ANÁLISE DOS TERMOS "DOR" E "GUAPO" PRESENTES NO ATLAS LINGUÍSTICO GALEGO E SUA VITALIDADE NO *TWITTER*: UMA PROPOSTA METODOLÓGICA[1]

ANALYSIS OF THE LINGUISTIC TERMS "DOR" AND "GUAPO" FROM THE GALEGO LINGUISTIC ATLAS AND ITS VITALITY IN TWITTER: A METHODOLOGICAL PROPOSAL

Daniela Barreiro Claro[2]
Universidade Federal da Bahia – Instituto de Computação

Silvana Soares Costa Ribeiro[3]
Universidade Federal da Bahia – Instituto de Letras

Luis Emanuel Neves de Jesus[4]
Universidade Federal da Bahia – Instituto de Computação

Resumo: O artigo tem por objetivo apresentar uma análise comparativa, com abordagem quantitativa e diatópica, referente a termos registrados no Atlas Linguístico Galego em relação aos coletados no *Twitter*. Especificamente, pretende-se analisar a vitalidade dos termos que constam no ALGa (volume V), verificando-se se tais termos continuam sendo utilizados para se comunicar nos *tweets*. Para concretização do objetivo, desenvolveu-se uma metodologia específica que foi testada com os dados selecionados. Os resultados obtidos revelam que é possível analisar a vitalidade de alguns termos, mas que alguns ajustes metodológicos são necessários a fim de alcançar o objetivo com os termos do ALGa.

Palavras-chave: Atlas; ALGa, Vitalidade; *Twitter*; Dialetologia.

[1] Agradecemos ao Instituto de Língua Galega da Universidade de Santiago de Compostela, na pessoa dos pesquisadores Rosário Alvarez Blanco, Maria Alvarez de la Granja, Xose Luis Regueira Fernandez e Xulio Cesar Sousa Fernandez, pelas contribuições fornecidas durante a realização deste trabalho e a leitora de galego na UFBA Araceli Luna Magarinõs durante a fase de análise dos dados.

[2] dclaro@ufba.br

[3] silvanar@ufba.br

[4] luis.emanuel@ufba.br

Abstract: *The article aims to present a comparative analysis, with a quantitative and diatopic approach, referring to terms registered in the Galician Linguistic Atlas in relation to those collected on Twitter. Specifically, it is intended to analyze the vitality of the terms contained in the ALGa (volume V), checking if these terms continue to be used to communicate in tweets. To achieve the objective, a specific methodology was determined and tested with the selected data. The results obtained reveal that it is possible to analyze the vitality of some terms, but that some methodological adjustments are made available in order to achieve the objective with the terms of the ALGa.*

Keywords: *Atlas; Vitality; Twitter; Dialetology; ALGa.*

INTRODUÇÃO

A linguagem natural é a maneira mais comum de se comunicar, seja pela fala ou pela escrita. A sociolinguística e a dialetologia estudam os padrões de comportamento linguístico dos falantes dentro de uma comunidade de fala. Diversos países e regiões realizaram o mapeamento das línguas e revelaram os resultados de pesquisa por meio de atlas linguísticos: produtos cartográficos, cujo principal objetivo é descrever a realidade linguística daquela região, de um país ou de um continente.

No início do século XX, ocorre a publicação do *Atlas Linguistique de la France* (*ALF*) de Jules Gilliéron (1902-1910), considerado o marco decisivo de consolidação da geolinguística. Tal publicação, além de dar início à cartografia linguística, promoveu a abertura de trilhas para a realização e publicação de trabalhos nesse campo, tanto na Europa, inicialmente, quanto no mundo.

Como dito, um dos elementos motivadores para a pesquisa dialetal foi a descrição da realidade linguística dos países. Desta forma, constata-se que os primeiros atlas linguísticos elaborados no mundo foram de caráter nacional, ou seja, retrataram a realidade linguística da língua majoritária de um país, circunscrita aos limites político-geográficos estabelecidos e não aos "limites expansivos" de uma dada língua. Vejam-se, por exemplo: (i) na Europa, um atlas como o *Atlas Linguistique de la France* (*ALF*) (GILLIÉRON. J.; EDMONT, E., 1902-1910); o *Sprach- und Sachatlas Italiens und der Südschweiz* (JABERG, K.; JUD. J., 1928-1940) ou o *Atlasul lingvistic român* (PETROVIC, 1956-1972); e (ii) na América,

o *Atlas lingüístico-etnográfico de Colombia* (FLÓREZ, 1981-1983) ou *Atlas Linguístico de México* (LOPE-BLANCH, 1990). Cada atlas nacional elaborado promoveu um melhor conhecimento linguístico e social dos países.

Outros atlas linguísticos foram sendo produzidos a partir daí e avançaram para a descrição de grandes áreas. Em 1998, surge, na Europa, a publicação de um atlas continental: o *Atlas Linguarum Europae* (*ALE*). Há também atlas de família de línguas, um exemplo de mapeamento das línguas românicas na Europa é o *Atlas Linguistique Roman – ALiR – (1996*), enquanto que há também atlas que descreveram línguas românicas em particular, tais como: Atlas Linguístico do Brasil (português), na América do Sul, e Atlas Linguístico do Galego, na Europa.

A partir de 1996, o Projeto Atlas Linguístico do Brasil (Projeto ALiB) foi criado, tendo o início da elaboração de instrumentos metodológicos e solicitação de recursos para execução se concretizado a partir de 1997. A pesquisa por meio dos inquéritos linguísticos realizou-se de 2001 a 2013. No ano de 2014, o Projeto ALiB publicou os seus dois primeiros volumes, que correspondem à análise da variação diatópica referente aos dados de falas de entrevistados oriundos de 25 capitais dos estados do Brasil considerados (CARDOSO et al., 2014a; 2014b).

O Atlas Linguístico Galego - ALGa (GARCÍA et al., 2005) iniciou a sua pesquisa em 1974. A coleta dos dados começou há mais de três décadas. Desde então, cada termo foi estudado em diversos níveis de análise da língua. O primeiro volume do ALGa foi lançado em 1990 e correspondeu ao estudo da morfologia verbal.

Durante a última década do século XX, observou-se um crescimento exorbitante das redes sociais que influenciam a maneira do falante se comunicar através da escrita. As redes sociais impuseram formas próprias de comunicação, incluindo aspectos específicos do léxico e da morfologia das línguas, dentre elas a Língua Galega. Assim como foi realizado com o Atlas Linguístico do Brasil

(NUNES et al., 2020), um novo desafio emergiu a partir da leitura do ALGa: a necessidade de avaliar a vitalidade linguística dos termos presentes no ALGa, com coleta de dados da década de 70 do século XX, e nas redes sociais, segunda década do século XXI, para o trabalho em curso. A rede social escolhida foi o *Twitter* devido à sua independência de domínio, seus documentos de texto e a localidade onde habitam as pessoas que o utilizam.

