Universidade Federal da Bahia - UFBA
Instituto de Matemática e Estatística - IME
Departamento de Ciência da Computação - DCC
Bacharelado em Sistemas de Informação - BSI

Análise da vitalidade dos itens lexicais do Atlas Linguístico do Brasil no Twitter

Arley Prates Mendes Nunes

Salvador - Bahia
Julho de 2019

Análise da vitalidade dos itens lexicais do Atlas Linguístico do Brasil no Twitter

Arley Prates Mendes Nunes

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal da Bahia como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientadora: Profª. Daniela Barreiro Claro.

Salvador - Bahia

Julho de 2019

À minha família

Agradecimentos

Primeiramente a Deus. A minha mãe Alvisa e ao meu pai Antenor, a razão da minha vida. Ao meu irmão Aécio, pelo apoio e irmandade.
Aos meus avós, em especial, Vó Tereza e Vô Bebe. Aos meus tios, em especial, minha tia Alda. Aos meus primos, em especial minha afilhada Maria Tereza e minha prima irmã Nana.
A Elienai Esquivel, meu amor linda.
Ao meu amigo Luís, por acreditar sempre em mim e a sua família que me acolheram como um filho. A Rodrigo pela amizade e disponibilidade em me ajudar no meu desenvolvimento.
Aos meus amigos do GERE - UFBA-STI, em especial a Carla Bahia. E aos inúmeros amigos que estiveram comigo nessa jornada.
Aos Professores, em especial Profª. e orientadora Daniela Claro, que desde o início da graduação sempre me motivou, inspirou e me orientou nas minhas escolhas. Ao Projeto ALiB pela parceria e por fim, ao grupo FORMAS pelo feedback, em especial a Babacar.

Resumo

O Projeto Atlas Linguístico do Brasil (ALiB) tem o objetivo de descrever a realidade linguística dos brasileiros. Os dois primeiros volumes do ALiB foram publicados em 2014 e correspondem a análise da variação sociolinguística em 25 capitais de estado. No entanto, a obtenção destes dados iniciou há mais de duas décadas, sendo assim necessário avaliar a vitalidade dos termos empregados. A vitalidade destes termos pode ser analisada no cotidiano por meio das redes sociais. Atualmente, 91,3% dos internautas utilizam diariamente uma rede social. Assim o presente trabalho tem por objetivo analisar a vitalidade dos termos das cartas semântico-lexical do ALiB em um corpus do Twitter. A metodologia consistiu em primeiramente identificar, nos tweets, os termos catalogados pelo Projeto ALiB, e em seguida analisar semanticamente cada termo através de um algoritmo de desambiguação de sentido por meio da OpenWordnet-Pt. Os experimentos validaram que 52,5% dos termos utilizados no Projeto ALiB são ainda utilizados nas redes sociais. Alguns termos tais como badogue, biloca, curica estão em desuso visto que não foram encontradas nenhuma ocorrência no Twitter.

Palavras-chave: sociolinguística, vitalidade, twitter, desambiguação lexical de sentido, openwordnet-pt.

Abstract

The Brazilian Linguistic Atlas Project (ALiB) has the goal to describe the linguistic reality of the Brazilian people.The two first volumes of ALiB was published in 2014 and represents an analysis of the sociolinguistic variation in 25 state capitals. However, the obtaining process of this data started more than two decades ago, being necessary to evaluate the vitality of the terms employed. The vitality of the terms may be analyzed in daily life through social networks. Today, 91,3% of the internet users use social networks. This document aims to analyze the vitality of the terms of the semantic-lexical letters of the ALiB in a Twitter corpus. The methodology consisted in first, identify, in tweets, the terms cataloged by the ALiB Project, and, after that, analyze semantically each term through a sense disambiguation algorithm through OpenWordnet-Pt. The experiments validated that 52,5% of the terms utilized in ALiB Project are used in social networks today. In addition it was possible to observe that some terms like "badogue", "bioloca" and "curica" are in disuse since 47.5% of the occurrences were obtained with the same sense of the Project.

Keywords: sociolinguistics, vitality, twitter, word sense disambiguation, openwordnet-pt.

Sumário

Lista de Figuras

Capítulo 1

Introdução

A maneira mais comum das pessoas se comunicarem é através da linguagem natural, seja pela fala ou escrita. De acordo com [e Silvana Araújo 2011], a sociolinguística tem por objeto de estudo os padrões de comportamento linguístico observáveis dentro de uma comunidade de fala. Dessa forma, linguistas e dialetólogos se reuniram para elaboração de um Atlas Linguístico do Brasil, cujo principal objetivo é descrever a realidade linguística do Brasil [Cabezudo 2015].

O projeto ALiB publicou no ano de 2014 os dois primeiros volumes, que corresponde a análise da variação sociolinguística em 25 capitais de estados (PROJETO ALiB)[1]. Segundo [CARDOSO 2009], os informantes somaram um total de 1100 e os dados foram catalogados através da aplicação de questionários linguísticos, como por exemplo, fonético-fonológico, prosódia e semântico-lexical.

A obtenção destes dados iniciou há mais de duas décadas, sendo necessário avaliar a vitalidade dos termos empregados. No mundo, o Brasil é um dos países que mais faz uso das redes sociais em seu dia a dia, o que possibilita o estudo da vitalidade linguística por meio das mídias sociais [Coelho 2018]. Entre elas o Twitter, um serviço de microblog e rede social muito utilizado pelos brasileiros, o qual possui uma característica particular: possibilita que seus usuários compartilhem mensagens curtas de forma rápida, objetiva e dinâmica.

Este trabalho analisou a vitalidade linguística dos termos catalogados do Projeto ALiB através de uma análise quantitativa e semântica, utilizando dados do Twitter, extraídos entre os meses de fevereiro a março de 2019. Na análise quantitativa verificou-se a ocorrência dos termos presentes nos dados desta rede social. A análise semântica consistiu no estudo de um método de desambiguação lexical de sentido, para identificação do mesmo sentido empregado no Projeto ALiB.

Este trabalho está organizado em capítulos como segue: o capítulo 2, apresenta

[1] https://alib.ufba.br/histórico

a fundamentação teórica; o capítulo 3 descreve a metodologia; o capítulo 4 apresenta os experimentos realizados e resultados obtidos; no 5 são abordadas algumas discussões em relação aos desafios encontrados; por fim, o capítulo de número 6 tem seu conteúdo voltado para conclusões e trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo é organizado da seguinte maneira: 2.1 O Projeto Atlas Linguístico do Brasil, a idealização e elaboração do projeto ALiB; 2.2 Vitalidade, apresenta o conceito de vitalidade; em 2.3 Twitter, descreve como funciona uma das redes sociais mais acessadas pelos brasileiros e que serviu como principal ferramenta para coleta de dados deste trabalho; 2.4 Algoritimo de Lesk, apresenta a idealização para construção e funcionamento do método de desambiguação de sentido; finalizando este capítulo, 2.5 Wordnet, conceitua e desenvolve sobre uma Wordnet para o português brasileiro.

2.1 O Projeto Atlas Linguístico do Brasil

Em março do ano de 1952 o governo brasileiro anunciou o decreto nº 30.643, que tem como um dos principais objetivos a elaboração de um atlas linguístico do Brasil. Apesar desta iniciativa, pesquisadores brasileiros, visando a necessidade de estudar a dialetologia no Brasil, somente deram início à criação do Projeto Atlas Linguístico do Brasil (Projeto ALiB) em 1996, que reuniu pesquisadores dos atlas regionais já publicados daquela época [Cardoso and Mota 2012].

Um atlas linguístico não é simplesmente uma coleção de mapas com indicações de caráter geográfico, geopolítico, social, econômico, entre outras, mas um atlas que, ao lado dessas indicações, traz, com evidência, informações sobre a realidade da língua, os diferentes usos, as diversas maneiras de sua realização e os processos de escolha que os membros de uma coletividade assumem [Jacyra Andrade Mota 2015].

O objetivo do Projeto ALiB é descrever a realidade linguística do Brasil, no que tange à língua portuguesa, com enfoque na identificação das diferenças diatópicas (fônicas, morfossintáticas, léxico-semântico e prosódia) consideradas na perspectiva da Geolinguística.

Em 2014, durante o III Congresso de Dialetologia e Sociolinguística, foram pu-

blicados o Volume I de introdução e o Volume II que apresenta 159 cartas linguísticas com dados de 25 capitais dos estados brasileiros, à exceção apenas de Brasília (Distrito Federal) - em vista da data de sua criação - e Palmas, capital do recém-criado do Estado de Tocantins [Ribeiro 2015].

Para construção do atlas do Brasil os pesquisadores contaram com um total de 250 localidades distribuídas pelo país e selecionadas de acordo com critérios demográficos, históricos e culturais, além de, considerar a extensão de cada Estado/região e a natureza de seu povoamento na delimitação do número de pontos da área [do Brasil 2001].

Segundo [CARDOSO 2009], os informantes somaram um total de 1100, atendendo a duas faixas etárias, entre 18 e 30 anos e 50 e 65 anos, habitantes das localidades pesquisadas e filhos de pais que pertencem a localidade, há uma variedade quanto a condições sociais, sexo e escolaridade. Em se tratando do nível escolar, os participantes são alfabetizados e nas capitais dos estados foram incluídos 4 informantes de nível universitário.

Os dados foram catalogados através da aplicação de três tipos de questionários linguísticos, organizados da seguinte maneira: (a) fonético-fonológico - 159 perguntas, às quais se juntam 11 questões de prosódia; (b) semântico-lexical - 202 perguntas; e (c) morfossintático - 49 perguntas [CARDOSO 2009].

2.2 Vitalidade

A diversidade linguística é parte do patrimônio da humanidade, cada língua consiste na real identidade de um povo. Sendo assim, considerar que uma língua pode morrer é uma negligência irrecuperável para os seres humanos [Drude et al. 2003].

Uma língua morre quando o seu último falante desaparece. Acontece a despedida de uma maneira de olhar e organizar o mundo, específica da comunidade que a fala. Com ela, extingue também o conhecimento e o saber acumulado por essa comunidade durante sua história [UNESCO 2019].

