecionados baseados em três critérios principais: (i) o uso direto dos dados do Projeto ALiB; (ii) a proposição de métodos computacionais aplicados à análise linguística; e (iii) a representatividade temática na interface entre Linguística e Computação.

O Projeto ALiB[2] publicou no ano de 2014 os dois primeiros volumes do atlas, que apresentam a análise da variação geossociolinguística em 25 capitais de estados (CARDOSO, S. A. M. S. et al., 2014; CARDOSO, S. A. M. et al., 2014), a partir de dados

[2]https://alib.ufba.br/histórico

coletados pela aplicação de questões fonético-fonológicas, prosódicas, semântico-lexicais, morfossintáticas, pragmáticas e metalinguísticas. Considerando oito informantes por capital, os dados analisados são de 200 dos 1100 entrevistados pelos ALiB (CARDOSO, S. A. M. S. et al., 2014; CARDOSO, S. A. M. et al., 2014).

A coleta dos dados do ALiB iniciou em 2001 e terminou em 2013, sendo necessário avaliar a vitalidade dos termos empregados.

Uma das possibilidades de analisar a vitalidade de termos é através do emprego do termo no cotidiano. Atualmente, o cotidiano no Brasil é retratado por milhares de pessoas por meio das redes sociais. Além disso, o Brasil é um dos países que mais faz uso das redes sociais em seu dia a dia, o que amplifica as potencialidades de uso destas redes sociais para o estudo da vitalidade (CARVALHO, 2022). Dentre as redes sociais mais utilizadas, o Twitter, agora denominado de X, era uma das principais redes sociais que faziam uso de mensagens curtas de forma rápida, objetiva e dinâmica.

Esse tema de pesquisa conjuga preocupações linguísticas e computacionais. Do ponto de vista linguístico, permite a análise da variação quanto à modalidade (oral ou escrita) de uso da língua e ao tipo de registro de uso. O exame de grandes quantidades de postagens em redes sociais permite a testagem de duas hipóteses linguísticas: (i) ou o gênero textual postagem curta favorece o uso de termos mais coloquiais e, por extensão, mais regionais; ou (ii) favorece a neutralização de marcas regionais, considerando que o público leitor potencial é difuso e formado por pessoas usuárias de diferentes dialetos. Do ponto de vista computacional, o tema levanta questões sobre como desenvolver técnicas para mineração e tratamento de um grande conjunto de dados.

Assim, o primeiro trabalho selecionado que faz uso direto dos dados do Projeto ALiB (i), que utiliza-se de métodos computacionais (ii) e representa a interface entre a Linguística e a Computação (iii), foi o Trabalho de Conclusão de Curso de Graduação em Ciência da Computação do discente *Arley Prates Mendes Nunes*, intitulado *Análise da vitalidade dos itens lexicais do Atlas Linguístico do Brasil no Twitter* (NUNES, 2019). Este trabalho foi publicado como resumo na 1ª Conferência Internacional de Linguística no Twitter - *Linguistweets* (**resumolinguis\textit {tweets}**) e no PROPOR 2020 - International Conference on the Computational Processing of Portuguese (NUNES et al., 2020) sob o título *Vitality analysis of the Linguistic Atlas of Brazil on Twitter*.

Embora exista outro trabalho no âmbito do FORMAS, também desenvolvido sobre

análise da vitalidade, em redes sociais, de termos de atlas linguísticos, tais como os do Atlas Linguístico Galego (ALGa), tal trabalho (CLARO; RIBEIRO; JESUS, 2021) não utiliza os dados do Projeto ALiB (i), sendo portanto descartado na seleção prévia feita para a apresentação neste artigo.

Em virtude das dificuldades de determinação das localidades de origem das postagens de redes sociais a serem analisados, observou-se que era necessário ampliar os estudos das geolocalizações com o intuito de dar mais confiança aos dados catalogados através das redes sociais. Assim, o segundo trabalho foi fruto da Iniciação Científica do discente *Pedro Guimarães Mendes Santos*, cujo título foi *Extração de Informação de Geolocalização em Redes Sociais para o Projeto Projeto ALiB*. Este trabalho também atendeu aos critérios definidos para o filtro prévio, no qual faz uso direto dos dados do ALiB (i), utiliza-se de métodos computacionais (ii) e representa a interface entre a Linguística e a Computação e foi publicado nos anais do SBSI 2023 - XIX Simpósio Brasileiro de Sistemas de Informação (SANTOS et al., 2023).

Diante dessas análises referentes à vitalidade dos termos, observou-se a necessidade de analisar e caracterizar as variações semântico-lexicais dos termos do ALiB. Assim, o terceiro trabalho culminou com uma dissertação de Mestrado intitulada *Análise da mudança semântica lexical: identificação e caracterização na língua portuguesa* da discente *Laila Pereira Mota Santos* (SANTOS, 2024). Este trabalho analisou diatopicamente a variação semântico-lexical dos termos do Projeto ALiB através de um *corpus* criado com divisões temporais, denominado *Tycholina*. Este trabalho também faz uso direto dos dados do ALiB (i), utiliza-se de métodos computacionais (ii) e representa a interface entre a Linguística e a Computação. Este último está em fase de revisão na revista Linguamática.

Os resultados obtidos em cada um dos trabalhos foram descritos em detalhes a seguir e, ao final, um panorama comparativo é fornecido com o intuito de melhor descrever as convergências e os desafios no uso das técnicas computacionais aplicadas aos dados linguísticos do Projeto ALiB, contribuindo para delinear os caminhos percorridos nessa interseção disciplinar.

## 3.1 Vitalidade dos termos em redes sociais

O objetivo de trabalho de (NUNES et al., 2020) foi desenvolver recursos computacionais que (a) buscassem grandes quantidades de *tweets* (postagens do X, antigo Twitter) com termos equivalentes aos documentados no volume II do Projeto ALiB, (b) verificassem se o sentido utilizado nas postagens correspondia aos sentidos registrados no atlas e (c) representassem a distribuição geográfica desses termos a partir da origem geográfica dos próprios *tweets*.

A abordagem proposta envolveu dois métodos: um método quantitativo e um método semântico. O **método quantitativo** consistiu em identificar automaticamente a presença de termos do Projeto ALiB no conjunto de dados obtidos e foi dividido em quatro etapas:

- FILTRO CAPITAIS: De todos os *tweets* coletados, são filtrados apenas os de usuários das capitais brasileiras;
- SEPARA EM TOKENS: Identificação de termos com palavra e termos multi-palavras;
- IDENTIFICA: Consiste na comparação da presença do termo no *tweet* de acordo o tipo de token (unica palvra ou multi-palavra);
- ARMAZENA: As seguintes informações são armazenadas do *tweet*: id, texto e localidade, assim como a carta do ALiB e o termo do ALiB correspondente.

O **método semântico** consistiu em analisar automaticamente os termos encontrados nos *tweets*, verificando se possuem o mesmo sentido empregado nas cartas semântico-lexicais do Projeto ALiB. Ele é composto de três etapas:

- AMBIGUIDADE: Identifica quais são os termos ambíguos do ALiB através do dicionário online[3] da língua portuguesa;
- DESAMBIGUAÇÃO: Desambigua o termo a partir das definições na OpenWN-PT (Open World Net - Portuguese);
- COMPARAÇÃO: Identifica se o termo desambiguado possui o mesmo sentido empregado pelo Projeto ALiB.