Neste primeiro momento, somente o Volume V do ALGa foi utilizado, o qual corresponde aos dados que se referem à coleta com base em questões que cobriam a área temática "o ser humano" com informações sobre as partes do corpo, as doenças, as ações e as características físicas dos informantes galegos situados na Galícia. O volume V do ALGa possui 300 cartas e foi constituído por 167 entrevistas realizadas em todo o território de fala galega entre os anos de 1974 e 1977. No ALGa, há 25.291 termos lexicais distintos que foram utilizados nas pesquisas no *Twitter*. Diante da quantidade de termos registrados no ALGa e uma prévia análise por uma nativa falante galega[5], elegeram-se para este trabalho sete cartas do ALGa: Carta 2 - *Cabeza* (cabeça), Carta 111 - *Dor* (dor), Carta 83 - *Esgarro* (cuspe), Carta 16 - *Nariz* (nariz), Carta 85 - *Guapo* (bonito), Carta 31 - *Pescozo* (pescoço) e Carta 104 - *Trenza* (trança de cabelo).

Assim, este trabalho analisou a vitalidade linguística dos termos destas sete cartas por meio de uma metodologia proposta para ser executada em quatro etapas utilizando os dados do *Twitter*, obtidos entre os meses de abril e setembro de 2020. A primeira etapa foi uma análise da língua Galega nos *tweets* que foram recuperados. A validação da língua descrita no *Twitter* foi realizada através de redes neurais artificiais[6]. A segunda etapa foi uma análise lexical que visava a obter os termos mais frequentes no *Twitter* que estavam presentes no ALGa para selecionar as cartas que seriam analisadas. Devido à alta ocorrência dos termos

[5] Araceli Luna Magariños foi leitora de Galego na UFBA.

[6] Redes neurais artificiais são modelos computacionais inspirados pelo sistema nervoso central que são capazes de realizar aprendizado de máquina para um reconhecimento de padrões.

com pouco significado semântico e a pouca seletividade dos termos em relação aos *tweets*, foram estabelecidos alguns filtros para que melhor se obtivessem os *tweets*, como se descreve na seção seguinte. A terceira etapa foi uma análise semântica que buscou a identificação do sentido dos termos como registrados no ALGa para analisar a correspondência de tais termos nos *tweets*. Essa etapa foi uma análise manual (*tweet* a *tweet*) com intuito de comprovar o mesmo significado do termo empregado no *tweet* e no ALGa.

Por fim, foi realizada uma análise numa perspectiva diatópica da vitalidade dos termos referentes às localidades informadas nos *tweets*, estabelecendo, assim, uma comparação geolocalizada com o ALGa. Os resultados obtidos destacaram que o processo metodológico precisa de novos ajustes para melhor refletir a realidade da língua Galega nos *tweets*. Além disso, observou-se que alguns termos não estão sendo utilizados nas redes sociais do *Twitter* com o sentido registrado no ALGa. Há casos, por exemplo, de termos que são homônimos na língua e que só o processamento de língua pelos humanos é capaz, num primeiro momento, de estabelecer a distinção e solucionar as ambiguidades. Por outro lado, alguns termos catalogados pelo ALGa continuam sendo empregados pelos galegos no *Twitter*.

O presente trabalho está organizado como segue: a seção 1 apresenta os trabalhos mais relacionados no que se refere à vitalidade dos termos. A seção 2 descreve a metodologia proposta que corresponde à aquisição dos dados e validação da língua galega, assim como à análise dos dados. A seção 3 apresenta os resultados diferenciando-os em quantitativo e diatópico, trazendo uma comparação entre os achados nos *tweets* e no ALGa; a seção 4 discute alguns dos aspectos observados durante o desenvolvimento deste trabalho e aborda algumas hipóteses em relação aos resultados obtidos e por fim, a seção 5 apresenta as considerações finais deste trabalho.

1 TRABALHOS RELACIONADOS

O principal trabalho relacionado a este é o artigo publicado por (NUNES et al., 2020), que descreve a vitalidade dos termos registrados no ALiB também utilizando o *Twitter* como rede social. Um subconjunto de cartas foi selecionado e os valores semânticos de cada termo presente nas cartas foram analisados quanto à correspondência dos termos presentes no *Twitter*. Para analisar a desambiguação dos sentidos, o *framework OpenWordNet-PT* (DE PAIVA et al., 2012) foi utilizado como meio semiautomatizado para auxiliar no processo.

Outro trabalho mais próximo referente à vitalidade foi apresentado por Peres (2011) que realizou uma análise da vitalidade da língua italiana em uma pequena comunidade de Araguaia, distrito de Marechal Floriano, Espírito Santo, tendo como objetivo analisar os fatores que levam à manutenção ou à perda de uma língua minoritária. Similarmente ao trabalho do Nunes et al. (2020), Peres (2011) também norteia sua investigação no documento *Language Vitality and Endangerment* da UNESCO 2003 (DRUDE, 2003). No entanto, ele difere deste trabalho principalmente por duas questões: a população escolhida é bem reduzida, imigrantes italianos residentes em uma pequena região do Brasil e a abordagem de análise da vitalidade foi feita com a aplicação de questionários com temas referentes ao histórico e à vida social dos italianos.

Alguns outros estudos anteriores, trabalharam a linguagem e a localização. Quinn (2016) propôs correlacionar as linguagens com os falantes e suas influências baseado nos perfis dos usuários em um determinado sistema. Montgomery & Stoeckle (2013) propuseram um estudo sobre a variação da linguagem, incluindo primeiramente a Europa através da geração manual de um mapa a fim de traçar as delimitações dos dialetos empregados.

Pavalanathan et al. (2015) fizeram uma análise dos efeitos na localização determinada pelo usuário ou identificada por meio de um GPS no *Twitter*. Um dos principais achados deste trabalho foi que as localizações que são informadas

manualmente têm uma melhor acurácia em termos de trabalhos que usam essa variável. Além disso, analisaram as idades e os gêneros que utilizam prioritariamente as informações de localização no Twitter, e observaram que os homens e os usuários de idade mais avançadas tendem a descrever manualmente a sua localização. Diferentemente do presente trabalho, nenhum desses analisam a vitalidade dos termos em relação a um Atlas Linguístico.

Assim, o presente trabalho tem como objetivo analisar a vitalidade de variantes lexicais do Galego que estão presentes no ALGa e que também ocorrem no *Twitter*.

2 METODOLOGIA PROPOSTA

A vitalidade dos termos referentes à área temática "o ser humano" apresentados no Atlas Linguístico Galego e presentes no *Twitter* foi analisada em duas principais abordagens: quantitativa e por localidade. O método tem caráter de pesquisa descritiva com foco em caracterizar certo fenômeno, que neste caso, é a detecção dos termos registrados no ALGa que são utilizados no *Twitter*, especificamente aqueles presentes em sete cartas do ALGa que foram analisadas.

A proposta metodológica aplicada consistiu em quatro etapas, após a aquisição dos *tweets*, denominadas: (1) validação da língua Galega; (2) tratamento dos dados, além da validação dos dados pré-processados por especialistas da área de Letras e Computação; (3) análise quantitativa que engloba a análise lexical e semântica dos termos registrados no ALGa com o mesmo significado dos termos registrados nos *tweets* e (4) análise diatópica da vitalidade dos termos, comparando-os com a ocorrência na mesma localidade no ALGa.

2.1 Aquisição dos dados e validação da língua Galega

O conjunto de dados adotado para aplicação do método foi delimitado pelos termos registrados no Atlas Linguístico Galego e em um conjunto de postagens realizadas no *Twitter*.

No que tange aos aspectos do Atlas, os pesquisadores da Galícia disponibilizaram uma planilha originada a partir do conteúdo das cartas linguísticas do ALGa e que contém os termos registrados na publicação. A planilha é composta por 7 colunas e 25.291 linhas, as quais alocam as informações coletadas referentes à variação linguística utilizada nos seis volumes já publicados do ALGa. Somente o volume V contém 8288 termos.