Em [Cristófaro-Silva 2002], apud, [Ngunga and Bavo 2011] afirma que o caminho que uma língua percorre está ligado ao ciclo da vida "pois a língua é um meio social que nasce, cresce, desenvolve-se e morre". Para o autor, a construção de uma língua faz parte de um processo gradual, lento e coletivo. E sua morte tem início com as divergências de transmissão para as futuras gerações.

Para avaliar o grau de vitalidade no processo de mudança sofrido pelas línguas, sociolinguistas têm elaborado escalas de vitalidade linguística. [Drude et al. 2003] propõe seis fatores no estudo de línguas ameaçadas de extinção, cada um com gradação de 0 (zero) a 5 (cinco) que descrevem o estágio de vitalidade ou o perigo de desaparecimento de uma determinada língua: fator 1: Transmissão da língua às gerações futuras; fator 2: Número

absoluto de falantes; fator 3: Porcentagem de falantes em meio ao total da população; fator 4: Mudança de domínios no uso da língua; fator 5: Resposta aos novos domínios e à mídia e por último; fator 6: Materiais para o ensino da língua e letramento.

Este trabalho utiliza como base o fator de número 5: Resposta aos novos domínios e à mídia. Entre esses meios, a internet e, especificamente, as mídias sociais, são os principais meios de comunicação devido a grande parcela de internautas brasileiros.

O estudo da vitalidade linguística, neste trabalho, é realizado pela utilização dos termo obtidos pelo Projeto ALiB, semelhante aos trabalhos [Teixeira 2015] e [Nunes 2014], é estudado a vitalidade, particularmente a de regionalismos madeirenses. O grau de vitalidade proposto pelos autores [Drude et al. 2003], são comumente utilizados para avaliar a vitalidade de línguas ameaçadas de extinção, como por exemplo, de dialetos ou línguas de regiões minoritárias, no entanto, a vitalidade avaliada neste trabalho é sobre os termos catalogados pelo projeto ALiB dos brasileiros de diferentes localidades.

2.3 Twitter

A internet é uma das mais relevantes ferramentas de comunicação dos dias atuais. No Brasil, cerca de 140 milhões de habitantes fazem uso das redes sociais, ou seja 66% da população. Por conta disto, os usuários brasileiros gastam em média 3h 34min diariamente utilizando as redes sociais [Coelho 2018]. As idades dos internautas variam entre 18 e 65 anos, entretanto, as faixas etárias se dividem em maior uso entre 25 a 34 anos; o segundo maior grupo de usuários entre 18 e 24 anos; o terceiro está a população entre 35 a 44 anos e por fim, os idosos a partir de 65 anos [Coelho 2018]; [Ribeiro 2019].

Lançado no final de 2006, o Twitter nasceu com o objetivo de oferecer aos seus usuários um serviço de microblog e rede social. As pessoas que fazem uso desta plataforma digital podem publicar tweets, ou seja, mensagens de até 280 caracteres. Com esta característica, o Twitter é conhecido como uma rede social de conteúdo rápido, objetivo e dinâmico. E nos dias de hoje está disponibilizado em 37 idiomas [Vinha 2017].

A utilização do Twitter é ainda caracterizada por picos durante eventos sociais populares, como competições esportivas, eleições, acontecimentos inesperados e fatos relacionados a celebridades. Um exemplo disso ocorreu no dia 25 de junho de 2009, na ocasião da morte do cantor Michael Jackson, momento em que os servidores da rede social caíram devido à alta quantidade de mensagens postadas em um curto período de tempo a respeito do acontecimento [ICMNews, 2009].

2.4 Algoritmo de Lesk

O algoritmo de Lesk desenvolvido em 1986 por Michael Lesk, utiliza tradicionais dicionários, como o Oxford Advanced Learner's [Banerjee and Pedersen 2002], com o objetivo de desambiguar palavras específicas presentes em frases curtas. No entanto, Benerjee e Pedersen, propuseram uma adaptação deste método, para que ele pudesse ser utilizado como Wordnet ao invés de dicionário. Wordnet é estruturado semanticamente diferente de dicionários que são alfabeticamente [Cabezudo 2015]. Este método é empregado na área de Word Sense Disambiguation ou como conhecida no Brasil, Desambiguação lexical de sentido [Nóbrega 2013].

Em [Nóbrega 2013], afirma que, muito provavelmente, o método de Lesk (1986) foi o primeiro método computacional de desambiguação de sentido. O mesmo se caracteriza pelo uso de dicionários e pela independência de língua, ou seja, apesar de ter sido originalmente desenvolvido para o Inglês pode ser portado para outros idiomas, como o Português, por meio da alteração do dicionário utilizado.

O funcionamento do algoritmo de Lesk, originalmente, consiste na escolha do sentido para uma palavra-alvo que possui um maior número de palavras comuns em sua definição do dicionário com os rótulos atribuídos às palavras no contexto dela.

Para exemplificar como funciona o algoritmo de Lesk, considera-se a seguinte sentença: "sentar no banco".

O significado de "sentar": (1) Pôr(se) num assento; (2) apoiar as nádegas num assento; (3) Colocar ou colocar-se em determinado lugar;

O significado de "banco": (1) Assento estreito e comprido; (2) Pranchão elevado que trabalham os carpinteiros; (3) Balcão de comércio; (4) Grande cardume de peixe; (5) Instituição financeira;

Nesse contexto o primeiro significado de sentar possui uma palavra em comum com o primeiro significado de banco. Os outros significados não apresentam nenhuma palavra em comum. Dessa forma, o algoritmo de Lesk define o sentido com maior número de palavras em comum entre os significados.

[Nóbrega 2013] exemplifica o 4 tipos de ambiguidade: lexical, estrutural, anafórica (ou referencial) e temática. Como ele, esta monografia se atém ao tipo lexical, que ocorre devido ao fenômeno da homonímia e da polissemia. Uma simples maneira de entender estes dois fenômenos é saber que há palavras que são ditas e escritas da mesma maneira, porém, seus significados divergem. De acordo com [Cereja and Magalhães 1999] polissemia é a propriedade de uma mesma palavra apresentar vários sentidos. Já palavras homônimas, tem a mesma escrita ou a mesma pronuncia, porém, são palavras diferentes, que possuem diferentes classificação gramatical, origem e significados [Lopes 2019].

Exemplo de polissemia: O ponto de ônibus é ali / Ele só marcou um ponto.

Analisando a palavra ponto nas duas sentenças acima, podemos identificar que elas são substantivos. E morfologicamente, possem o mesmo radical "pont"e a mesma vogal temática "o". Assim, não há indicação que essas palavras são diferentes e portanto, apresenta polissemia, uma palavra com mais de um sentido [Lopes 2019].

Exemplo de Homonímia: Ele casa amanhã e vai para casa domingo.

Em relação a classe gramatical da sentença acima. A primeira ocorrência de "casa"é um verbo (casar), conjugado na terceira pessoa do singular do presente do indicativo, "ele casa". Já a segunda ocorrência é um substantivo, o lugar onde ele mora. Logo, temos duas palavras diferentes que tem o mesmo nome, sendo assim classificadas como homonímia [Lopes 2019].

2.5 Wordnet

A partir da década de 1985, na Universidade de Princeton, pesquisadores, sob a liderança de George Miller, deram início ao que hoje conhecemos como Wordnet. A definição utilizada [Navega 2004] diz que o Wordnet é um banco de dados léxico no qual há informações de "palavras, palavras compostas, verbos, frases idiomáticas, relações hierárquicas entre palavras e outras propriedades". A Wordnet-PR (de Princeton), criada no início da década 1990 para a língua inglesa, vem sendo utilizada como modelo pela sua ampla utilização e adaptação a diferentes línguas [Bond and Paik 2012] , apud, [Oliveira et al. 2015].

WordNet é uma base de dados lexical amplamente utilizada. Nesta base, as palavras estão classificadas em quatro classes morfossintáticas: substantivo, adjetivo, verbo e advérbio. Além disso, as palavras são organizadas em synsets, isto é, conjuntos de formas sinônimas que representam um conceito. Por fim, os synsets estão interligados por meio de relações semânticas como hiponímia / hiperonímia, meronímia/holonímia e antonímia, formando uma taxonomia de conceitos ([Fellbaum and Miller 1998], apud, [Marinho]).

Wordnets para o português começaram a ser construídas por volta dos anos 2000. No entanto, somente era possível o acesso online e não disponível para download, impossibilitando adaptações e modificações por parte de outros pesquisadores. Dez anos depois, surge então, iniciativas de construção de Wordnets para o português brasileiro. Por conseguinte, nasceram algumas Wordnets, a exemplo disto a OpenWordnet-PT motivado pela necessidade de uma Wordnet acessível e disponível, permitindo o compartilhamento de conhecimento com a comunidade científica [Oliveira et al. 2015].

A OpenWordnet-PT, abreviada como OpenWN-PT, vem sendo desenvolvida desde 2010. Tendo como ponto de partida a Wordnet-PR, realizando mapeamento entre os synset do inglês para o correspondente em português. Além disso, informações de artigos do

Wikipedia em português são utilizados para agregar informações relevantes na OpenWN-PT [de Paiva et al. 2012].

Atualmente a OpenWN-PT, segundo [Oliveira et al. 2015], possui 43.925 synsets, entre os quais 32.696 correspondem a substantivos, 4.675 verbos, 5.575 a adjetivos e 979 advérbios. Tanto os dados[1] quanto às definições do modelo Resource Description Framework (RDF)[2] e Web Ontology Language (OWL)[3], estão disponíveis para download e também, os seus dados podem ser consultados por meio de três formas: (a) pela interface Web[4]; (b) por meio do Query Language for RDF (SPARQL[5]) Endpoint[6]; (c) via (SPARQL) utilizando uma biblioteca disponível em uma linguagem de programação.