[3]https://www.dicio.com.br

Dentre os principais resultados obtidos por meio deste trabalho, observa-se na Figura 1 que, do total dos *tweets* que foram identificados como postados a partir das capitais, menos de 2% continham termos registrados no volume II do ALiB, perfazendo um total de 3.504 *tweets*..

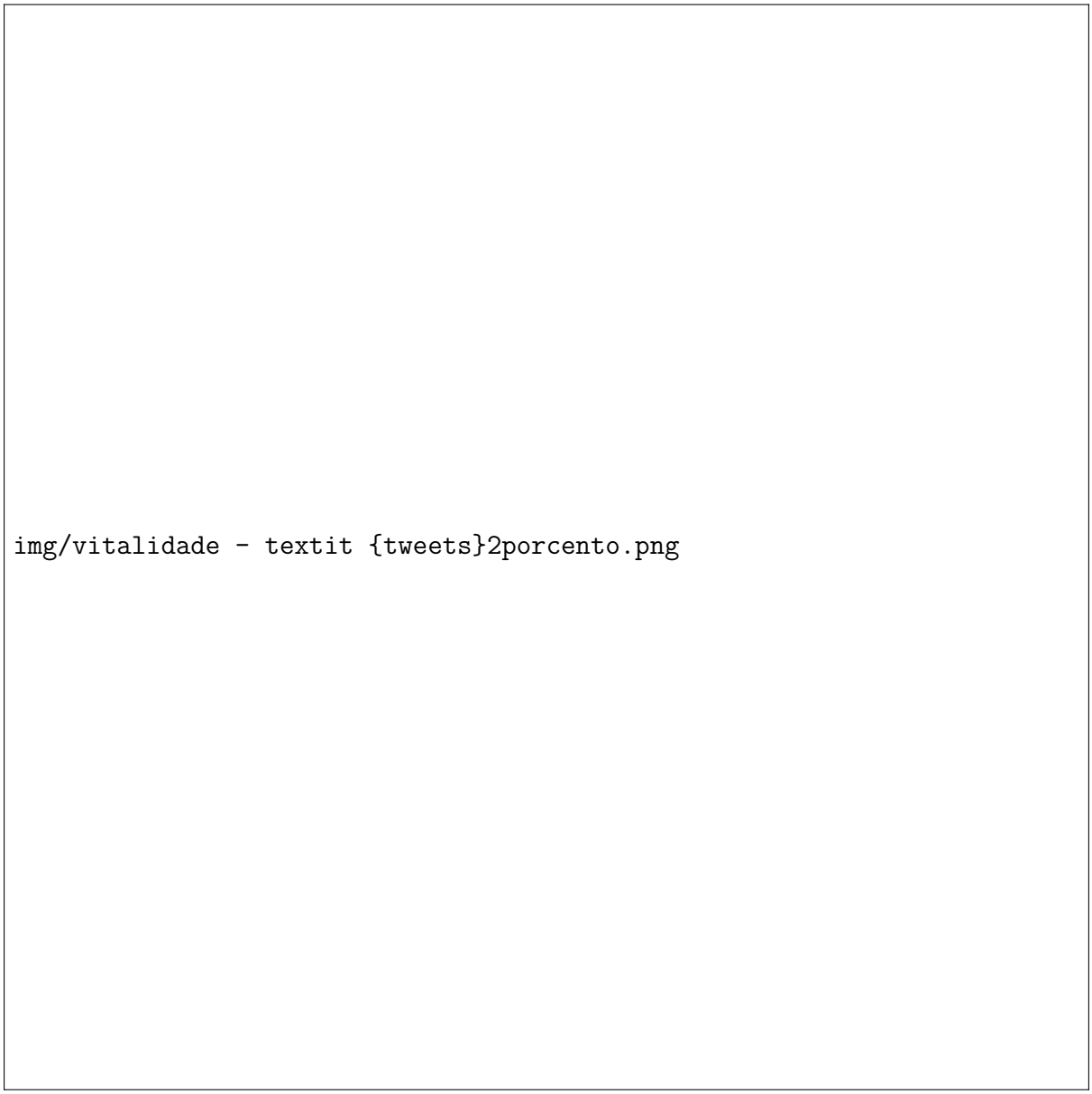


**Figura 1:** Percentual dos *tweets* das capitais que contém termos do ALiB publicado em 2014(NUNES et al., 2020).

E, por fim, dos 203 termos do ALiB, 90 estão distribuídos nestes 3.504 *tweets*, sendo os termos com mais de 50 ocorrências destacados na Figura 2.

Dentre esses, destaca-se a ocorrência, nos *tweets*, de termos registrados sob carta lexical "Prostituta". Os *tweets* da capital do Rio de Janeiro, com 563 casos, e da capital de Alagoas, com 31 dados, foram os que apresentaram, respectivamente, a maior e a menor

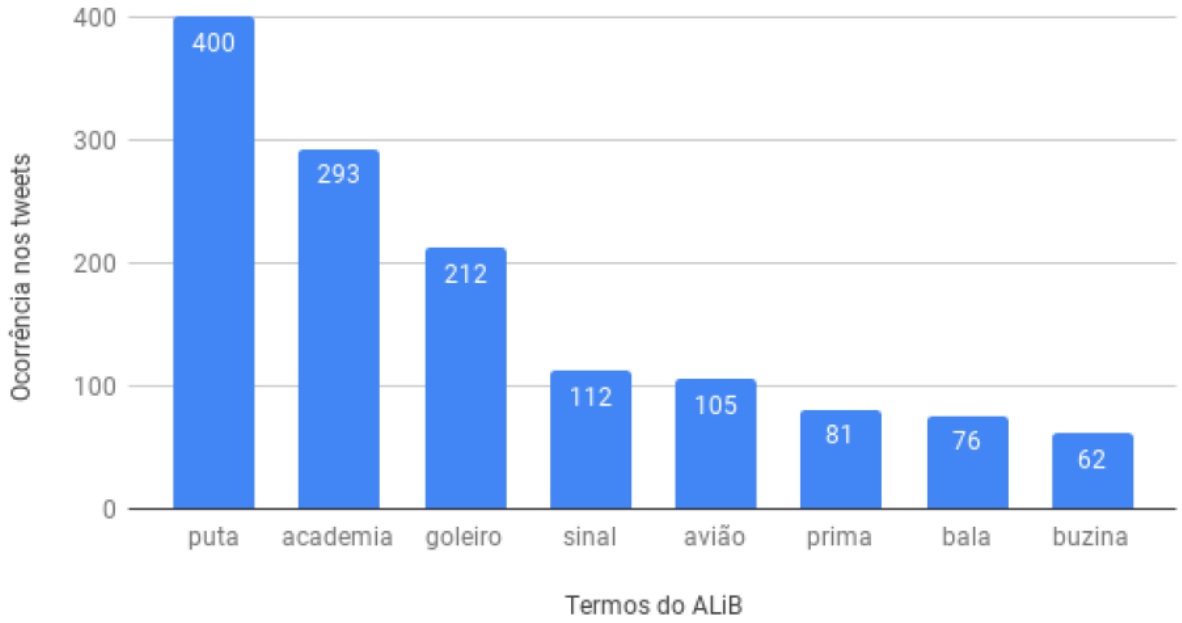


**Figura 2:** Termos do ALiB com mais de 50 ocorrências(NUNES et al., 2020)

quantidades de ocorrência de termos para o conceito “Prostituta”, conforme Figura 3.