Em relação ao *Twitter*, os dados foram coletados em postagens que abrangiam o período de 23 de abril de 2020 a 28 de setembro de 2020 com intervalos de execução, totalizando oito coletas. Com o auxílio de um algoritmo de programação na linguagem Python[7] e com o uso de um modelo de representação de palavras multilingual denominado *fastText* (JOULIN et al., 2016) foram detectados os *tweets* que de fato eram da língua Galega. Nestes intervalos foram filtrados 10.000 *tweets* da língua Galega que correspondem a aproximadamente 44 GB de armazenamento para os *tweets* em Galego.

A aquisição destes dados na sua forma direta, sem nenhum tratamento, impõe desafios para normalização tanto nos aspectos textuais quanto nos aspectos lexicais. Fez-se necessária a realização do tratamento dos dados específicos (diferenciando conjuntos textuais e conjuntos lexicais). Não há ainda na literatura uma metodologia pré-estabelecida para realização de tal tratamento. A metodologia aqui desenvolvida será abordada na seção 2.2

[7] Disponível em: https://www.python.org/. Acesso em: mar. 2021.

2.2 Tratamento dos dados

Como os conjuntos de dados possuíam características heterogêneas, técnicas computacionais foram aplicadas para o refinamento das configurações dos dados com o objetivo de prepará-los para adotar uma metodologia para a extração de informação.

Nos *tweets* foram avaliados os aspectos inerentes à integração das informações disponibilizadas pelo atlas e as que correspondiam ao objetivo da investigação. Dentre os dados gerais dos *tweets* foram selecionados: (i) o texto do *tweet*; e (ii) a respectiva localização geográfica do usuário que realizou a publicação (*o post)*. Os demais dados foram descartados inicialmente, mas podem ser utilizados em pesquisas futuras.

Durante o arranjo dos dados e a junção de planilha ALGa[8] e *tweets* coletados, foi necessário promover uma limpeza na planilha ALGa, notou-se que alguns termos possuíam algumas características específicas, e para o tratamento foram definidas cinco regras de aplicação para excluí-los:

1) Remoção de termos com tamanho menor que 3 caracteres;
2) Remoção de termos nos quais ocorriam sufixo, prefixo ou radical isolado;
3) Remoção de termos com caracteres especiais;
4) Remoção de termos duplicados e
5) Remoção de termos que são classificados como conjunção e/ou preposição (conectivos).

A regra 1) corresponde à remoção de termos com tamanho menor que 3 caracteres, tais como: 'pé' (vol. 5/mapa 73); 'ca' (vol. 2/mapas 313 e 204); 'ao' (vol. 2/mapa201); 'un' (vol. 3/mapa 41; (vol. 2/mapa 218; (vol. 3/mapa 262); 'os' (vol. 2/mapa

[8] A planilha considerada, nesse momento da coleta de dados contemplava os sete volumes do ALGa.

200). Isso ocorreu, pois, esses termos estavam diminuindo o fator de seletividade[9] dos *tweets* no momento da consulta.

A regra 2) correspondeu à remoção dos termos em que apareciam raiz, sufixos ou prefixos, itens que estavam presentes, sobretudo, nos volumes 1 e 2 que tratam de morfologia. Seguem exemplos retirados do volume II: '-doiro' (mapas 118/119); '-eco' (mapa 89); '-meles' (mapa 78); '-ás' (mapas 64, 62 e 74) e 'íns' (mapa 67).

A regra 3) removeu alguns dos termos que tinham caracteres especiais, tais como: + e. Esses itens dificultaram o processo de busca de suas ocorrências, por não possuírem valores semânticos imediatamente detectáveis, relacionavam-se às possibilidades combinatórias. Seguem exemplos retirados do volume V: 'v. + fatigado' / 'v. + asfixiado'' / v. + afatigado' / 'v. + sofocado' presentes no mapa 220.3 e 'v. + a rodar' localizado no mapa 249.1.

A regra 4) removeu os termos duplicados, deixando na base somente uma ocorrência de cada. Cada termo poderia ter ocorrido em uma ou "n" localidades, e dessa forma figurava na planilha mais de uma vez. Deixou-se apenas um registro para cada carta. Apresentam-se alguns exemplos do exposto retirados dos volumes V e VI: 'alto' (vol. 6/mapa 15); 'aloqueiro' (vol. 6/mapa 60.2); 'ameixa' (vol. 5/mapa 80.3); 'órgano' (vol. 5/mapa 77) e 'vagante' (vol. 6/mapa 176).

Por fim a regra 5) removeu alguns conectivos com ou sem aglutinação com artigos. As ocorrências localizadas nos volumes II e III foram: 'como' (vol. 3/mapa 399); 'que' (vol. 2/mapa 366); 'para' (vol. 2/mapa 208) e 'por' (vol. 2/mapa 357). Itens que faziam parte do conjunto de termos do ALGa foram preservados, a exemplo de 'nas', tais como a aglutinação da preposição 'em' com o artigo definido 'a' (plural 'as') = 'nas' que é uma das ocorrências de variação lexical para

[9] Fator de seletividade representa o percentual de *tweets* que podem ser retornados em virtude de um termo. Quanto maior (100%), melhor é a identificação do *tweet* dado para aquele termo.

'nariz' registrada na carta 16 (*Nariz*) ALGa (vol. V). Logo, certos conectivos ou casos de preposições aglutinadas não foram removidos no pré-processamento caso se caracterizassem como ocorrência de homonímia.

Como mostra a Tabela 1, o tratamento dos dados de todos os termos do Atlas gerou uma redução de 21,59% dos termos, diminuindo assim o espaço amostral para 19.832 itens. Consequentemente, o tratamento dos dados do volume V gerou uma redução de 11,41% dos termos, culminando em um espaço amostral de 7.343 termos.

Tabela 1: Quantitativo das ocorrências após aplicação das regras de filtragem

Regra	**Exemplo**	**Quantidade Geral (Todos os volumes)**		**Quantidade Específico (Volume V)**	
		valor absoluto	valor relativo	valor absoluto	valor relativo
1	*pé, ca, un, os*	149	0.58%	16	0.19%
2	*-doiro, -eco, -meles*	83	0.32%	0	0%
3	*v. + fatigado, v. + a rodar*	41	0.16%	23	0.27%
4	*ameixa, órgano, vagante*	5.179	20.47%	906	10.93%
5	*como, que, para, por*	7	0.02%	0	0%
Total de termos para análise		19.832	78.41%	7.343	88.59%
Total Geral		25.291	100%	8.288	100%

Fonte: Elaborado pelos autores.

2.2.1 Validação dos dados pré-processados

O processo de validação dos dados foi composto por discussões pontuais e por reuniões virtuais entre os especialistas, dentre os quais alguns são nativos da Galícia ou descendentes e outros são linguistas e pesquisadores da Computação. Essa característica interdisciplinar dos especialistas favoreceu a validação dos dados de forma coerente.

As discussões pontuais foram direcionadas à validação das regras de remoção e ajustes dos termos. O objetivo foi avaliar os impactos dos tratamentos realizados nos dados para determinar a mínima perda de informações nesses termos.

Nas reuniões virtuais foram avaliados os termos que tiveram uma maior incidência nos *tweets*. O objetivo foi validar quais cartas, dentre as sete pré-selecionadas, seriam as mais relevantes para uma análise detalhada dos aspectos lexicais e semânticos.