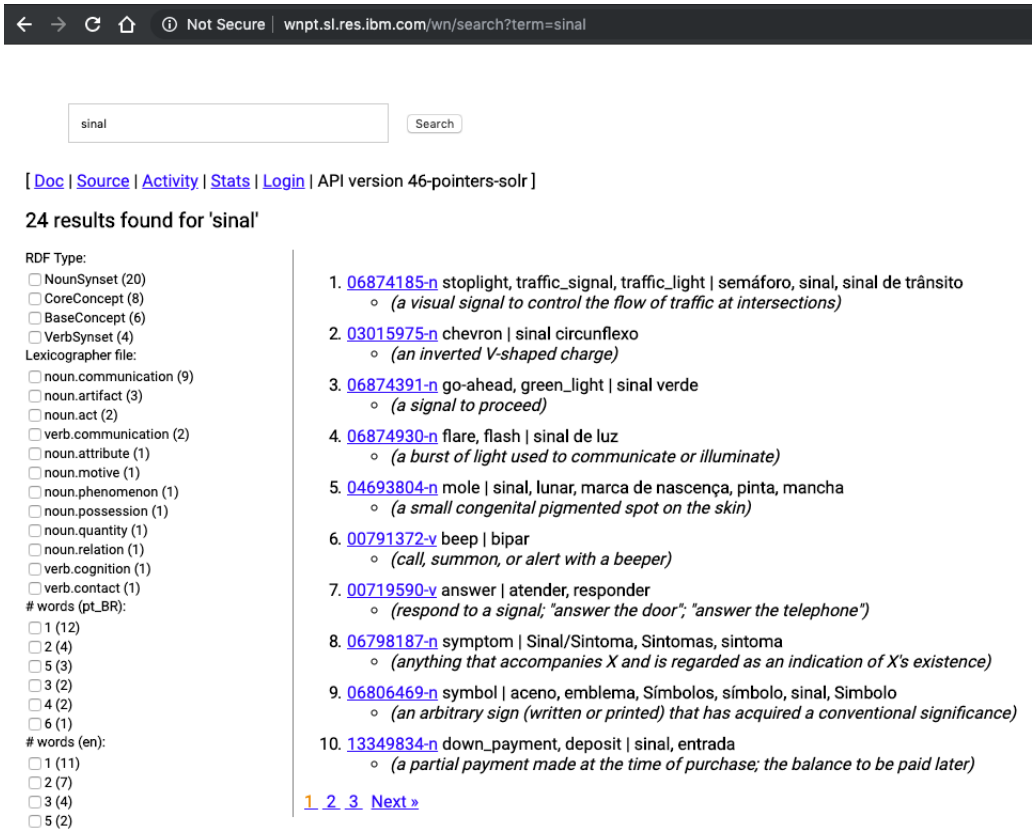

Figura 2.1: (a) Consulta pela interface Web.

A figura 2.1 mostrar o resultado da consulta pelo termo "sinal"na OpenWN-PT através da interface Web. Sendo obtidos 24 resultados de synsets que possui relação com o termo.

[1]https://github.com/own-pt/openWordnet-PT
[2]https://www.w3.org/RDF/
[3]https://www.w3.org/TR/owl-guide/
[4]http://wnpt.sl.res.ibm.com/wn/
[5]https://www.w3.org/TR/rdf-sparql-query/
[6]http://wnpt.sl.res.ibm.com:10035

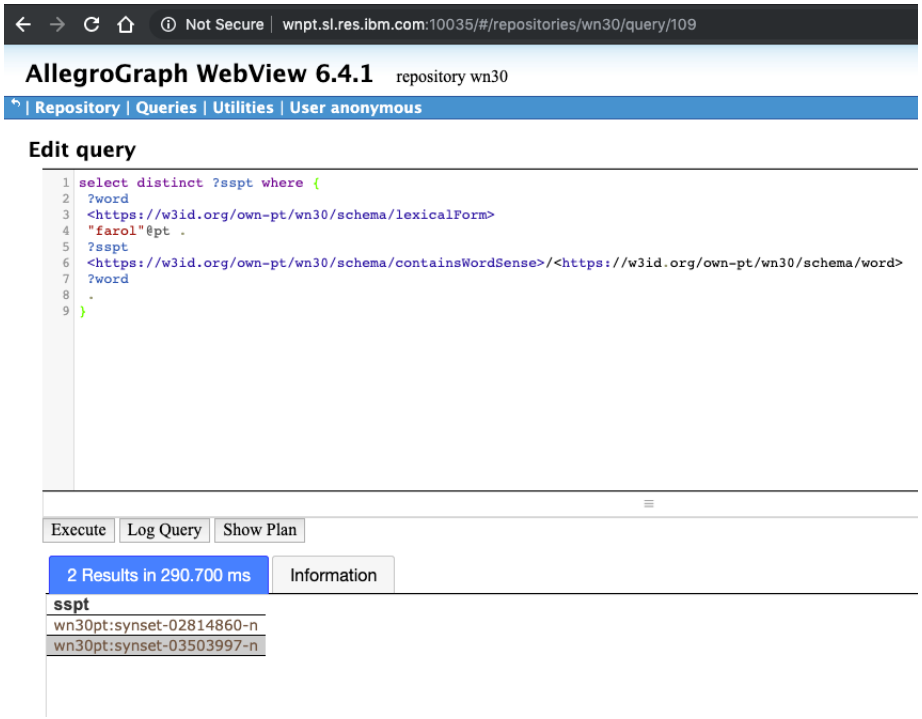

Figura 2.2: (b) Consulta por meio do SPARQL EndPoint.

A figura 2.2 mostra o resultado da consulta em SPARQL por meio do SPARQL EndPoint disponibilizado pela OpenWN-PT.

A OpenWN-PT é disponibilizada em um tipo de arquivo RDF, o que permite a interoperabilidade de sistemas que utilizam diferentes tecnologias. Em RDF é possível representar informações na Web e uma afirmação em RDF consiste em uma tripla que possui três elementos básicos: sujeito, predicado e objeto.

SUJEITO: ?word

PREDICADO: https://w3id.org/own-pt/wn30/schema/lexicalForm

OBETO: "avião"@pt

```
def synsets(self, word):
    synsets = []
    stringSparqlQuery = """
        select distinct ?sspt where {
            ?word
            <https://w3id.org/own-pt/wn30/schema/lexicalForm>
            \"""+word+"""\"@pt .
            ?sspt
            <https://w3id.org/own-pt/wn30/schema/containsWordSense>/<https://w3id.org/own-pt/wn30/schema/word>
            ?word
            .
        }
    """
    queryString = sparql.prepareQuery(stringSparqlQuery)
    syns = graph.query(queryString)
    for s in syns:
        synsets.append(str(s[0]))
    return synsets #return array of synsets
```

Figura 2.3: (c) Consulta SPARQL utilizando a blioteca RDFLib em Python.

A figura 2.3 mostrar o exemplo de uma consulta em RDF através da biblioteca RDFLib desenvolvida em linguagem Python. Dessa maneira, é possível executar consultas automatizadas a partir de um conjunto de termos.

Capítulo 3

Métodos de Análise da vitalidade dos termos do ALiB

Os métodos foram divididos em análise quantitativa e análise semântica. A subseção 3.1 descreve o conjunto de dados; 3.2 o método quantitativo; e por fim, 3.3 o método semântico.

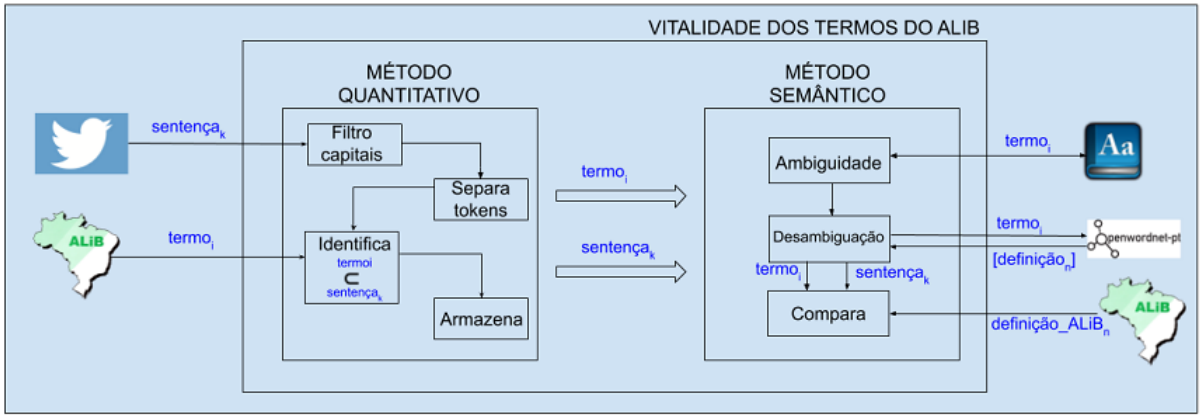

Figura 3.1: Análise da vitalidade dos termos do ALiB.

A figura 3.1 acima apresenta o esquema dos métodos utilizados para analisar a vitalidade dos termos do Projeto ALiB.

Tendo como entrada dois conjuntos de dados, tweets e termos, o método quantitativo, tem como objetivo armazenar um subconjunto desses dados, armazenando apenas os tweets que possuem a presença do termo do Projeto ALiB e divididas em quatro etapas que serão descritas com mais detalhes nas próximas sub-sessões. Nesse contexto, um tweet é compreendido como uma sentença.

A partir deste armazenamento realizado pelo método quantitativo, o subconjunto contendo os termos identificados nos tweets, é utilizado como entrada no processo do método semântico.

O método semântico tem como objetivo verificar se os termos do Projeto ALiB encontrados nos tweets, possuem o mesmo sentido empregado pelo Projeto ALiB. Assim, este método é subdividido em três etapas que serão descritas nas próximas sub-sessões.

3.1 Conjunto de dados - Tweets

O conjunto de dados foi obtido através do Twitter, por meio das APIs (Application Programming Interface) que disponibiliza os dados públicos que são compartilhados diariamente nesta plataforma digital.