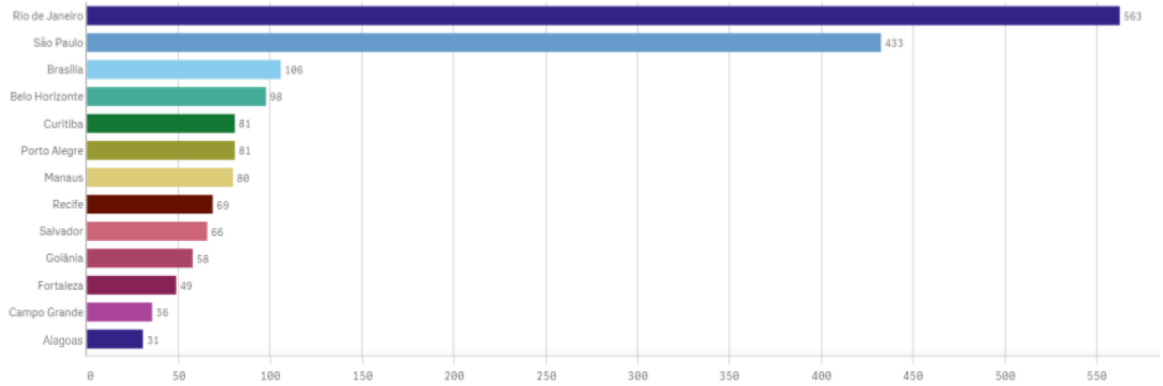


**Figura 3:** Frequência de ocorrência dos termos da carta L15 - ”Prostituta”do ALiB (NUNES et al., 2020).

Novas definições (*synsets*) foram inseridas na OpenWN-PT (PAIVA; RADEMAKER; MELO, 2012) a partir das perguntas utilizadas durante as entrevistas e que geraram as cartas semântico-lexicais do Projeto ALiB.

Considerando a carta L14 - Pernilongo do ALiB publicada em 2014 (CARDOSO, S. A. M. S. et al., 2014; CARDOSO, S. A. M. et al., 2014), após a aplicação dos métodos quantitativo e semântico, observou-se a presença de algum dos termos do ALiB em um total de 293 *tweets*. Por meio do método quantitativo, 42 termos foram identificados. Através do método semântico, esse número foi reduzido para 16, visto que alguns termos

como *praga* e *muriçoca* não foram usados representando os conceitos/referentes do Projeto ALiB. No primeiro caso, o referente é uma cidade, não um tipo de inseto; no segundo caso, trata-se do título de uma música.

Para exemplificar, seguem dois exemplos de trechos dos *tweets* dos termos *muriçoca* e *praga* encontrados nos *tweets* que não se referem ao significado do ALiB.

> Desambiguação do termo **praga**:
> Tweet: *De **Praga** até Viena de bike - 03 a 09 de junho, hospedados em lindos e charmosos Chateaux! Bem-vindo à localização m…*

> Desambiguação do termo **muriçoca**:
> Tweet: *essa música aí da **muriçoca** podia tocar o carna todo*

E, por fim, a carta Pernilongo (NUNES et al., 2020) tendo as duas ocorrências dos termos em comparação com o ALiB e no Twitter, conforme Figura 4.

## 3.2 Geolocalização

A análise da vitalidade dos termos do ALiB, ou seja, a compreensão de quais expressões permanecem em uso e sua propagação nas redes sociais, despertou interesse da comunidade, porém, um dos principais desafios inerentes ao Twitter, que dificulta o processo de análise, é a não-obrigatoriedade da marcação de localização. Outros trabalhos já analisaram a vitalidade dos termos do ALiB (NUNES et al., 2020), porém, a falta de informação referente à localidade nos *tweets* não permitiu uma comparação mais aprofundada, visto que a localidade da pessoa que *tweeta* pode não corresponder a sua verdadeira localização. Além disso, menos de 3% dos *tweets* gerados por usuários possuem a geolocalização (SERERE; RESCH, 2024).

Diante deste contexto, o trabalho aqui descrito em síntese teve por principal objetivo desenvolver uma abordagem para extração de informação de geolocalização diretamente do conteúdo textual produzido nos *tweets*, com a finalidade de maximizar as ocorrências de geocódigo (coordenadas geográficas) e, consequentemente, processar os termos do ALiB com as informações de geolocalização com maior precisão.

O desenvolvimento do método de extração da geolocalização do tweet utilizou o modelo de linguagem baseado em *Transformers*, o BERT (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2018) em sua variação treinada para o