As discussões validaram as regras aplicadas ao conjunto de dados e definiram a análise dos termos referente ao Volume V do ALGa, o qual aborda, como dito anteriormente, os aspectos referentes "ao ser humano", tais como: partes do corpo, sentimentos, dentre outros.

Os dez termos mais frequentes nos *tweets* e existentes no ALGa são descritos no Quadro 1.

Quadro 1: Quantidade de termos em valores absolutos registrados no ALGa que ocorrem nos *tweets*

Termo	**quantidade**
'dor'	590
'ren'	499
'ver'	499
'man'	409
'ene'	367
'ante'	358
'can'	340
'gran'	301
'sen'	300
'nas'	275

Fonte: Elaborado pelos autores.

Após análise dos termos encontrados com maior frequência nos *tweets*, alguns deles foram descartados por serem anglicismos (salvo casos em que há homonímia, exemplo 'man' e 'door') e/ou terem pouca representatividade para avançarem para as próximas etapas.

Após a rodada de filtragem, somente a *Carta 111 - Dor* foi mantida para análise dos dados. Em seguida, novos critérios foram definidos para determinar outro subconjunto de cartas: diversidade das cartas e a prévia análise das variantes lexicais nos *tweets*.

Nessa etapa foi delimitado o escopo da investigação dos termos registrados no atlas nos *tweets* e foi disponibilizado o arcabouço essencial para análise dos dados, que é apresentado na seção 2.3.

2.3 Análise dos dados

O tratamento dos dados (listagem dos dados originados do ALGa - volume V -, confronto com dados dos *tweets* e aplicação das regras de filtragem) resultou em um conjunto de 7.343 termos para análise. Esses termos foram aplicados em um algoritmo computacional que quantifica o número de aparições dos termos nos *tweets* e armazena tais *tweets*.

Em um segundo momento, realizou-se uma reunião virtual com um galego falante para validar as escolhas das cartas e analisar algumas polissemias que pudessem ocorrer devido à natureza do *Twitter*, principalmente pela sua informalidade. Neste caso, duas análises prévias foram essenciais para a definição das cartas: a distribuição dos termos da carta e a natureza do *Twitter*. Assim, as cartas *2- Cabeza* (Cabeça), *83 - Esgarro* (Cuspe), *85 - Guapo* (Bonito), *104 - Trenza* (Trança), *16 - Nariz* (Nariz), *31 - Pescozo* (Pescoço) e *111 - Dor* foram selecionadas para serem analisadas conforme os resultados na seção 3.

3 RESULTADOS

Com o objetivo de avaliar a vitalidade dos termos registrados no ALGa e presentes no *Twitter*, dois grupos foram definidos: resultados quantitativos e resultados diatópicos. Os resultados quantitativos objetivam determinar a ocorrência dos termos nos *tweets* que tenham o mesmo valor semântico presente

na carta analisada. Os resultados diatópicos visam a determinar as localidades em que os termos foram "tweetados" e validar as equivalências com as localidades do ALGa.

3.1 Resultados Quantitativos

A Carta 2 - Cabeza no ALGa é composta por seis variantes lexicais (*testa; moucha; cocena; cachola; cacharula; cabeza*). Dentre os termos originados dessa carta, apenas o termo *cabeza* com sete ocorrências foi encontrado nos *tweets*. Em todas as localidades descritas nesta carta no ALGa (com seis variantes) não houve no *Twitter* uma variação deste termo, ou seja, todos os informantes mencionaram 'cabeza' como resposta. No *Twitter* não houve nenhuma variação do termo, então a vitalidade deste termo é mantida, visto que o uso em redes sociais revela o mesmo sentido presente no ALGa.

A *Carta 83 - Esgarro* é composta por 25 termos (*polo; pollo; pito; garro; garneallo; gargaxo; galapa; esputo; esgoira; esgarro; esgarrios; esgarrio; esgarrexar; esgarreo; esgarrear; esgarrar; esgarrafallar; escuto; escupitagho; escupir; escarro; egarro; cuspe; bostezar; asgarro*) que correspondem às variantes lexicais registradas no ALGa para o termo 'cuspe'. Dentre elas, duas variantes lexicais foram encontradas nos *tweets*. Os termos com maior incidência nos *tweets* foram: (i) 'polo' com 72 ocorrências; e (ii) 'pollo' com apenas uma ocorrência. Em uma análise referente à semântica dos termos, observou-se que dos 73 termos encontrados, dois *tweets* se referiram ao termo 'polo' com significado do nome de um usuário do *Twitter*, como no exemplo que segue, veja-se (01):

(01) **Tweet 42**
@poloxxxxxxx - Buen día polo !!!...

Todas as 70 ocorrências se referem à preposição 'polo', tais como se descreve no *tweet* exemplificado em (02):

(02) Tweet 2

RT @XXXXXXX: "Galiza é un país rico, empobrecido polo tratamento do Estado" @XXXXXx @XXXXXXXX

O único termo 'pollo' encontrado nos *tweets*, se refere às 'patas de pollo', no sentido de patas ou pés do animal conhecido como galinha, veja-se (03).

(03) Tweet 1

"Esta lavadera de manos me tiene es mamada eche, las tengo ya como patas de pollo" - Mi mamá la que se lava las man…

De acordo com esses dados, nenhuma ocorrência de 'esgarro' com o sentido de 'cuspe' foi encontrada nos *tweets*. Assim, essa carta não foi selecionada para uma análise diatópica.

A *Carta 104 - Trenza* é composta por 54 termos distintos com o sentido de penteado (*trinca; trenzas; trenza; tranza; tirabuzón; roquete; rolete; rodete; restras; restra; rastra; priquete; piriquito; piriquetes; piriquete; perriquito; periquitos; periquito; periquite; periquetes; periquete; periqueta; pericos; perico; pericacho; pelo tecido; pelo enrastrado; pelo encaracolado; pelo encadillado; paneira; moño; mollo; mazo; gadello; enrestrar; enrastrar; cordas; corda; coletitas; coleta; cola de cabalo; cola de caballo; cola; cinta; chichas; chicha; cebolla; cebola; carrapito; carrapicho; carapito; caracol, cadrelo; cabelo*). Dentre eles, foram encontradas três ocorrências distintas nos *tweets*. Os termos identificados são 'cola' (três ocor.), 'cabelo' (duas ocor.) e 'cinta' (três ocor.).

O termo 'cabelo' ao qual o *tweet* se refere corresponde a um doce e não à trança, conforme *tweet* que segue exemplificado em (04):

(04) Tweet 1

RT @xxxxx: "lambiscar unha ducia de doces das monxas de Santa Clara, catro tartas de Mondoñedo recheas de améndoas e cabelo de anxo e…

Nos contextos de uso dos termos 'cola' e 'cinta' presentes nos *tweets* coletados, observou-se que tais vocábulos não possuem o significado de 'trenza'

conforme a carta 104. Assim, essa carta não foi selecionada para ser analisada do ponto de vista diatópico.