O acesso a essas APIs é disponibilizado aos desenvolvedores, pelo qual, cada API fornece um tipo de informação específica. De acordo com Twitter[1], as APIs são divididas em cinco grupos principais: conta e usuários, tweets e respostas, mensagens diretas, anúncios e, por fim, ferramentas de publicação de compartilhamento.

Para ter acesso a uma API do Twitter é necessário se cadastrar em um aplicativo através do portal do desenvolvedor do mesmo. Por padrão, somente são disponibilizadas informações que os usuários escolheram compartilhar publicamente. Ao se registrar no aplicativo, o Twitter então fornece as credenciais, o que permite realizar a troca de informações.

O conjunto de dados obtido para esse trabalho, com intuito de analisar a vitalidade dos termos do Projeto ALiB no twitter, foi coletado através da API de tweets e respostas, entre o período de fevereiro a março de 2019, somando um total de 2.692.460 tweets, publicados no idioma português brasileiro.

Esses tweets, estão salvos em arquivos JSON, um formato de arquivo com informações estruturadas como atributo-valor, compacto para troca de informações simples e rápida entre sistemas. Cada linha de um arquivo JSON é representada por um tweet e seus respectivos atributos. Esse conjunto de dados foi organizado em 275 arquivos JSON, e cada arquivo possui aproximadamente 16 MB (MegaBytes), o qual, totaliza em um conjunto de dados com mais de 17 GB (GigaBytes) armazenados.


```
{"created_at": "Mon Feb 18 11:10:47 +0000 2019",
    "id":1097453369940168704,
    "id_str":"1097453369940168704",
    "text":"Desempenho da track nos charts coreanos durante 6pm - 7pm KST
    "source":"href=\"http:\/\/twitter.com\/download\/android\" rel=\"nofo
    "truncated":false,"in_reply_to_status_id":null,
    "in_reply_to_status_id_str":null,
    "in_reply_to_user_id":null,
    "in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":
```


Figura 3.2: Exemplo da estrutura de um arquivo JSON - tweet

[1]https://about.twitter.com/pt.html

3.2 Conjunto de dados - Termos do Projeto ALiB

Os dados do Projeto ALiB utilizados neste trabalho correspondem aos dados do II volume do Atlas Linguístico do Brasil publicado no ano de 2014, com dados de 25 capitais de estados brasileiros, à exceção apenas de Brasília - em vista da data de sua criação e Palmas, capital do recém criado do estado de Tocantins.

Dos questionários linguístico aplicado pelo Projeto ALiB, este trabalho utilizou o questionário semântico-lexical que possui 202 perguntas que recobrem 14 áreas semânticas, o qual contou com a colaboração de vários pesquisadores na área da linguística [do Brasil 2001].

A baixo, um exemplo das cartas semântico-lexical, aplicada aos informantes com intuito de estudar a variação linguística:

CARTA:

- MANDIOCA / AIPIM:

  ALiB: "... aquela raiz branca por dentro, coberta por uma casca marrom, que se cozinha para comer?"

- VARIAÇÕES (termos):

  INFORMANTES: "macaxeira, mandioca, aipim, mandioca brava, macaxeira brava".

As cartas e suas variações (termos) do Projeto ALiB utilizadas neste trabalho estão em anexo.

3.3 Método Quantitativo

O método quantitativo consistiu em identificar os termos do Projeto ALiB no conjunto de dados obtidos para analisar a vitalidade. Nesse sentido, o método é estruturado em 4 etapas, conforme é apresentado na figura abaixo:

A partir dos conjuntos de dados, tweets e termos do Projeto ALiB, é realizado as seguintes etapas:

1. FILTRO CAPITAIS: É filtrado os tweets dos usuários das capitais brasileiras.

   (a) Esta etapa foi realizada com intuito de analisar dados de informantes das mesmas localidades catalogadas pelo Projeto ALiB, publicado no ano de 2014 com dados de 25 de capitais de estado;

2. SEPARA EM TOKENS: É dividido em duas sub-etapas:

   (a) termos com uma única palavra, por exemplo: "macaxeira";

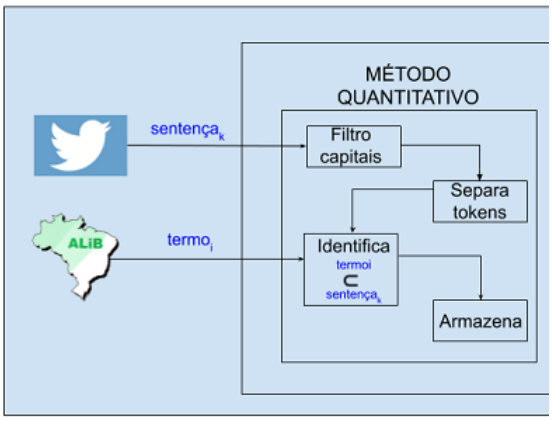

Figura 3.3: Análise da vitalidade dos termos do ALiB - Método quantitativo.

(b) termos multi-palavras, por exemplo: "chuva de gelo";

3. IDENTIFICA: Consiste na comparação da presença do termo no tweet de acordo o tipo de token: única palvra ou multi-palavra.

4. ARMAZENA: As seguintes informações são armazenadas do tweet: id, texto e localidade. E o termo do Projeto ALiB identificado no tweet: a carta e o seu termo correspondente.

   Por exemplo:
   ID do tweet: 1073616656465833987
   Carta: Amarelinha
   Termo: "avião"
   Sentença: "Sentei no banco do avião"
   Localidade: "São Paulo"
   Armazenamento: Carta, Termo, Sentença, Localidade.

3.4 Método Semântico

O método semântico consistiu em analisar os termos encontrados nos tweets, verificando se possuem o mesmo sentido empregado nas cartas semântico-lexical do Projeto ALiB. Este método é estruturado em 4 etapas conforme é apresentado na figura abaixo:

A partir do armazenamento realizado pelo método quantitativo, o método semântico tem como entrada o termo e a respectiva sentença a qual possui a presença deste termo. Assim, é realizado as seguintes etapas:

1. AMBIGUIDADE: Identifica quais são os termos ambíguos do ALiB através do dicionário[2] online da língua portuguesa;

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

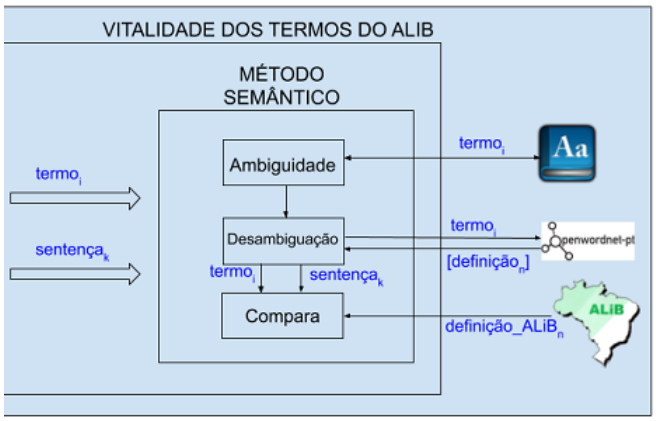

Figura 3.4: Análise da vitalidade dos termos do ALiB - Método semântico.

2. DESAMBIGUAÇÃO:
   (a) Consulta às definições e exemplos dos termos ambíguos na OpenWN-PT;
   (b) Atualiza a OpenWN-PT com definições e exemplos para os termos ambíguos do Projeto ALiB.
   (c) Desambigua o termo a partir das definições na OpenWN-PT:
      i. Termo a ser desambiguado;
      ii. Sentença que possui a presença do termo;
      iii. Consulta as definições na OpenWN-PT;
      iv. O método de Lesk define um sentido do termo na sentença.
3. COMPARA: Identifica se o termo desambiguado possui o mesmo sentido empregado pelo Projeto ALiB.

Para exemplificar: dado um termo do Projeto ALiB em uma sentença (tweet), o algoritmo de desambiguação de sentido tem como objetivo escolher o sentido correto do termo presente na sentença.

3.4.1 Ambiguidade dos termos do ALiB

A primeira etapa do método semântico foi identificar quais os termos estão presentes na OpenWN-PT. Dos 203 termos presentes em 23 cartas semântico-lexical utilizados neste trabalho, 87 estão na OpenWN-PT. E verificar os significados dos 87 termos através do Dicionário Online de Português. Assim, foram identificados 37 termos (ambíguos) com mais de um sentido e 52 (não ambíguos) com um único sentido ou que não estão presentes neste dicionário. Dos significados não encontrados, em sua grande maioria, possuem mais de uma palavra, como por exemplo: chuva de granizo, flor de banana e mandioca brava.

3.4.2 Atualização da OpenWordnet-PT

A segunda etapa deste método consistiu em inserir as definições dos termos ambíguos do Projeto ALiB que estão presentes na OpenWN-PT. Apesar dela possui 43.925 sinônimos, [Oliveira et al. 2015] apenas 10 termos foram encontrados significados do total de 203 termos do Projeto ALiB e 116 termos não estão na OpenWN-PT. No entanto, a OpenWN-PT é disponibilizada livremente para download, pelo qual, permite a sua atualização. Dessa forma, para realização do experimento 2 proposto por este trabalho foram inseridas definições para execução do método semântico.

As definições inseridas, foram obtidas a partir dos significados encontrados no dicionario online, conforme é descrito na etapa de verificação da ambiguidade. Essas definições possuem característica de frases curtas e informais que descrevem o sentido de uma palavra. Dos 37 termos presentes na OpenWN-PT e identificados como ambíguos, foram analisados 13 termos os quais possuem no mínimo 2 synsets na OpenWN-PT e inseridos 24 novas definições.

Para atualização das definições na OpenWN-PT, por exemplo, o termo "papagaio"da carta "papagaio de papel / pipa"foi atualizado com as seguintes triplas em RDF na OpenWN-PT:

Atualização para definição 1:

- SUJEITO: https://w3id.org/own-pt/wn30-pt/instances/03621473-n
- PREDICADO: https://w3id.org/own-pt/wn30/schema/gloss
- OBJETO: "pessoa que repete alguma coisa, que fala copiosamente"@pt

Atualização para definição 2:

- SUJEITO: https://w3id.org/own-pt/wn30-pt/instances/01816887-n
- PREDICADO: https://w3id.org/own-pt/wn30/schema/gloss
- OBJETO: "Objeto voador, usado como brinquedo de crianças"@pt

Todas as definições inseridas na OpenWN-PT para realização do experimento 2. Estão descritas no apêndice e as listas dos termos encontrados, não encontrados, ambíguos e não ambíguos estão em anexos.