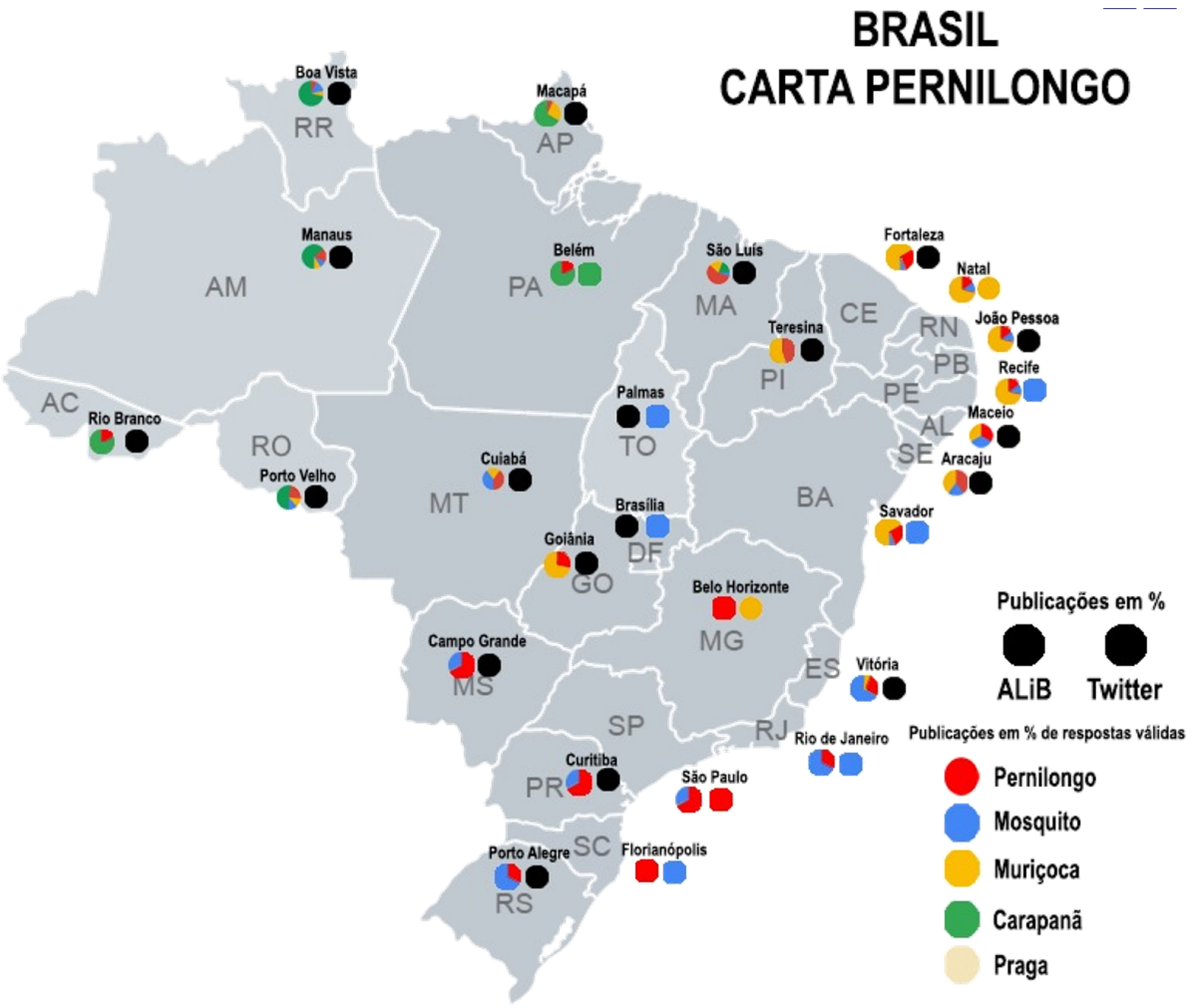


**Figura 4:** Carta Pernilongo - Comparação entre o ALiB e o Twitter (NUNES et al., 2020).

português brasileiro, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). O modelo foi treinado para a tarefa de *Named Entity Recognition* (NER, em português, Reconhecimento de Entidades Nomeadas), com o intuito de facilitar a identificação das localidades de uma maneira automatizada. O método desenvolvido foi avaliado em dois datasets totalizando mais de 720.000 *tweets*. Por fim, o mapa com as capitais foi gerado para a comparação e visualização dos termos em suas respectivas localidades.

O método para Extração de geolocalização em *tweets* com os termos do ALiB foi desenvolvido em cinco etapas: Corpora, Segregação do geocódigo, Normalização, Inferência das localizações por meio do modelo de linguagem e Plotagem no mapa, conforme a Figura 5.

Este método foi denominado GeoALiB, cujo objetivo é extrair a informação da geolocalização de um tweet através do BERTimbau para a tarefa de NER (CLARO; SANTOS


**Figura 5:** Etapas do método de extração de geolocalização(SANTOS et al., 2023).

et al., 2024). Na tarefa de NER, somente o rótulo LOCAL foi utilizado.

A maior parte dos *tweets* que continham os termos do ALiB no Corpus A não tinha nenhum dado de geolocalização (somente 4,6% dos *tweets* já possuíam geocódigo). Para o Corpus B, no qual todos os *tweets* continham termos do ALiB, cerca de 3,8% dos *tweets* possuíam geocódigo.

A etapa de inferência da localização ocorreu através do BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), processando o conjunto de *tweets* normalizados e anotando as entidades nomeadas que ocorrem no texto desses *tweets*. Ao final do processamento, os dados foram rotulados para cada tweet individualmente e avaliados por humano, caso a caso, as entidades que foram inferidas, considerando apenas os rótulos de *LOCAL* identificados. Através destes experimentos é possível observar (i) quais termos do ALiB estão em uso no Twitter e com qual a frequência eles ocorrem na plataforma; (ii) visualizar os locais dentro do território brasileiro onde os termos estão sendo usados.

Este primeiro experimento utilizou o Corpus A, com poucos *tweets* válidos, excesso de re*tweets* e ausência dos termos de interesse deste trabalho. Adiciona-se um grande número de ocorrências duplicadas e em outros idiomas. Apesar do grande volume de dados, após análise, o número foi reduzido para 1.161 *tweets* válidos, dos quais somente quinze retornaram entidades de localização após o processamento do modelo de linguagem.

No segundo experimento foram selecionados 63 termos do ALiB de diferentes cartas, o que resultou em 49.845 novos *tweets* coletados. Após o pré-processamento destes *tweets* para segregação de georreferenciamento e normalização, o modelo de linguagem foi capaz de inferir a localização de 501 *tweets*, conforme descrito na Tabela 1.

Neste estudo (SANTOS et al., 2023), somente foram descritos os resultados referentes aos termos de duas cartas do ALiB: as cartas Semáforo e Prostituta. Na análise dos termos da carta Semáforo, após a execução do BERTimbau, o processo de inferência

| **Resultados do Corpus A** | | **CorpusB-v1** | **CorpusB-v2** |
|---|---|---|---|
| Total | 500 000 | 49 845 | 176 643 |
| *tweets* não-válidos | 498 839 | | |
| *tweets* válidos | 1 161 | | |
| Sem Geocódigo | 1 107 | 48 305 | 171 226 |
| Com Geocódigo | 54 | 1 540 | 5 417 |
| BERTimbau | 3 744 | 501 | 1 603 |

**Tabela 1:** Resultados gerais com os Corpora(SANTOS et al., 2023)

do modelo de linguagem conseguiu obter 163 ocorrências do termo *semáforo*, sendo 43 delas no Rio de Janeiro (RJ), 68 em São Paulo (SP), 21 em Brasília (DF), 13 em Belo Horizonte (MG), 8 em Fortaleza (CE), 3 em Duque de Caxias (RJ), 2 em Barreiras (BA) e apenas uma ocorrência nos municípios de Contagem, Itaquera, Itatiaia, São Miguel e Mogi Guaçu. Já o termo *sinal* teve suas 16 ocorrências registradas na região sudeste, mais precisamente nas cidades de São Francisco (MG), Belo Horizonte (MG) e São Paulo (SP). Por fim, o termo *sinaleira* teve 10 registros catalogados no estado do Rio Grande do Sul, mais precisamente em Porto Alegre e Campo Bom. O mapa de calor descrito na Figura 6 permite analisar a frequência com que esses termos são aplicados no Twitter entre as cidades.

Em relação à Carta Prostituta, o modelo conseguiu inferir 151 novos termos. A localização de 11 ocorrências do termo *garota de programa* em Brasília (DF), São Paulo (SP), Rio de Janeiro (RJ), Cuiabá (MT), São Bento (PB), Curitiba (PR) e São José das Palmeiras (PR). O termo *prostituta* teve 27 ocorrências nas cidades de São Paulo (SP), Rio de Janeiro (RJ), Brasília (DF), Joinville(SC) e Curitiba (PR). Já os termos *puta* e *prima* tiveram 89 e 24 ocorrências, respectivamente, nas cidades de Belo Horizonte (MG), São Paulo (SP), Rio de Janeiro (RJ), São Bernardo e Cuiabá (MT).

O mapa de calor na Figura 7 permite analisar a frequência com que esses termos são aplicados no Twitter entre as cidades em que eles foram identificados. É possível identificar uma maior incidência dos termos desta carta da região sudeste do país, além de alguns casos identificados em algumas cidades da região nordeste.