A *Carta 16 - Nariz* é composta por 29 termos (*ventas; trompeta; trompas; trompa; punteiro; peta; percha; navalla; nas; nariz de trompeta; nariz de loro; nariz de cañada; nariz de burro; nariz; narices; napias; napia; nacho; lacena dos bormos; fucius; fuciños; fuciñoco; fuciño de cambota; fuciño; fociño; focín; focico; cañada; cachiporra*). Dentre eles, foram encontradas cinco ocorrências nos *tweets*, a saber: 'nas' (37), 'nacho' (1), 'fuciño' (duas), 'trompa' (uma) e 'ventas' (quatro) com frequência nos *tweets*.

O termo 'trompa' foi o único que teve o seu significado registrado como 'nariz' de acordo com o ALGa. Os demais termos não apresentam o significado de 'nariz' quando aparecem nos *tweets*. O termo 'ventas', nos *tweets*, aparece com o sentido de 'vendas' conforme *tweet* trazido em (05):

(05) **Tweet 4**
BUSCO TRABAJO EN VENTAS - MONTERREY N.L.

Todas as 37 ocorrências do termo 'nas' nos *tweets* selecionados se referem à preposição, conforme o *tweet* que segue (06):

(06) **Tweet 3**
RT @XXXXXX: A CIG realiza protestas simbólicas nas sete cidades para esixir ingresos e salarios dignos, protección social e seguridade no…

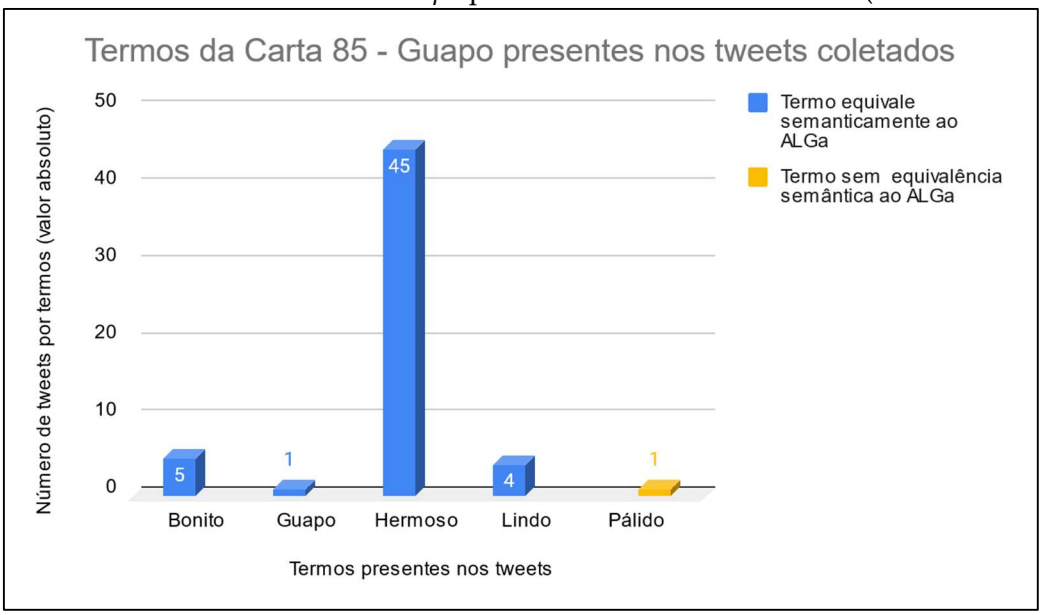
Diante do exposto, a Carta 16 - *Nariz* também não foi considerada elegível para uma análise diatópica em que figura a comparação ALGa e *Twitter*.

A *Carta 31 - Pescozo* é composta por 22 termos utilizados para designar tal parte do corpo humano (*poscozo; pezcozo; pescuezo; pescozo; pescocio; papo; man; ir coa man ao colo; ir; gorxapo; gorxa; goleira; garganta; gañote; cuello; coveira; cospozo; colo; coliga; cogote; cocote; cachazo*). Dentre eles, foram registrados nos *tweets* os termos 'colo' com uma ocorrência e 'man' com 14 ocorrências. Nenhum dos termos encontrados nos *tweets* corresponde a 'pescozo' conforme descrito no

ALGa. As principais aparições de 'man' nos *tweets* se referem às 'mãos' e a 'man', um anglicismo utilizado para designar 'homem'. Assim, essa carta do ALGa não foi eleita para ser utilizada para a representação dos *tweets* por localidade.

A *Carta 85 - Guapo* é composta por 22 termos distintos que possuem o significado em português de 'bonito' (*xeitoso; xeitosa; tá ben; que está ben; pálido; magho; lindo; hermoso; guapo; guapa; gopo; gasalloso; garboso; formoso; denoso; caravel; bonito; bon mozo; bo mozo; ben feito; ben feita; aquelado*). Dentre eles, foram encontrados cinco nos *tweets*, os quais a seguir são apresentados: 'bonito' (5), 'guapo' (1), 'hermoso' (45), 'lindo' (4) e 'pálido' (1), conforme Gráfico 1.

Gráfico 1: Termos da Carta 85 - *Guapo* presentes nos *tweets* coletados (valor absoluto)



Fonte: Elaborado pelos autores.

O termo 'pálido' no contexto do *tweet* apresentado não pode ser considerado com o mesmo valor semântico do encontrado na carta 'Guapo'. Conforme *tweet* selecionado como exemplo e exposto em (07), percebe-se que "um punto azul pálido" pode ser interpretado como "um ponto azul claro".

(07) **Tweet 1**
Un punto azul pálido

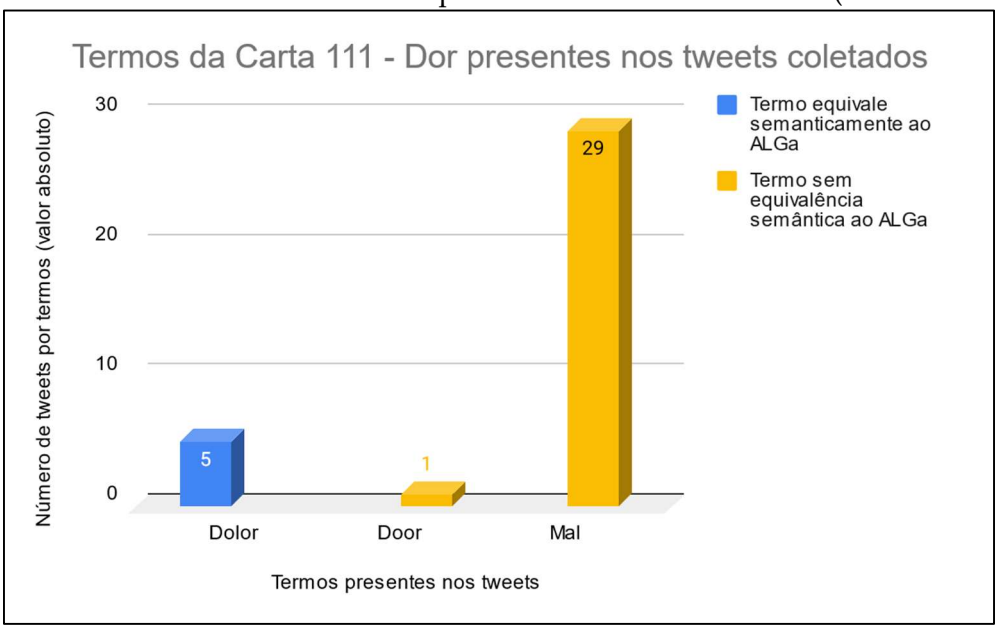
O termo ‘bonito’, em todas as suas ocorrências nos *tweets*, corresponde ao valor semântico que aparece na carta ‘guapo’ do ALGa. As quatro ocorrências do termo ‘lindo’, assim como as 45 do termo ‘hermoso’ empregadas nos *tweets* também correspondem ao sentido expresso na carta ‘guapo’ do ALGa. E o próprio termo ‘guapo’ corresponde ao valor semântico da carta ‘guapo’, conforme o *tweet* que segue (08):

(08) **Tweet 1**
Que guapo ereees😍

Devido à variedade dos termos encontrados nos *tweets* com os mesmos valores semânticos registrado no ALGa, essa carta será analisada do ponto de vista diatópico com o intuito de verificar a localidade de origem dos falantes que “tweetaram” com esses termos e analisar a correspondência com o ALGa.