3.4.3 Algoritmo de Lesk adaptado para o português

Na terceira etapa, é computado a desambiguação de sentido para palavra-alvo (termo) no contexto (tweet). O algoritmo adaptado de Lesk para o português é baseado na

suposição de que as palavras vizinhas de uma palavra-alvo, em um determinado contexto, tendem a compartilhar um assunto em comum. Este versão do algoritmo foi implementado em linguagem de programação Python[3] e utilizada a OpenWN-PT como repositório de sentido. Além disso, para consultar as informações na OpenWN-PT, foi necessário utilizar a biblioteca em Python RDFLib[4] para o carregamento do arquivo RDF[5] e execução das consultas em SPARQL[6].

Antes do desenvolvimento do método de Lesk, foi necessário a criação de métodos de consultas a OpenWN-PT que pudessem ser automatizadas em conjunto com o algoritmo. A figura 3.5 mostra uma consulta em RDF na OpenWN-PT, retornando todas as definições de um synset.

```
def definition(self, synset):
    definition = []
    stringSparqlQuery = """
        select distinct ?sgloss where {
            <"""+synset+"""> 
            <https://w3id.org/own-pt/wn30/schema/gloss>
            ?sgloss .
        }
    """
    queryString = sparql.prepareQuery(stringSparqlQuery)
    defis = graph.query(queryString)
    for d in defis:
        definition.append(str(d[0]).lower())
    if len(definition) > 0:
        return definition[0] #return definition of synset
    return definition = []
```

Figura 3.5: Método de consulta a OpenWN-PT

A seguir o algoritmo adaptado de lesk é apresentado as alterações para utilização da OpenWN-PT, o que tem como objetivo encontrar o sentido de uma palavra baseada no sentido das palavras vizinhas ao da palavra a ser desambiguada.

[3]https://www.python.org/
[4]https://rdflib.readthedocs.io
[5]https://www.w3.org/TR/rdf-concepts/
[6]https://www.w3.org/TR/rdf-sparql-query/

Algoritmo 1: Algoritmo de Lesk

```
1  início
2      disambiguate(word, context)
3      início
4          word_senses = wordnetpt.synsets(word)
5          best_sense = "SENSE NOT FOUND"
6          se (len(word_senses) > 0) então
7              para sense In word_senses faça
8                  signature = tokenized_gloss(sense)
9                  overlap = compute_overlap(signature, context)
10                 se (overlap > max_overlap) então
11                     max_overlap = overlap
12                     best_sense = wordnetpt.definition(sense)
13                 fim
14             fim
15         senão
16             retorna "SENSE NOT FOUND"
17         fim
18         retorna best_sense
19     fim
20     tokenized_gloss(sense)
21     início
22         tokens = set(word_tokenize(wordnetpt.definition(sense)))
23         para example In wordnetpt.examples(sense) faça
24             tokens = tokens.union(set(word_tokenize(example)))
25         fim
26         retorna tokens
27     fim
28     compute_overlap(signature, context)
29     início
30         gloss = signature.difference(stopwords)
31         retorna len(gloss.intersection(context))
32     fim
33 fim
```

Capítulo 4

Experimentos e Resultados

O presente trabalho realizou dois experimentos. O primeiro, quantitativo dos termos do Projeto ALiB no conjunto de dados do Twitter, e o segundo, análise semântica destes termos encontrados nos tweets.

4.1 Experimentos

4.1.1 Experimento 1

Os termos catalogados pelo Projeto ALiB foram publicados em 2014 nos dois primeiros volumes do Atlas Linguístico do Brasil. Nessa publicação, estão os dados coletados dos informantes de 25 capitais de estado que são utilizados neste experimento. A figura 4.1 abaixo mostra a quantidade desses termos presentes em 23 cartas semântico-lexical. E o método quantitativo é utilizado neste experimento conforme descrito na sessão 3.3 deste trabalho.

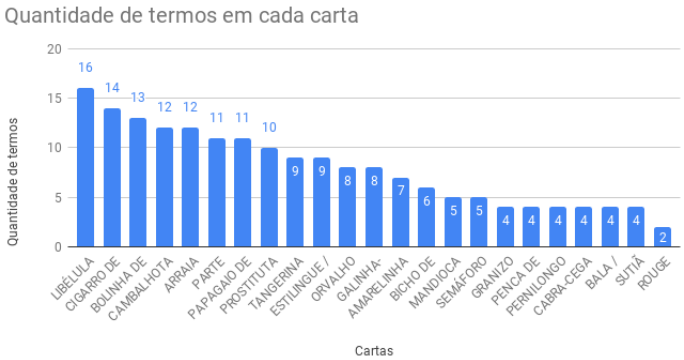

Figura 4.1: Quantidade de termos catalogado pelo Projeto ALiB por carta semântico-lexical

4.1.2 Experimento 2

O experimento 2 consistiu na avaliação da análise semântica dos termos do ALiB encontrados no Twitter. Para realização desta etapa, foi preciso identificar quais termos, são ambíguos e estão presentes na OpenWordnet-PT, o que se faz necessário para aplicação do algoritmo de desambiguação de sentido. E o método semântico é aplicado para este experimento conforme descrito na sessão 3.4.

4.2 Resultados

4.2.1 Resultado do Experimento 1

A figura 4.2 apresenta o tamanho do conjunto de dados obtidos para realização do experimento. O conjunto de dados possui 2.514.237 tweets, sendo 178.223 (6,2%) com tweets de usuários das capitais de estado.

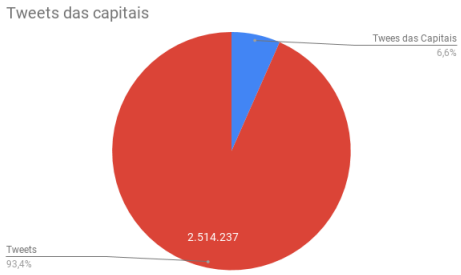

Figura 4.2: Quantidade de tweets do conjunto de dados e tweets das capitais

Dos tweets filtrados das capitais de estado, conforme a figura 4.3, 1,9% dos tweets, 3.504 termos, possuem termos do Projeto ALiB. Conforme descrito na sessão 3.2, esta etapa foi realizada com intuito de analisar dados de informantes das mesmas localidades catalogadas pelo Projeto ALiB, publicado no ano de 2014 com dados de 25 de capitais de estado.

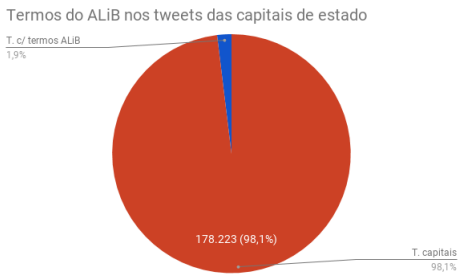

Figura 4.3: Quantidade de tweets das capitais com termos do Projeto ALiB

Dos 203 termos do ALiB, 90 termos estão distribuídos nos 3.504 tweets das capitais brasileiras. A figura 4.4 mostra os termos que tiveram até 50 ocorrências nos tweets.

Figura 4.4: Quantidade de termos do ALiB nos tweets com até 50 ocorrências

A figura 4.5 a baixo mostra os termos que tiveram mais de 50 ocorrências nos tweets.

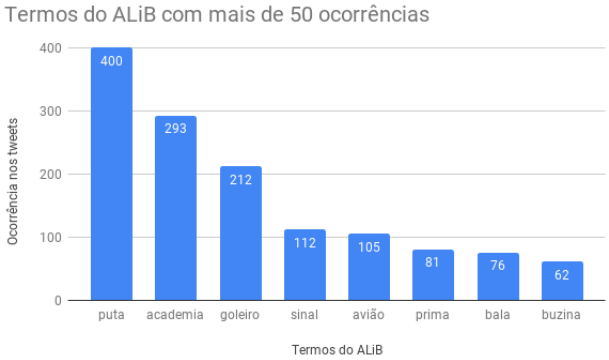

Figura 4.5: Quantidade de termos do ALiB nos tweets com mais de 50 ocorrências

Há 90 termos distintos do Projeto ALiB que estão presentes nos tweets e 113 não houveram nenhuma ocorrências. A lista completa de todos os termos ausentes e presentes estão no apêndice.

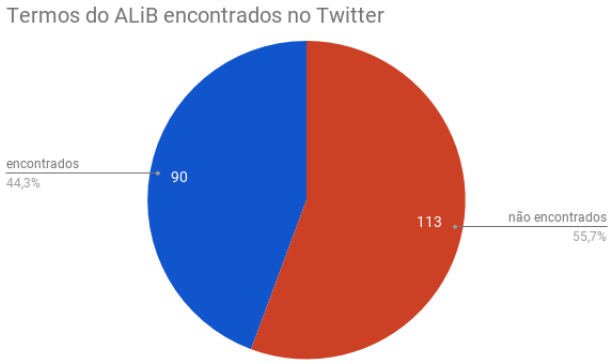

Figura 4.6: Quantidade de termos distintos presentes e ausentes nos tweets

A figura 4.7 a seguir, apresenta a ocorrência da carta "prostituta"nos tweets por capitais de estado. Sendo, a capital do Rio de Janeiro com 563, a de maior ocorrência e o estado de Alagoas com 31, a de menor ocorrência.