Comparando os resultados obtidos com as duas versões do Experimento B, houve um aumento expressivo na quantidade de localidades inferidas pelo BERTimbau, de 501 para 1.603, representando uma expansão de 220% no geocódigo do CorpusB-v2 em relação ao CorpusB-v1. Esse número expressivo se deve principalmente ao crescimento da

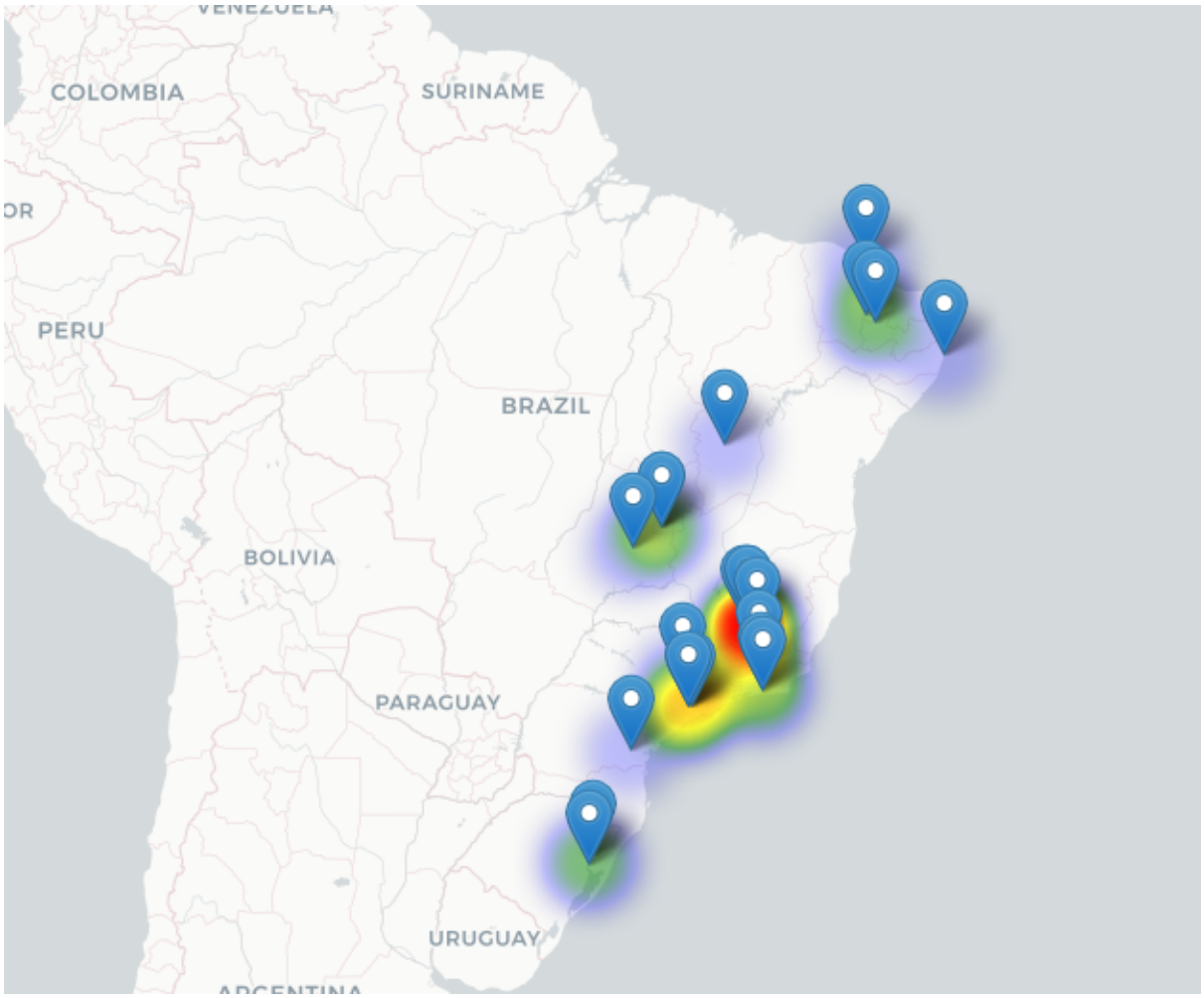


**Figura 6:** Mapa de calor da Carta Semâforo(SANTOS et al., 2023)

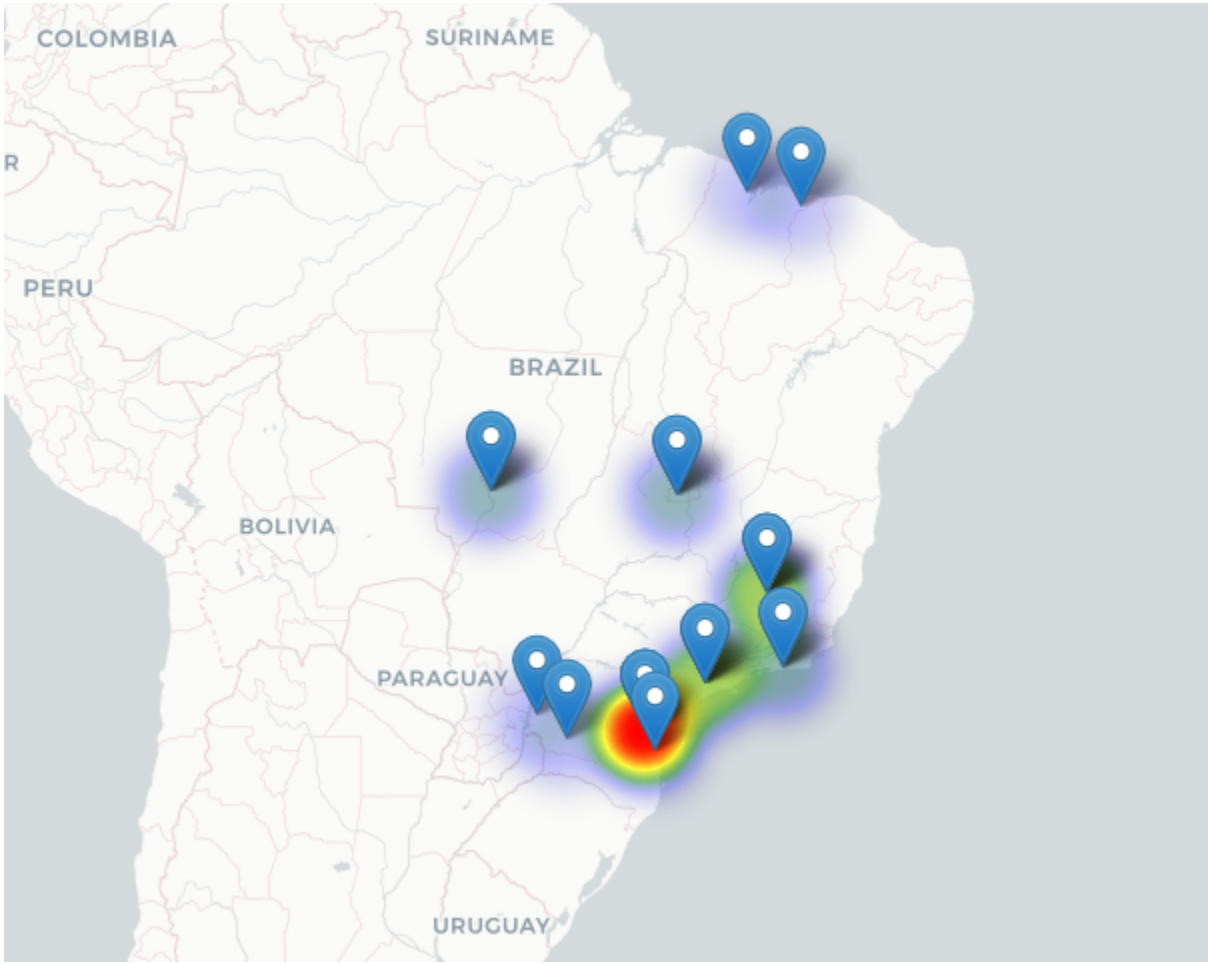


**Figura 7:** Mapa de calor da Carta Prostituta(SANTOS et al., 2023)

base. Além disso, somando-se ambos os corpora (CorpusB-v1 e CorpusB-v2), foi possível expandir o geocódigo para 2.104 *tweets* dentro do grupo de *tweets* originalmente sem geolocalização.