A *Carta 111 - Dor* revela a ocorrência de quatro termos (*mal; dor; door; dolor*) para representar o sentimento expresso em língua portuguesa pelo vocábulo ‘dor”. Três deles estão presentes nos *tweets* selecionados, a saber: ‘dolor’ (5), ‘door’ (1) e ‘mal’ (29), conforme exposto no Gráfico 2.

Gráfico 2: Termos da Carta 111 - *Dor* presentes nos *tweets* coletados (valor absoluto)



Fonte: Elaborado pelos autores.

A análise empreendida para as cinco ocorrências do termo 'dolor' nos *tweets* revela que tais usos correspondem aos valores semânticos encontrados no ALGa, conforme *tweet* selecionado para exemplificação (09) que segue:

(09) **Tweet 2**
"Una escucha honesta es la mejor medicina que podemos ofrecer al que pasa dolor" - XXX

O termo 'door' correspondeu a um anglicismo no valor de 'porta' em português, segundo o *tweet* exposto em (10) e não corresponde ao valor semântico presente no ALGa.

(10) **Tweet 1**
Mike Shinoda - Open Door (Demo XXXX)

Os 29 termos 'mal' encontrados nos *tweets* não correspondem aos sentidos da carta do ALGa, conforme pode-se constatar em (11) que segue.

(11) **Tweet 1**
¿Acaso eres perfecta pa' juzgar? Por más que lo hago bien tu lo ves mal -

Das sete cartas pré-selecionadas, apenas duas mostraram-se relevantes para serem submetidas a uma análise do ponto de vista diatópico. A Carta 111 - *Dor* e Carta 85 - *Guapo* foram submetidas a uma análise diatópica para verificar em qual ocorrência esses termos apareciam. Essa análise está descrita na seção seguinte.

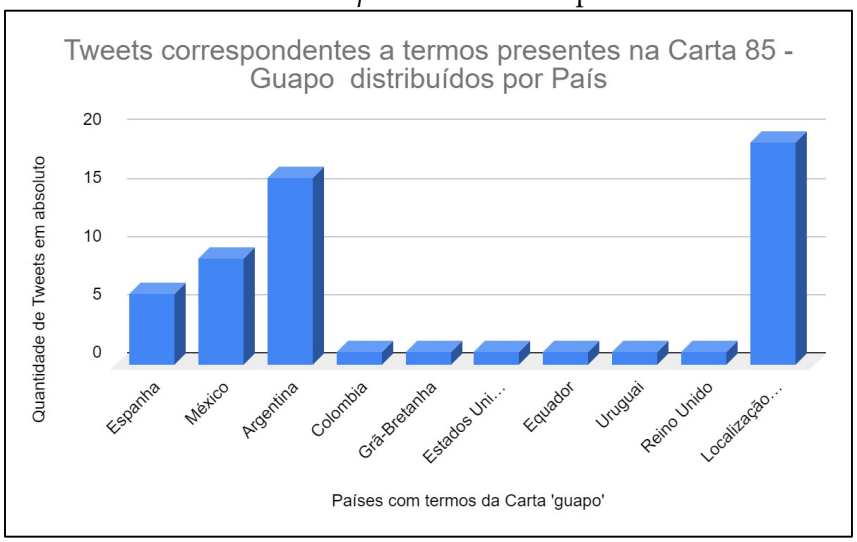
3.2 Resultados Diatópicos

A análise diatópica visa a verificar a ocorrência dos termos presentes nos *tweets* selecionados, cujos significados correspondem ao mesmo apresentado no ALGa e representados por localidade. Neste sentido, almeja-se analisar se os termos presentes nas localidades têm correspondência com as localidades descritas nas cartas publicadas no ALGa. Cada variante de cada carta foi analisada separadamente.

3.2.1 *Carta 85 - Guapo*

Iniciada a análise, observou-se que embora uma validação da língua Galega tenha sido realizada em etapas anteriores, diversas ocorrências dos termos referente à carta *Guapo* tiveram suas ocorrências originadas de localidades fora da Espanha, como se pode observar no Gráfico 3.

Gráfico 3: *Tweets* correspondentes a termos presentes na Carta 85 - *Guapo* - distribuídos por País



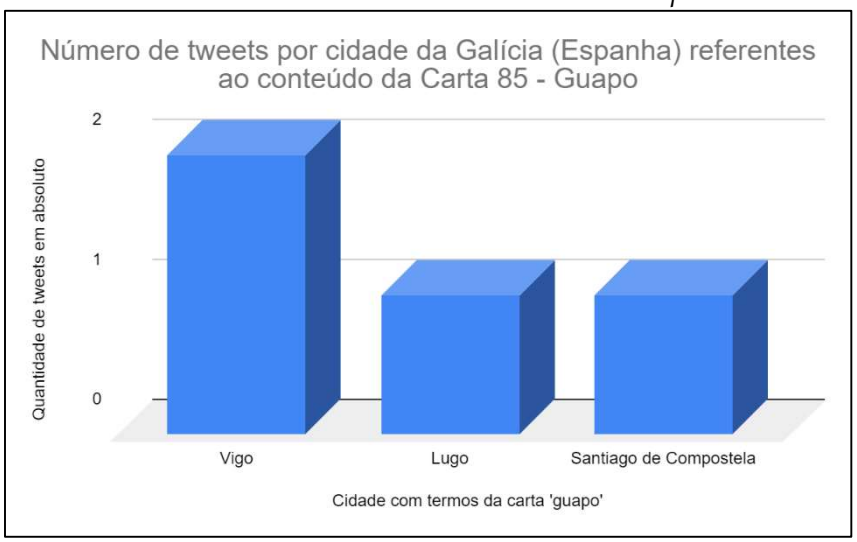
Fonte: Elaborado pelos autores.

Observa-se uma grande ocorrência de países da América do Sul e da América Central, mas constatam-se a presença de países da Europa. Esse fato pode ser justificado pela grande emigração que ocorreu no início do século XX para essas localidades por galegos e que podem continuar usando a sua língua materna como meio de comunicação nas redes sociais. É importante observar a grande quantidade de *tweets* (19) que possuem termos da Carta 85 - *Guapo* com os seus respectivos valores semânticos, mas não descrevem a localidade do usuário do *Twitter*. No *Twitter* não é obrigatório que o usuário informe o seu local de origem. Além disso, ele pode ser originalmente da Galícia, mas estar vivendo

na Argentina, então ele pode informar no local de origem o nome do seu atual país: a Argentina.

Considerando o extrato somente da Espanha, observa-se que três localidades da Galícia (Vigo, Lugo e Santiago de Compostela) tiveram termos "tweetados", conforme revela o Gráfico 4.