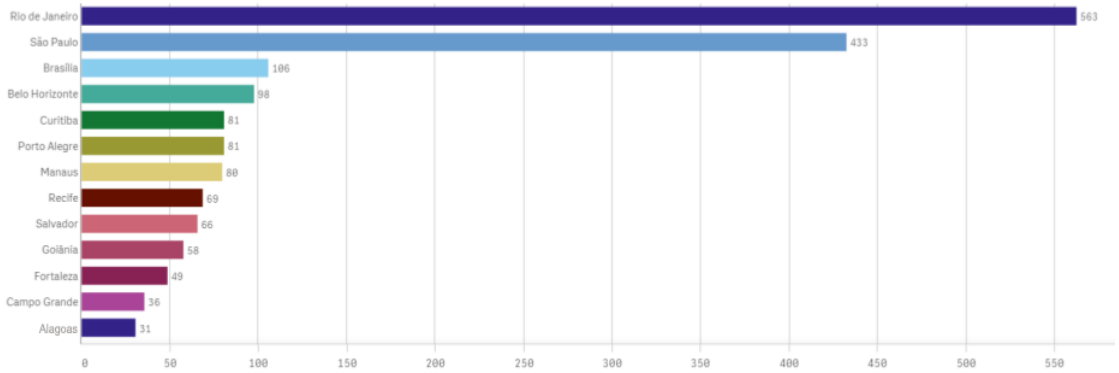

Figura 4.7: Ocorrência da carta "prostituta"por capitais de estados

4.2.2 Resultados do Experimento 2

Para obtenção dos resultados do experimento 2 foi necessário identificar quais dos termos do Projeto ALiB estão presentes na OpenWN-PT. Conforme a figura 4.7, dos 203 termos do Projeto ALiB, 91 estão presentes e 112 termos não houveram nenhum registros.

Na aplicação do método de semântico é necessário que a OpenWN-PT contenha definições, o que na wordnet é chamado de gloss. No entanto, dos 91 termos do Projeto ALiB que estão na OpenWN-PT, apenas 10 termos possuem definição. Além disso, com estas definições não foi possível executar o método semântico devido a necessidade de um

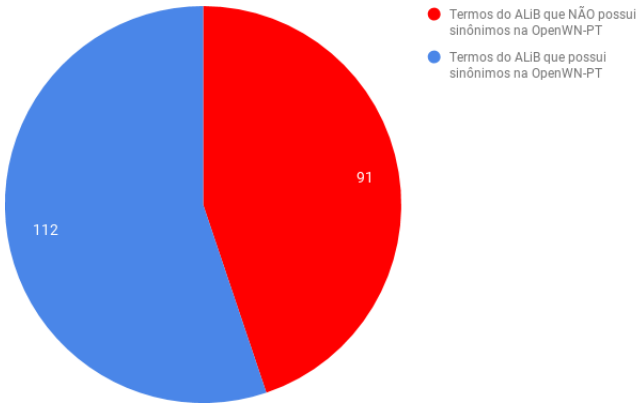

Figura 4.8: Quantidade de termos do ALiB presentes e ausentes na OpenWN-PT

termo possuir no mínimo dois synsets. A figura 4.8 apresenta os 10 termos presentes na OpenWN-PT e suas definições.

TERMO	OCORRÊNCIA OPENWN-PT	AMBÍGUO	DEFINIÇÃO OPENWN-PT	OCORRÊNCIA NO TWITTER
puta	4	NÃO	['uma mulher que se envolve em relações sexuais por dinheiro']	400
capão	2	NÃO	['frango macho castrado']	29
rapariga	9	NÃO	['mulher jovem', 'uma pessoa jovem do sexo feminino'] ['uma mulher que se envolve em relações sexuais por dinheiro']	9
rameira	1	NÃO	['uma mulher que se envolve em relações sexuais por dinheiro']	0
concha	6	NÃO	['o revestimetno duro e em grande parte calcário de um molusco ou brachiopoda']	6
garoa	2	NÃO	['um breve período de precipitação']	3
meretriz	2	NÃO	['uma mulher que se envolve em relações sexuais por dinheiro']	2
quenga	1	NÃO	['uma mulher que se envolve em relações sexuais por dinheiro']	1
avião	3	SIM	['uma mulher que parece muito atraente ou sedutora']	105
sinal	10	SIM	['movimento das mãos ou corpo para ajudar a enfatizar ou para expressar um pensamento ou sentimento']	112

Figura 4.9: os 10 termos presentes na OpenWN-PT e suas definições

Apesar da pouca quantidade de termos encontrados com significados, a OpenWN-PT é disponibilizada livremente para download, o que possibilitou a modificação para aplicação da desambiguação de sentido. Para execução desta etapa, o método semântico identificou através do dicionário Online para Português[1], 35 termos ambíguos do Projeto ALiB e 56 termos não ambíguos. Todos esses termos são listados no apêndice.

4.2.3 Resultado desambiguação OpenWN-PT atualizada

Para verificar o sentido empregado dos termos do Projeto ALiB no Twitter através do método semântico, foram selecionados dos 35 termos ambíguos, 13 termos que possuem no minímo dois synsets. Nestes, foram inseridas 24 definições na OpenWN-PT. Os 22 termos ambíguos restantes possuem um ou nenhum synset.

As novas definições inseridas na OpenWN-PT, foram criadas a partir das perguntas utilizadas pelas cartas semântico-lexical do Projeto ALiB e das consultas realizadas no dicionário Online para o Português. Para cada termo foi inserido uma definição em cada synset.

Dos 13 termos ambíguos que foram inseridos novas definições, 8 foram desambigua-

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

dos com um sentido correto e 5 termos não desambiguados. A tabela a seguir apresenta os termos e as definições em itálico que foram desambiguadas pelo método semântico.

TERMO	QUANTIDADE DESAMBIGUADA	OCORRÊNCA NOS TWEETS	SENTIDOS INSERIDOS
GOLEIRO	48	212	*Jogador de futebol que defende o gol de sua equipe e pode tocar com a mão na bola dentro da grande área.* A peça do vestuário que serve para segurar os seios.
ACADEMIA	17	293	*Local onde há aulas de ginástica, de musculação, estabelecimento e há prática de esportes.* Brincam aos saltos ou pulando com uma só perna.
AVIÃO	4	105	*Máquina voadora, que tem hélice, ou turbina.* Uma mulher que parece muito atraente ou sedutora.
PAPAGAIO	4	17	*Pessoa que repete alguma coisa, que fala copiosamente.* Objeto voador, usado como brinquedo de crianças.
BALA	4	76	*Projétil das armas de fogo.* Pequena quantidade de açúcar misturada com substâncias aromáticas e solidificada, que se deixa dissolver na boca.
SINAL	2	112	*Semáforo de trânsito.* Movimento das mãos ou corpo para ajudar a enfatizar ou para expressar um pensamento ou sentimento.
SEMÁFORO	2	8	*Semáforo de trânsito.* Telégrafo aéreo, estabelecido em pontos elevados da costa ou junto de portos.
QUEIMADO	1	1	*Que sofreu ação de queimar, que arde, incendiado, tostado.* Bala caramelada.

Figura 4.10: Termos desambiguados e os sentidos inseridos

Verificou-se através dos resultados do método semântico que os sentidos dos termos encontrados nos tweets são distintos dos significados catalogados pelo projeto ALiB ou para estes termos pouco se diversificou os assuntos dos tweets, obtendo-se apenas um sentido para cada termo pelo método semântico e/ou não definição em um sentido.

4.3 Discussões e Desafios

Atualmente o Twitter disponibiliza o acesso as APIs através do registro de um aplicativo no site do laboratório dos desenvolvedores do mesmo. No entanto, o registro do aplicativo é analisado pela equipe do Twitter, sendo necessário justificar no cadastro qual a intenção e como serão utilizados os dados obtidos. O registro só é concluído após o processo de análise, dessa forma, o presente trabalho segue todas as recomendações solicitadas pelo Twitter e pela preservação das informações dos seus usuários.

Os tweets analisados neste trabalho são mensagens publicadas de até 280 caracteres, cada mensagem é representada como um contexto na análise semântica. Porém,

outros estilos de textos sem a restrição na quantidade de caracteres, por exemplo, textos de notícias e textos do Wikipédia, com mais informações, melhoraria os resultados da tarefa de desambiguação de sentidos para o algoritmo utilizado.

O método de Lesk (1986) possui duas vantagens importantes, a primeira é que pode ser utilizada independente da língua e a segunda é a utilização de recurso externo, conforme utilizado neste trabalho a OpenWN-PT. No entanto, é dependente de um conjunto de definições e exemplos, os quais foram encontrados poucas dessas informações sobre os termos utilizados.

Em [Drude et al. 2003] os autores propõem seis fatores que descrevem o estágio de vitalidade de uma língua ou seu perigo de desaparecimento. Contudo, analisar a vitalidade de uma língua, ou, especificamente neste trabalho, analisar a vitalidade de variações de regionalismos brasileiros, demandam estudos mais aprofundados. O objetivo deste trabalho teve como escopo analisar a utilização desses termos nos novos meios de comunicação como a rede social Twitter. Não sendo possível afirmar, apenas através dos resultados obtidos nesse trabalho, que os termos catalogados pelo Projeto ALiB não são mais utilizados pelos internautas brasileiros. Além disso, outro desafio é o tratamento sobre a naturalidade dos usuários que utilizam o Twitter. É uma informação imprecisa, o que pode acarretar em enviesamento dos dados.

O presente trabalho desenvolveu métodos de análise de dados que utilizou diferentes recursos computacionais, entre eles, conjunto de dados do twitter em arquivos JSON e o repositório de sentido em triplas (RDF). Permitindo a interoperabilidade entre diferentes recursos, que mesmo não sendo diretamente um dos objetos de estudo propostos, foi possível interagir a partir da disponibilização de recursos abertos e interoperáveis.

A OpenWordnet-PT é uma das wordnets para o português brasileiro baseada na Wordnet-PR, que está constantemente em aprimoramento, além disso, é um projeto Open Source o que permite a utilização e modificação. No entanto, o método de Lesk escolhido para desambiguação de sentidos requer grande quantidade de definições existentes numa Wordnet. O que se tornou um fator limitante para atingir um dos objetivos deste trabalho, pela pouca quantidade dessas informações na OpenWordnet-PT.