## 3.3 Variação Semântico-Lexical

De acordo com os resultados obtidos nestas etapas anteriores de pesquisa com a equipe do FORMAS, observou-se uma necessidade de explorar as mudanças semântico-lexicais (MSL) dos termos do ALiB ao longo do tempo. Assim, o trabalho de (SANTOS, 2024) desenvolveu um método estruturado em cinco etapas principais: (i) seleção das unidades lexicais (lexias) a serem investigadas, (ii) seleção dos corpora para treinamento e ajuste fino dos modelos, (iii) pré-processamento dos dados, (iv) seleção dos modelos de linguagem e geração de representações vetoriais (embeddings) e, por fim, (v) a análise e avaliação da MSL ao longo do tempo (Figura 8).

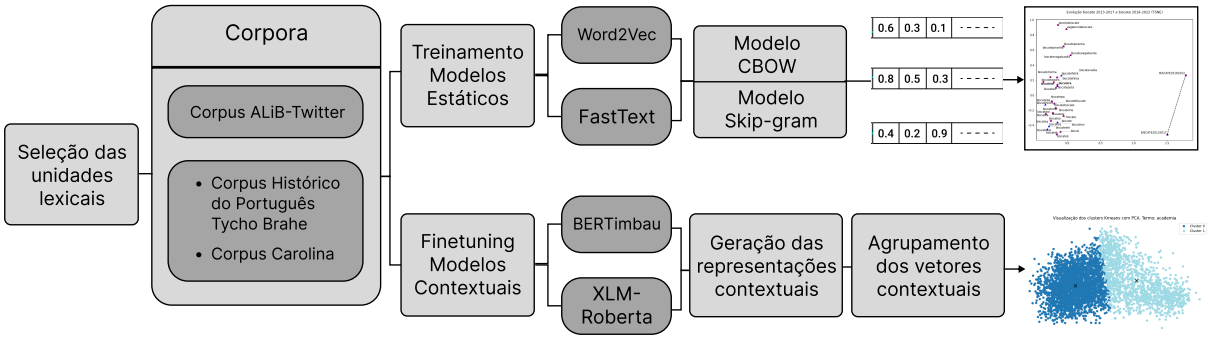


**Figura 8:** Fluxo geral dos métodos estático e contextual (SANTOS, 2024).

Para os experimentos, foram utilizados dois conjuntos de palavras (lexias $\alpha$ e $\beta$). O conjunto $\alpha$ continha itens lexicaiss polissêmicos registrados no volume II do ALiB, enquanto o conjunto $\beta$ incluía termos com provável mudança de sentido ligada a referentes recentes ou a transformações sociais.

Foram construídos dois corpora diacrônicos: o Corpus ALiB-Twitter (composto de *tweets* que continham citações as unidades lexicais investigadas no período de 2013 a 2022) e o Corpus Tycholina (união do Corpus Histórico do Português Tycho Brahe (GALVES, 2018) e do Corpus Carolina (CRESPO et al., 2023)), abrangendo um período muito mais longo (1380 a 2021).

A pesquisa comparou duas grandes abordagens de Modelos de Linguagem:

- **Modelos Estáticos (Word2Vec e FastText):** Utilizados para quantificar transformações globais de sentido, tratando cada palavra como um vetor fixo.
- **Modelos Contextuais (BERTimbau e XLM-Roberta):** Ajustados (*fine-tuned*) em corpora diacrônicos, que geram representações dinâmicas e são mais sensíveis a mudanças sutis de uso e polissemia.

Os modelos estáticos mostraram-se eficazes na quantificação de transformações globais. Contudo, sua sensibilidade é limitada por corpora ruidosos (como o ALiB-Twitter) e pela dificuldade em representar a polissemia, pois calculam uma média vetorial global para cada palavra. As Figuras 9 e 10 apresentam uma representação gráfica do espaço semântico gerado a partir da abordagem com modelo estático. A análise dos resultados buscou avaliar, por meio da examinação do espaço semântico, a vitalidade e surgimento de novos sentidos similares.

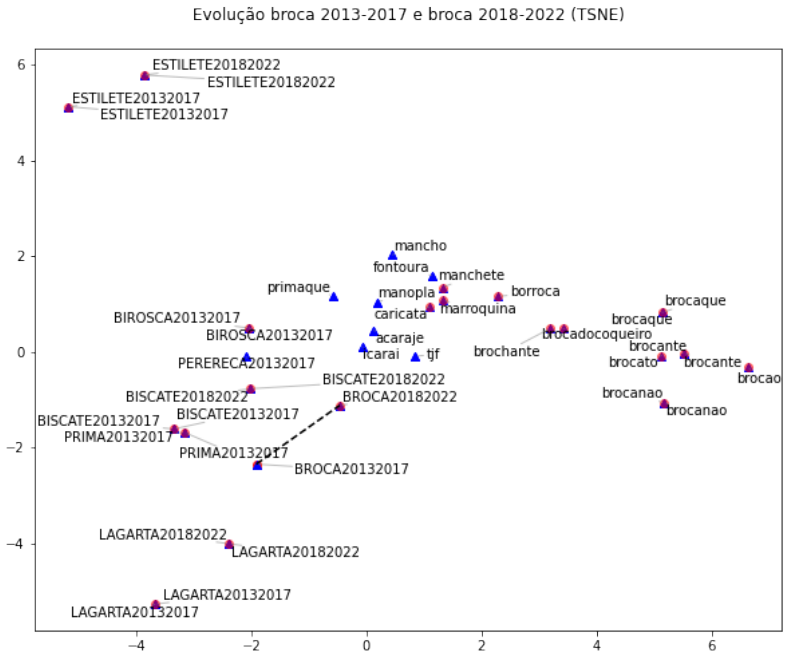


**Figura 9:** Representação do léxico "broca" (Carta Bicho da Goiaba do ALiB) com modelo FastText e Corpus Tycholina (SANTOS, 2024).

Os modelos contextuais combinados com um algoritmo de agrupamento (K-means), revelaram-se capazes de capturar mudanças de uso mais sutis. Isso ocorre porque produzem *embeddings* distintos para a mesma palavra, refletindo seu uso em sentenças específicas. Essa capacidade permitiu uma categorização mais precisa de fenômenos como

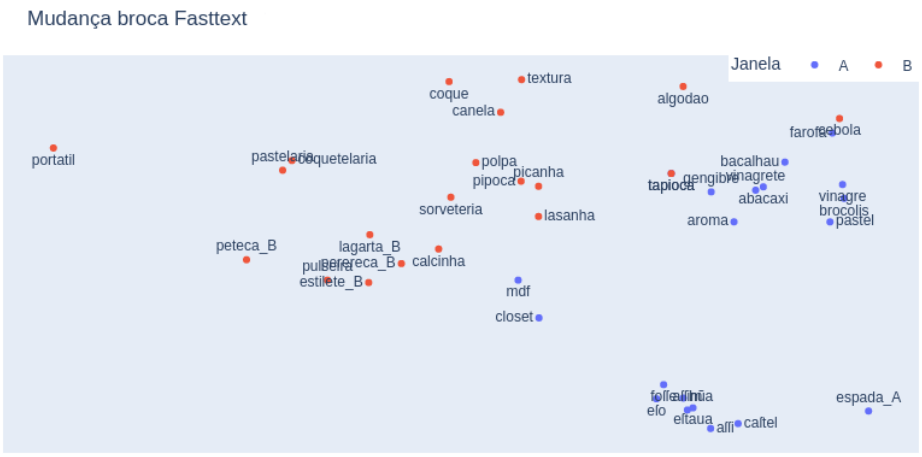


**Figura 10:** Representação do léxico "broca" (Carta Bicho da Goiaba do ALiB) com modelo FastText e Corpus ALiB-Twitter (SANTOS, 2024).