Gráfico 4: Número de *tweets* por cidade da Galícia (Espanha) referente ao conteúdo da Carta 85 - *Guapo*



Fonte: Elaborado pelos autores.

É importante observar que um *tweet* cuja localização foi Azagra em Navarra continha um termo da Carta 'guapo', porém por se tratar de uma localidade fora da Galícia, o dado foi retirado do resultado.

Ainda em relação à carta 85 - *Guapo*, observou-se que, embora o termo '*hermoso*' tenha sido o de maior ocorrência nos *tweets*, esse termo não se encontra presente no Dicionário da Real Academia Galega[10]. Porém, o mesmo termo está presente no ALGa, como uma das variantes lexicais do termo 'guapo'. Dentre as possibilidades mencionadas de uso do termo por um galego falante é a de que esse termo seja oriundo de um castelhanismo.

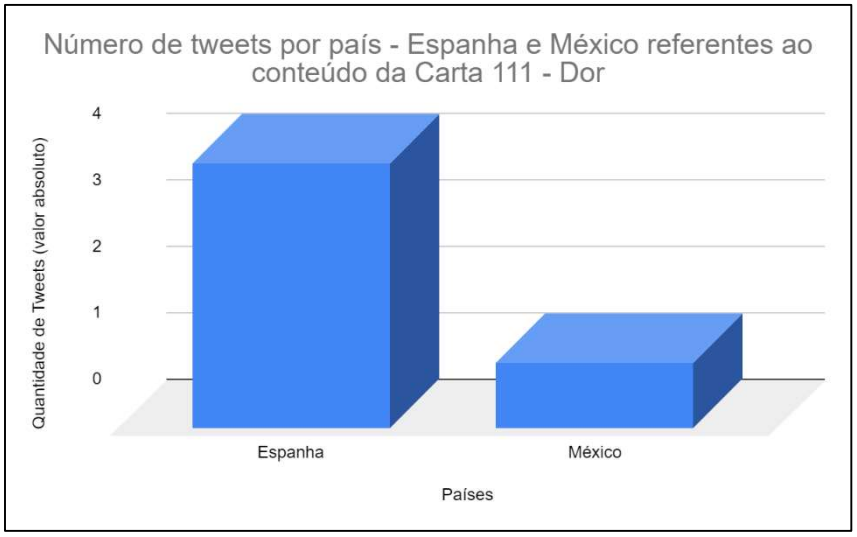
[10] https://academia.gal/dicionario

3.2.2 *Carta 111 - Dor*

As localidades referentes à carta 'dor' somente correspondem às ocorrências dos termos 'dolor' nos *tweets*, visto que os demais termos não tiveram uma correspondência semântica aos termos registrados no ALGa.

Observa-se também que alguns termos foram encontrados em outras localidades que não somente a Espanha, como se visualiza por meio do Gráfico 5.

Gráfico 5: Número de *tweets* por país - Espanha e México
referente ao conteúdo da Carta 111 – *Dor*

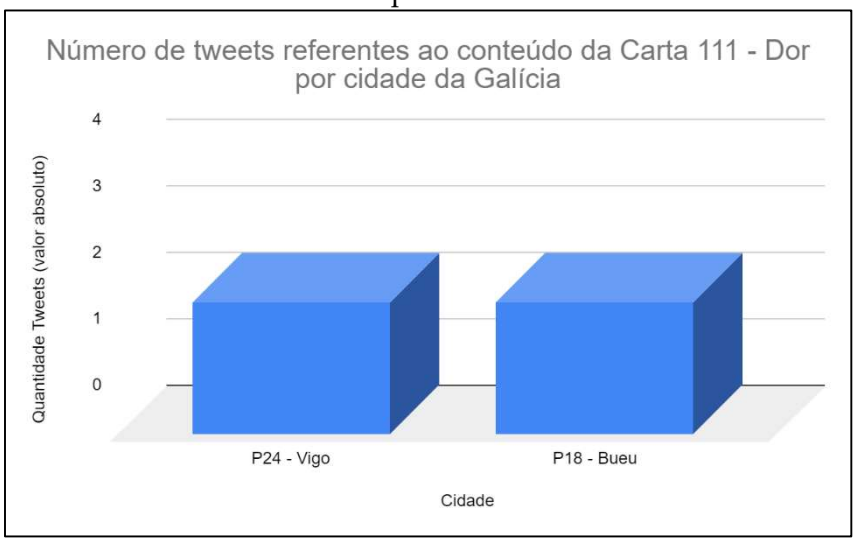


Fonte: Elaborado pelos autores.

Encontrou-se um termo referente à carta 'dor' com localidade de registro do falante que "tweetou" a partir do México. Esse texto também pode ser fruto de emissão realizada por um galego emigrado e que faz uso da língua Galega nas redes sociais.

Expressivamente na Espanha, temos a ocorrência de termos nas localidades Vigo e Bueu, como se observa no Gráfico 6.

Gráfico 6: Número de *tweets* referente ao conteúdo da Carta 111 - *Dor* por cidade da Galícia



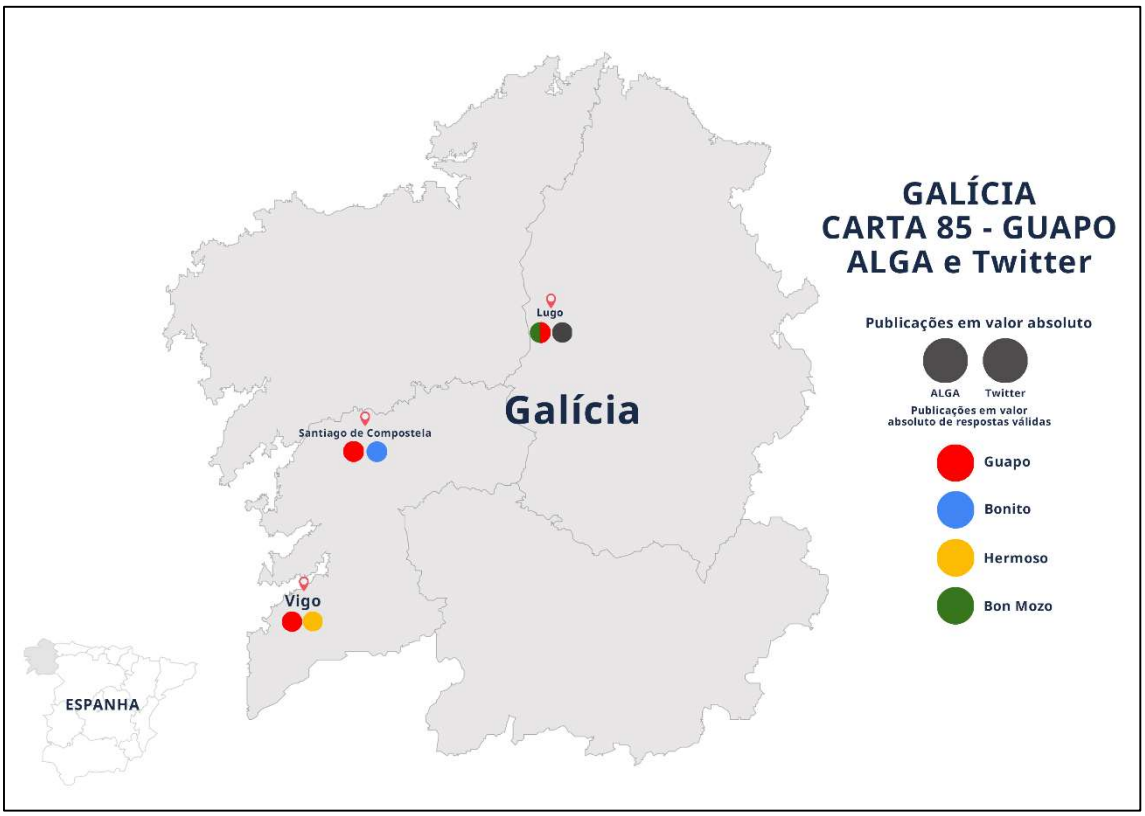
Fonte: Elaborado pelos autores.