A eficiência do método de Lesk é dependente de um grande conjunto de significados e exemplos presente na Wordnet utilizada, o que, mesmo com a atualização da OpenWN-PT é preciso levar em consideração alternativas para o aperfeiçoamento dos resultados.

Capítulo 5

Conclusões e trabalhos futuros

O presente trabalho buscou analisar diferentes conjuntos de dados a partir de informações coletadas pelo projeto ALiB, estudando quantitativamente e semanticamente estes dados através da rede social Twitter. Dessa maneira, este trabalho investigou conhecimentos da área da linguística e computação para atingir os objetivos propostos, estando em evidência a área de processamento de linguagem natural, no que compete a linguagem textual processada por máquinas.

O método utilizado comprovou a oportunidade de agregar conhecimento para a sub-área de processamento de linguagem natural, em específico, a tarefa de desambiguação de sentido, utilizando por si, uma abordagem baseada em conhecimento comumente utilizada como um dos primeiros algoritmos de desambiguação de sentido: o algoritmo de Lesk adaptado para OpenWN-PT, uma Wordnet para o português brasileiro.

Desse modo, os resultados aqui obtidos, puderam verificar que há grande dependência de informações obtidas nos repositórios de sentido e há necessidade de aprimorar o método utilizado com outras técnicas de desambiguação de sentido.

Como trabalho futuro, em grande potencial, pode ser considerado a predição de novos termos através de técnicas de aprendizado de máquina, o que possibilitaria aperfeiçoamento das análises propostas.

Referências Bibliográficas

[Banerjee and Pedersen 2002] Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, pages 136–145. Springer.

[Bond and Paik 2012] Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. *Small*, 8(4):5.

[Cabezudo 2015] Cabezudo, M. A. S. (2015). *Investigação de métodos de desambiguação lexical de sentidos de verbos do português do Brasil*. PhD thesis, Universidade de São Paulo.

[Cardoso and Mota 2012] Cardoso, S. A. and Mota, J. A. (2012). Projeto atlas linguístico do brasil: antecedentes e estágio atual. *Alfa: Revista de Linguística*, 56(3).

[CARDOSO 2009] CARDOSO, S. A. M. (2009). Projeto atlas linguístico do brasil - projeto alib: Descrição e estágio atual. *Revista da ABRALIN*, 18(1).

[Cereja and Magalhães 1999] Cereja, W. R. and Magalhães, T. A. C. (1999). *Gramática reflexiva: texto, semântica e interação*. Atual.

[Coelho 2018] Coelho, T. (2018). 10 fatos sobre o uso de redes sociais no Brasil que você precisa saber. `https://www.techtudo.com.br/noticias/2018/02/10-fatos-sobre-o-uso-de-redes-sociais-no-brasil-que-voce-precisa-saber.ghtml/`. [Online; acessado 19 Junho 2018].

[Cristófaro-Silva 2002] Cristófaro-Silva, T. (2002). Morte de língua ou mudança lingüistica? uma revisão bibliográfica. *Revista do Museu Antropológico*, 5/6(1).

[de Paiva et al. 2012] de Paiva, V., Rademaker, A., and de Melo, G. (2012). Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

[do Brasil 2001] do Brasil, A. L. (2001). questionário 2001/comitê nacional do projeto alib. *Londrina: Ed. UEL*.

[Drude et al. 2003] Drude, S. et al. (2003). Language vitality and endangerment. `https://unesdoc.unesco.org/ark:/48223/pf0000183699`. [Online; acessado 20 Abril 2019].

[e Silvana Araújo 2011] e Silvana Araújo, D. L. (2011). A Teoria da Variação Línguística. `http://www.vertentes.ufba.br/a-teoria-da-variacao-linguistica`. [Online; acessado 18 Abril 2019].

[Fellbaum and Miller 1998] Fellbaum, C. and Miller, G. (1998). *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press.

[Jacyra Andrade Mota 2015] Jacyra Andrade Mota, Marcela Moura Torres Paim, S. S. C. R. (2015). Documentos 5 - Projeto Atlas Linguístico do Brasil Avaliação e Perspectivas. `https://alib.ufba.br/sites/alib.ufba.br/files/documentos_5.pdf`. [Online; acessado 11 Janeiro 2019].

[Lopes 2019] Lopes, S. A. (2019). Diferenças entre polissemia e homonímia.

[Marinho] Marinho, R. S. *Desambiguação lexical de revisões de itens aplicada em sistemas de recomendação*. PhD thesis, Universidade de São Paulo.

[Navega 2004] Navega, S. C. (2004). Manipulação semântica de textos:"os projetos wordnet e lsa". *Anais do Infoimage*.

[Ngunga and Bavo 2011] Ngunga, A. and Bavo, N. N. (2011). Práticas linguísticas em moçambique: avaliação da vitalidade linguística em seis distritos. *Maputo: CEA*.

[Nóbrega 2013] Nóbrega, F. A. A. (2013). *Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento*. PhD thesis, Universidade de São Paulo.

[Nunes 2014] Nunes, N. (2014). Variação social e vitalidade de alguns regionalismos madeirenses no português falado na cidade do funchal. *Confluência*, 1(46):335–370.

[Oliveira et al. 2015] Oliveira, H. G., de Paiva, V., Freitas, C., Rademaker, A., Real, L., and Simões, A. (2015). As wordnets do português. *Oslo Studies in Language*, 7(1).

[Ribeiro 2019] Ribeiro, C. (2019). Conheça as redes sociais mais usadas no Brasil e no mundo em 2018. `https://www.techtudo.com.br/noticias/2019/02/conheca-as-redes-sociais-mais-usadas-no-brasil-e-no-mundo-em-2018.ghtml`. [Online; acessado 19 Abril 2019].

[Ribeiro 2015] Ribeiro, J. A. M. M. M. T. P. S. S. C. (2015). Documentos 5: projeto atlas linguístico do brasil, avaliações e perspectivas. *XI Workshop do Projeto Atlas Linguístico do Brasil*.

[Teixeira 2015] Teixeira, C. S. N. (2015). *Calheta e Funchal: estudo dialetal e sociolinguístico de alguns regionalismos madeirenses (comparação da sua vitalidade nos dois concelhos)*. PhD thesis.

[UNESCO 2019] UNESCO (2019). UNESCO Atlas de las lenguas en periglo en el mundo. `http://www.unesco.org`. [Online; acessado 18 Março 2019].

[Vinha 2017] Vinha, F. (2017). Faça o download do Twitter e comunique-se em 280 caracteres de qualquer lugar. `https://www.techtudo.com.br/tudo-sobre/twitter.html`. [Online; acessado 10 Janeiro 2019].

Apêndice A

Apêndice

```
1  def disambiguate(word, sentence):
2      word_senses = wordnetpt.synsets(word)
3      if (len(word_senses) > 0):
4          best_sense = word_senses[0]
5          max_overlap = 0
6          context = set(word_tokenize(sentence.lower()))
7          for sense in word_senses:
8              signature = tokenized_gloss(sense)
9              overlap = compute_overlap(signature, context)
10             if overlap > max_overlap:
11                 max_overlap = overlap
12                 best_sense = wordnetpt.definition(sense)
13     else:
14         return "SENTIDO NÃO ENCONTRADO"
15     if best_sense in word_senses:
16         return "SENTIDO NÃO ENCONTRADO"
17     return best_sense
18
19 def tokenized_gloss(sense):
20     tokens = set(word_tokenize(wordnetpt.definition(sense)))
21     for example in wordnetpt.examples(sense):
22         tokens = tokens.union(set(word_tokenize(example)))
23     return tokens
24
25 def compute_overlap(signature, context):
26     gloss = signature.difference(stopwords)
27     return len(gloss.intersection(context))
```

Figura A.1: Algoritmo de Lesk adaptado para PT

Métodos para consulta a OpenWN-PT.

```
def synsets(self, word):
synsets = []
stringSparqlQuery = """
select distinct ?sspt where {
?word
<https://w3id.org/own-pt/wn30/schema/lexicalForm>
\"""+word+"""\"@pt .
?sspt
<https://w3id.org/own-pt/wn30/schema/containsWordSense>/<https://w3id.org/own-pt/wn
?word
.
}
"""
queryString = sparql.prepareQuery(stringSparqlQuery)
syns = graph.query(queryString)
for s in syns:
synsets.append(str(s[0]))
return synsets #return array of synsets

def definition(self, synset):
definition = []
stringSparqlQuery = """
select distinct ?sgloss where {
<"""+synset+"""><https://w3id.org/own-pt/wn30/schema/gloss>
?sgloss .
}
"""
queryString = sparql.prepareQuery(stringSparqlQuery)
defis = graph.query(queryString)
for d in defis:
definition.append(str(d[0]).lower())
if len(definition) > 0:
return definition[0] #return definition of synset
return definition
```

```
def examples(self, word_sense):
examples = []
stringSparqlQuery = """
select distinct ?sexample where {
<"""+word_sense+"""><https://w3id.org/own-pt/wn30/schema/example>
?sexample .
}
"""
queryString = sparql.prepareQuery(stringSparqlQuery)
exmps = graph.query(queryString)
for e in exmps:
examples.append(str(e[0]).lower())
return examples #return array of examples
```