a pejoração (ex: "homem" para sentido de 'agressor' no grupo Lexias $\beta$) e a ampliação de sentido (ex: "memória" de 'lembrança' para 'componente eletrônico' no grupo Lexias $\beta$). No entanto, a interpretação dos agrupamentos gerados exige validação qualitativa humana para garantir a validade linguística. A Figura 11 apresenta uma representação gráfica do espaço semântico gerado a partir da abordagem com modelo contextual. A visualização representa os agrupamentos semânticos em duas épocas distintas e mostra o surgimento de um novo sentido a partir do surgimento de um novo grupo semântico.

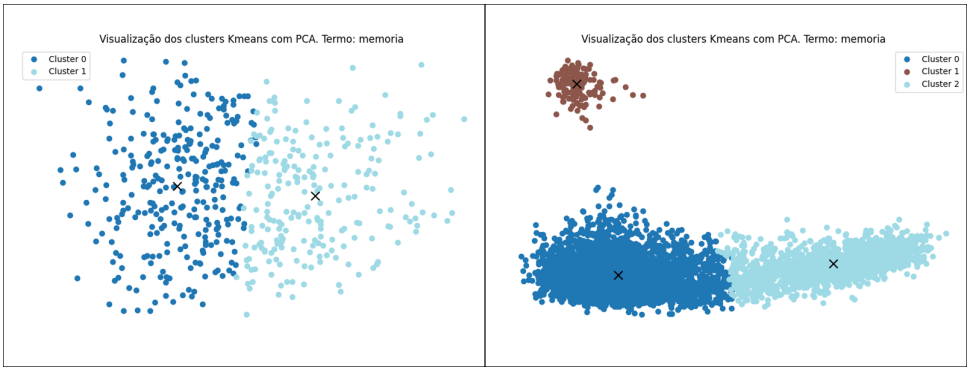


**Figura 11:** Representação do léxico "memória" com modelo BERTimbau e Corpus Tycholina (SANTOS, 2024).

O FastText demonstrou melhor desempenho do que o Word2Vec na detecção de vitalidade (se os itens lexicaiss mantêm o uso no sentido catalogado no ALiB) , especialmente em corpora ruidosos de mídias sociais (ALiB-Twitter), devido à sua capacidade de incorporar informações de subpalavras nas geração dos *embeddings*.

No panorama geral, enquanto os modelos estáticos forneceram uma visão agregada e

quantificável da MSL, os modelos contextuais permitiram uma análise mais granular dos diferentes sentidos de uma palavra ao longo do tempo.

Apesar dos resultados obtidos, a principal limitação do estudo reside na necessidade de melhoria da identificação e interpretação dos agrupamentos contextuais, que ainda dependem da avaliação qualitativa humana.

## 3.4 Panorama Comparativo

De acordo com cada estudo descrito, os três trabalhos foram analisados sob quatro dimensões: (a) o objetivo específico, (b) os dados e ferramentas utilizadas, (c) os métodos de análise empregados e (d) os principais resultados alcançados. A Tabela 2 detalha esses critérios e apresenta o panorama dos trabalhos destacados.

**Tabela 2:** Panorama dos trabalhos destacados

| **Objetivo específico** | **Dados** | **Ferramentas** | **Métodos** | **Resultados** |
|---|---|---|---|---|
| Vitalidade dos termos do ALiB | 2.692.460 *tweets* catalogados em 2019 em PT-BR | Twiter; OpenWN-PT ; Dicio | Quantitativo e Semântico | Carta Comparativa ALiB x Twitter |
| Extração da geolocalização | 720.000 *tweets* | Enelvo; BERTimbau; NER | Pipeline da extração do geocódigo | Cartas com mapa de calor da frequencia dos termos por localidade |
| Mudança Semântica Lexical | 62.269.473 tokens do Corpus ALiB-Twitter e 16.221.967 tokens do Corpus Tycholina | Word2Vec, FastText, BERTimbau e XLM-RoBERTa | Identificação de MSL por modelos estáticos e contextuais (com aplicação de algoritmo de agrupamento para identificação dos grupos semânticos) | Representações gráficas do espaço semântico das unidades lexicais e análise qualitativas da MSL |

A Tabela 2 apresenta um panorama das três frentes de investigação relacionadas ao ALiB e ao uso de dados de mídias sociais para análise linguística. O primeiro eixo aborda a vitalidade dos termos do ALiB, utilizando quase 2,7 milhões de *tweets* em português brasileiro coletados em 2019. Com o apoio de recursos lexicais como o OpenWN-PT e o Dicio online, além da API do Twitter, foram aplicados métodos quantitativos e semânticos

para comparar a presença e circulação dos termos do ALiB em redes sociais, resultando em uma carta comparativa entre ALiB e Twitter.

O segundo eixo diz respeito à extração de geolocalização, baseada em 720 mil *tweets*. Para isso, empregaram-se ferramentas de PLN como Enelvo, BERTimbau e técnicas de NER, compondo um pipeline dedicado à identificação de geocódigos. A aplicação desse fluxo permitiu mapear a distribuição espacial dos termos, gerando cartas em formato de mapa de calor que evidenciam a frequência lexical por localidade.

Por fim, o terceiro eixo concentra-se na mudança semântico-lexical, a partir de dois grandes corpora: o ALiB-Twitter, com mais de 62 milhões de tokens, abrangendo um período de dez anos, e o Corpus Tycholina, com mais de 16 milhões, abrangendo um período de seis séculos. Utilizando modelos de representação vetorial estáticos (Word2Vec, FastText) e contextuais (BERTimbau, XLM-RoBERTa), foram aplicadas técnicas de agrupamento para a identificação de grupos semânticos. Os resultados incluem representações gráficas do espaço semântico das unidades lexicais e análises qualitativas que discutem evidências de mudança e variação semântica.

# 4 Considerações Finais

A partir do percurso apresentado, observa-se que a colaboração entre pesquisadores do ALiB e do FORMAS, para a informatização do ALiB, inicialmente centrada no desenvolvimento do Banco de Dados e do sistema ALiBWeb, desencadeou uma ampliação significativa das possibilidades de pesquisa na interface entre Linguística e Computação. O ALiBWeb não apenas modernizou a gestão, a validação e a publicação dos dados do atlas, mas também consolidou uma infraestrutura capaz de sustentar novas frentes de pesquisa da Linguística computacional. Esse movimento permitiu que pesquisadores da Computação, em diferentes níveis de formação, se aproximassem da Dialetologia (e da Linguística de modo mais amplo) e utilizassem dados do ALiB como base para experimentação, inovação metodológica e desenvolvimento de ferramentasespecializadas.