Lista dos termos que não foram encontrados nos TWEETS

CARTAS termos

CHUVA DE PEDRA

chuva de granizo

chuva de gelo

chuva de neve

chuva de granito

chuva de pedra

chuva de pedra de gelo

ORVALHO / SERENO

cerração

névoa

TANGERINA / MEXERICA

mexerica

maricote

laranja-cravo

tanja

carioquinha

maricote

bergamota

PENCA cacho

PARTE TERMINAL DA INFLORESCÊNCIA DA BANANEIRA / UMBIGO / CORAÇÃO

mangará

flor da banana

flor da bananeira

coração da banananeira

buzo da bananeira

mangai

coração do boi

MANDIOCA / AIPIM

mandioca brava

macaxeira brava

GALINHA-D'ANGOLA / GUINÉ / COCAR

galinha d angola

tô-fraco

capote

picote

saqué
catraia
angolista
LIBÉLULA
    libélula
bate-bunda
lava-bunda
lava-cu
jacinta
zigue-zigue
cigarra
cavalo-do-cão
lavadeira
chachimbal
catirina
mané-magro
besouro
assa-peixe
olho-de-peixe
BICHO DE FRUTA
    bicho da goiaba
gongolô
bicho da fruta
coró
PROSTITUTA
    prostituta
rampeira
rameira
CIGARRO DE PALHA
    cigarro de palha
cigarro de fumo
cigarro de tabaco
porronca
cigarro de papel
mata-rato
pacaia
boró

bororó
cigarro de corda
brejeiro
pé-de-burro
pé-duro
mato-rato
CAMBALHOTA
    carambela
bunda-canastra
pirueta
cangapé
cabriola
cambona
cambota
maria-escambonaaú
BOLINHA DE GUDE
    bola de gude
bolinha de gude
bola de vidro
bolinha de vidro
biloca
bolita
bola de fona
cabeçulinha
marraio
ximbra
búrica
ESTILINGUE / SETRA / BODOQUE
    badogue
estilete
peteca
setra
PAPAGAIO DE PAPEL/PIPA
    cangula
curica
papeta
PIPA/ARRAIA

curica
caixotinho
caixotinha
capocheta
jereco
CABRA-CEGA
cabra-cega
cobra-cega
pata-cega
pega-pega
gata-cega
AMARELINHA
amarelinha
BALA / CONFEITO / BOMBOM
confeito
SUTIÃ
califom
porta-seio
ROUGE
carmim
vermelho
SINALEIRO / SEMÁFORO / SINAL
sinaleiro
sinaleira
luminoso

LISTA DOS TERMOS PRESENTES NA OPENWN-PT

CARTAS TERMOS

CHUVA DE PEDRA chuva de granizo

ORVALHO / SERENO sereno

neblina

garoa

neve

nevoeiro

névoa

fumaça

TANGERINA / MEXERICA mexerica

bergamota

mimosa

tangerina

PENCA penca

palma

cacho

concha

PARTE TERMINAL DA INFLORESCÊNCIA DA BANANEIRA / UMBIGO / CORAÇÃO umbigo

pendão

pêndulo

buzina

MANDIOCA / AIPIM macaxeira

mandioca

aipim

GALINHA-D'ANGOLA / GUINÉ / COCAR capote

capão

saqué

galinhola

LIBÉLULA libélula

helicóptero

cigarra

lavadeira

cavalo

macaco

besouro

BICHO DE FRUTA larva

lagarta
broca
PERNILONGO mosquito
praga
PROSTITUTA mulher
puta
meretriz
rameira
rapariga
prima
quenga
biscate
CAMBALHOTA carambola
mortal
cambota
perereca
BOLINHA DE GUDE bola de gude
peteca
ESTILINGUE / SETRA / BODOQUE estilingue
baladeira
atiradeira
funda
estilete
peteca
setra
PAPAGAIO DE PAPEL/PIPA papagaio
pipa
raia
pandorga
coruja
quadrado
PIPA/ARRAIA pipa
papagaio
raia
avião
periquito
CABRA-CEGA pega-pega

AMARELINHA amarelinha

macacão

macaco

academia

maré

avião

BALA / CONFEITO / BOMBOM bala

bombom

caramelo

queimado

SUTIÃ sutiã

goleiro

ROUGE blush

carmim

SINALEIRO / SEMÁFORO / SINAL sinal

semáforo

farol

luminoso

LISTA DOS TERMOS AMBÍGUOS

CARTAS	TERMOS
CIGARRO DE PALHA	palheiro
	pito
CAMBALHOTA	cambalhota
	carambola
BOLINHA DE GUDE	birosca
	bila
PAPAGAIO DE PAPEL / PIPA	papagaio
	pipa
	raia
	pandorga
	rabiola
	coruja
	peixinho
PIPA / ARRAIA	avião
	periquito
	ratinho
	bolachinha
AMAR ELINHA	macaca
	macacão
	macaco
	academia
	cancão
	maré
BALA / CON-FEITO / BOMBOM	bala
	bombom
	queimado
SUTIÃ	sutiã
	corpete
	goleiro
ROUGE	ruge
	blush
SINALEIRO / SEMÁFORO / SINAL	sinal
	semáforo
	farol
	pito

Apêndice B

Anexos

	A	B	C	D	E	F	G	H	I	J
1	**carta**	**pergunta**								
2	CHUVA DE PEDRA	Durante uma chuva, podem cair bolinhas de gelo. Como chamam essa chuva?								
3	ORVALHO / SERENO	De manhã cedo, a grama geralmente está molhada. Como chamam aquilo que molha a grama?								
4	TANGERINA / MEXERICA	... as frutas menores que a laranja, que se descascam com a mão, e, normalmente, deixam um cheiro na mão? Como elas são?								
5	PENCA	... cada parte que se corta do cacho da bananeira para pôr para madurar/amadurecer?								
6	PARTE TERMINAL DA INFLORESCÊNCI	... a ponta roxa no cacho da banana?								
7	MANDIOCA / AIPIM	... aquela raiz branca por dentro, coberta por uma casca marrom, que se cozinha para comer?								
8	GALINHA-D'ANGOLA / GUINÉ / COCAR	... a ave de criação parecida com a galinha, de penas pretas com pintinhas brancas?								
9	LIBÉLULA	... o inseto de corpo comprido e fino, com quatro asas bem transparentes, que voa e bate a parte traseira na água?								
10	BICHO DE FRUTA	... aquele bichinho branco, enrugadinho, que dá em goiaba, em coco?								
11	PERNILONGO	... aquele inseto pequeno, de perninhas compridas, que canta no ouvido das pessoas, de noite? Imitar o zumbido.								
12	PROSTITUTA	... a mulher que se vende para qualquer homem?								
13	CIGARRO DE PALHA	Que nomes dão ao cigarro que as pessoas faziam antigamente, enrolado à mão?								
14	CAMBALHOTA	... a brincadeira em que se gira o corpo sobre a cabeça e acaba sentado? Mímica.								
15	BOLINHA DE GUDE	... as coisinhas redondas de vidro com que os meninos gostam de brincar?								
16	ESTILINGUE / SETRA / BODOQUE	... o brinquedo feito de uma forquilha e duas tiras de borracha (mímica), que os meninos usam para matar passarinho?								
17	PAPAGAIO DE PAPEL/PIPA	... o brinquedo feito de varetas cobertas de papel que se empina no vento por meio de uma linha?								
18	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?								
19	CABRA-CEGA	... a brincadeira em que uma criança, com os olhos vendados, tenta pegar as outras?								
20	AMARELINHA	... a brincadeira em que as crianças riscam uma figura no chão, formada por quadrados numerados, jogam uma pedrinha (mímica) e vão pulando com uma perna só?								
21	BALA / CONFEITO / BOMBOM	... aquilo embrulhado em papel colorido que se chupa? Mostrar.								
22	SUTIÃ	... a peça do vestuário que serve para segurar os seios?								
23	ROUGE	... aquilo que as mulheres passam no rosto, nas bochechas, para ficarem mais rosadas?								
24	SINALEIRO / SEMÁFORO / SINAL	Na cidade, o que costuma ter em cruzamentos movimentados, com luz vermelha, verde e amarela?								

Figura B.1: Cartas semântico-lexical do Projeto ALiB

	A	B	C	D	E	F	G	H	
1	GRANIZO	chuva de granizo	chuva de gelo	chuva de pedra	chuva de neve	chuva de granito	chuva de pedra de gelo		
2	ORVALHO	sereno	neblina	garoa	neve	cerração	nevoeiro	névoa	fu
3	TANGERINA	mexerica	poncã	mariconte	laranja-cravo	tanja	carioquinha	maricote	b
4	PENCA DE BANANA	penca	palma	cacho	concha				
5	EXTREMIDADE DA INFLORESCÊNCIA DA BANANEIRA	mangará	umbigo	flor da bananeira	flor da banana	pendão	buzo da bananei	coração da bana	m
6	AIPIM	macaxeira	mandioca	aipim	mandioca brava	macaxeira brava			
7	GALINHA D'ANGOLA	galinha d'angola	tô-fraco	capote	guiné	picote	capão	saqué	c
8	LIBÉLULA	libélula	helicóptero	bate-bunda	lava-bunda	lava-cu	jacinta	zigue-zigue	c
9	BICHO DA GOIABA	bicho da goiaba	larva	tapuru	lagarta	broca	gongolô	bicho da fruta	c
10	PERNILONGO	pernilongo	mosquito	muriçoca	carapanã	praga			
11	PROSTITUTA	prostituta	mulher	puta	meretriz	rameira	rampeira	garota de progra	p
12	CIGARRO DE PALHA	cigarro de palha	cigarro de fumo	cigarro de tabac	porronca	palheiro	cigarro de papel	mata-rato	p
13	CAMBALHOTA	cambalhota	carambela	carambola	bunda-canastra	pirueta	mortal	cangapé	c
14	BOLINHA DE GUDE	bola de gude	bolinha de gude	peteca	bola de vidro	bolinha de vidro	biloca	birosca	b
15	ESTILINGUE	estilingue	baladeira	atiradeira	badogue	funda	atiradeira	estilete	p
16	BRINQUEDO DE EMPINA (COM VARETAS)	papagaio	pipa	raia	pandorga	cangula	curica	rabiola	p
17	BRINQUEDO DE EMPINA (SEM VARETAS)	pipa	papagaio	curica	raia	avião	periquito	ratinho	b
18	CABRA-CEGA	cabra-cega	cobra-cega	pata-cega	pega-pega	gata-cega			
19	AMARELINHA	amarelinha	macaca	macacão	macaco	academia	cancão	maré	a
20	BALA	bala	bombom	caramelo	confeito	queimado			
21	SUTIÃ	sutiã	corpete	califom	porta-seio	goleiro			
22	RUGE	ruge	blush	carmim					
23	SEMÁFORO	sinal	semáforo	sinaleiro	farol	sinaleira	luminoso		

Figura B.2: Variações das cartas