O panorama dos três trabalhos descritos demonstra concretamente como essa aproximação se materializou em contribuições distintas, mas complementares. O estudo sobre a vitalidade dos termos do ALiB no ambiente digital, apoiado em mais de dois milhões de *tweets*, revela como dados de redes sociais podem ampliar a compreensão do uso lexical

contemporâneo através das redes sociais. A pesquisa dedicada à extração de geolocalização evidencia o potencial de modelos de PLN para mapear a distribuição espacial de unidades lexicais em larga escala, o que, inclusive, abre novas possibilidades para a análise da difusão geolinguística de itens lexicais. Já a investigação sobre mudança semântico-lexical, ancorada em corpora extensos e modelos estáticos e contextuais, mostra como técnicas modernas de representação vetorial podem evidenciar as continuidades e transformações nos significados ao longo do tempo.

Dessa forma, as contribuições apresentadas revelam que a intersecção entre Linguística e Computação, impulsionada pelo ALiBWeb, não se limita ao apoio tecnológico, mas configura um espaço oportuno para o desenvolvimento de novas metodologias, abordagens analíticas e interpretações linguísticas. Os resultados alcançados evidenciam a relevância dessa colaboração interdisciplinar e apontam para um campo ainda em expansão, no qual a integração entre bases dialetológicas tradicionais e técnicas computacionais avançadas promete aprofundar, de maneira inédita, o conhecimento sobre a variação e a mudança linguística no Brasil.

## Agradecimentos

Os autores gostariam de agradecer à FAPESB, à CAPES, ao CNPQ e à UFBA pelo auxilio financeiro através dos Projetos TIC 002/2015 e CCE 023/2024, além do programa de Bolsas de Iniciação Científica da UFBA e da CAPES.

## Referências

CARDOSO, Suzana A. M. et al. **Atlas linguístico do Brasil**. Londrina: Eduel, 2014. v. 2. ISBN 978-85-7216-709-3.

CARDOSO, Suzana A. M. S. et al. **Atlas linguístico do Brasil**. Londrina: Eduel, 2014. v. 1. ISBN 978-85-7216-705-5.

CARVALHO, Vívian de Nazareth Santos. A criatividade linguística nas Redes Sociais: os processos de criação de novas palavras na internet. **Palimpsesto - Revista do Programa de Pós-Graduação em Letras da UERJ**, v. 21, n. 38, p. 114–127, maio 2022. DOI: 10.12957/palimpsesto.2022.62857. Disponível em: <https://www.e-publicacoes.uerj.br/palimpsesto/article/view/62857>.

CHEN, Peter Pin-Shan. The entity-relationship model—toward a unified view of data. **ACM transactions on database systems (TODS)**, Acm New York, NY, USA, v. 1, n. 1, p. 9–36, 1976.

CLARO, Daniela Barreiro; OLIVEIRA, Josane Moreira de;
PAIM, Marcela Moura Torres. ALiBWeb. **Work. Pap. em Linguística**, Universidade Federal de Santa Catarina (UFSC), v. 23, n. 1, p. 75–90, set. 2022.

CLARO, Daniela Barreiro; PAIM, Marcela Moura Torres;
JESUS, Luis Emanuel Neves de. O sistema do projeto atlas linguístico do Brasil: análise linguística automatizada. **Rev. Diadorim**, Programa de Pos-Graduacao em Letras Vernaculas - PPGLEV, v. 23, n. 1, p. 273–287, jan. 2021.

CLARO, Daniela Barreiro; RIBEIRO, Silvana Soares Costa;
JESUS, Luis Emanuel Neves de. ANÁLISE DOS TERMOS “DOR” E “GUAPO” PRESENTES NO ATLAS LINGUÍSTICO GALEGO E SUA VITALIDADE NO TWITTER: UMA PROPOSTA METODOLÓGICA. **Estudos Linguísticos e Literários**, n. 71, p. 108–136, dez. 2021. DOI: 10.9771/ell.i71.48189. Disponível em: <https://periodicos.ufba.br/index.php/estudos/article/view/48189>.

CLARO, Daniela Barreiro; SANTOS, Joaquim et al. Extração de Informação. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. [S.l.]: BPLN, 2024. cap. 22. ISBN 978-65-01-20581-6. Disponível em: <https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-ie/cap-ie.html>.

CRESPO, Maria Clara Ramos Morales et al. **Carolina: a General Corpus of Contemporary Brazilian Portuguese with Provenance, Typology and Versioning Information**. [S.l.: s.n.], 2023. arXiv: 2303.16098 [cs.CL].

DEVLIN, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **CoRR**, abs/1810.04805, 2018. arXiv: 1810.04805. Disponível em: <http://arxiv.org/abs/1810.04805>.

GALVES, Charlotte. The Tycho Brahe Corpus of Historical Portuguese: Methodology and results. **Linguistic Variation**, v. 18, p. 49–73, jul. 2018. DOI: 10.1075/lv.00004.gal.

NUNES, Arley Prates M et al. Vitality analysis of the linguistic atlas of brazil on twitter. In: SPRINGER. INTERNATIONAL Conference on Computational Processing of the Portuguese Language (PROPOR 2020). [S.l.: s.n.], 2020. P. 184–194.

NUNES, Arley Prates Mendes. **Análise da vitalidade dos itens lexicais do Atlas Linguıstico do Brasil no Twitter**. 2019. F. 53. Monografia (TCC) – Universidade Federal da Bahia, Salvador. Disponível em: <https://alib.ufba.br/producoes-lexicologia>.

PAIVA, Valeria de; RADEMAKER, Alexandre; MELO, Gerard de. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In: PROCEEDINGS of COLING 2012: Demonstration Papers. Mumbai, India: The COLING 2012 Organizing Committee, dez. 2012. P. 353–360. Published also as Techreport http://hdl.handle.net/10438/10274. Disponível em: <http://www.aclweb.org/anthology/C12-3044>.

SANTOS, Laila P. M. **Análise da mudança semântica lexical: identificação e caracterização na língua portuguesa**. ago 2024. Dissertação – Universidade Federal da Bahia, Salvador.

SANTOS, Pedro et al. Ampliando a vitalidade dos termos do ALiB através da Extração de Informação Geolocalizada nas mídias sociais. In: ANAIS Estendidos do XIX Simpósio Brasileiro de Sistemas de Informação. Maceió/AL: SBC, 2023. P. 178–183. DOI: 10.5753/sbsi_estendido.2023.229362. Disponível em: <https://sol.sbc.org.br/index.php/sbsi_estendido/article/view/24615>.

SERERE, Helen Ngonidzashe; RESCH, Bernd. Understanding the impact of geotagging on location inference models for accurate generalization to non-geotagged datasets. **Geomatica**, v. 76, n. 1, p. 100004, 2024. ISSN 1195-1036. DOI: https://doi.org/10.1016/j.geomat.2024.100004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1195103624000041>.

SMART, Mark. **Ruby on rails 5: web app development for beginners**. [S.l.]: CreateSpace Independent Publishing Platform, 2016.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: CERRI, Ricardo; PRATI, Ronaldo C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020. P. 403–417. ISBN 978-3-030-61377-8.

**Contribuições.** Os autores contribuíram igualitariamente para o trabalho.

**Conflito de interesse.** Os autores declaram que não possuem interesses financeiros ou relações pessoais que possam ter influenciado o trabalho relatado neste artigo.

**Declaração de disponibilidade dos dados da pesquisa.** Os dados utilizados neste trabalho fazem parte do Projeto ALiB, projeto nacional, com cadastro no CNPq, cujos dados não podem ser divulgados abertamente, tendo em vista que estão passando por um processo de limpeza das informações que possam identificar os envolvidos.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.