3.2.3 *Comparativo por localidade em relação ao ALGa e ao Twitter*

Considerando os resultados das duas cartas em relação às localidades na Galícia, observa-se que na Figura 1 – Cartograma representativo da análise comparativa *Twitter* x ALGa (Carta 85 - *Guapo*) – os termos encontrados no *Twitter* diferem dos termos registrados no ALGa. Enquanto no ALGa, nas cidades de Vigo e Santiago de Compostela o termo empregado foi 'guapo', no *Twitter*, os usuários desta localidade mencionaram os termos 'hermoso' e 'bonito' respectivamente.

Em Lugo, por sua vez, o termo 'pálido' apareceu no *tweet*, mas no ALGa os termos 'guapo' e 'bon mozo' são os que correspondem a essa localidade. Neste caso, não foi observada a vitalidade dos termos nessas ocorrências para a amostra considerada nessa análise.

Figura 1: Cartograma representativo da análise comparativa
Twitter x ALGa - Carta 85 - *Guapo*

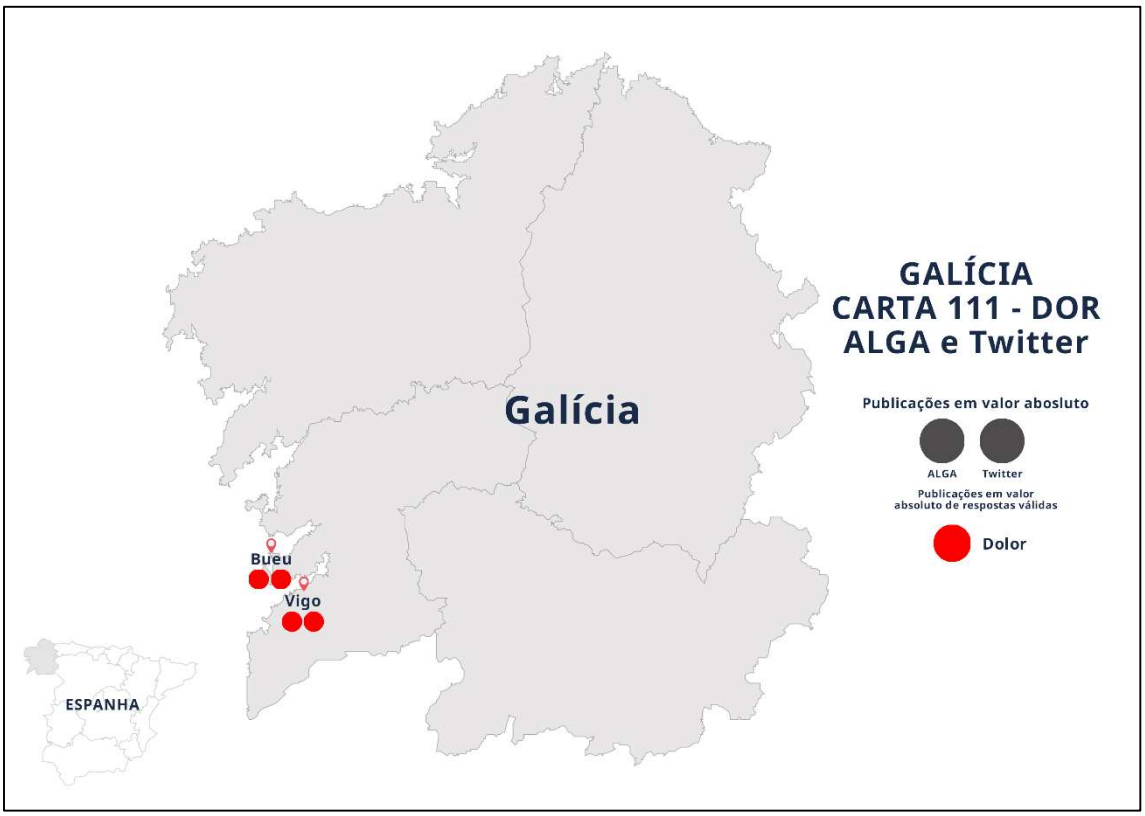


Fonte: Elaborada pelos autores.

Em se tratando da *Carta 111 - Dor*, somente o termo 'dolor' foi contabilizado e consequentemente suas localizações na Galícia podem ser indicadas.

As duas cidades galegas que "tweetaram" com o termo 'dolor' foram Bueu e Vigo, as quais mantiveram a vitalidade deste termo como sendo o valor para se referir a uma dor. Neste caso, demonstra-se a vitalidade deste termo por meio do exposto na Figura 2 - Cartograma representativo da análise comparativa *Twitter* x ALGa - Carta 111 – *Dor*.

Figura 2: Cartograma representativo da análise comparativa
Twitter x ALGa - Carta 111 - *Dor*



Fonte: Elaborada pelos autores.

4 DISCUSSÕES

Algumas discussões sobre a metodologia proposta se fazem necessárias em relação aos achados deste trabalho.

Primeiramente em relação à proposta metodológica idealizada e aplicada na etapa de Aquisição e Validação dos dados. A estratégia adotada foi filtrar via API do *Twitter* os *tweets* que fossem da língua espanhola, visto que o galego não é utilizado como língua no *Twitter*. Em seguida foi utilizado um modelo de representação de palavras, denominado *fastText* na sua versão multilingual para contemplar a língua galega. Assim foi possível validar se os *tweets* foram escritos em galego ou castelhano. Quando ocorreu a análise diatópica, observou-se que embora os *tweets* estivessem escritos em galego, a localização dos falantes que "tweetaram" estava fora da Galícia. Para esse estudo, esses *tweets* foram

descartados. Porém, reavaliando a metodologia, pretende-se redimensionar a busca dos *tweets* por meio da localização 'Espanha' e em seguida fazer análise da língua se o texto coletado no *Twitter* é ou não em galego. Assim pretende-se obter todas as ocorrências da Espanha e não mais de outros países.

Para aquisição dos dados do ALGa, como exposto, partiu-se da planilha geral de ocorrências e acrescentaram-se as filtragens. Objetivava-se na fase de "coleta por planilha" como um ponto de partida que um número x de cartas se mostrasse relevante para continuidade da análise. Como também visto no item 2.2.1, tal caminho apenas indicou como relevante o estudo da Carta 111- *Dor* e da Carta 85 - *Guapo*. O resultado conduziu a pesquisa para um novo ponto de partida, ou seja, observar o atlas e selecionar cartas específicas. Diante do exposto, numa nova fase de trabalho, a metodologia deverá ser ajustada e o passo 1 será ou poderá ser o de elencar um número predefinido de cartas, elencar os termos cartografados para a posterior coleta nos *tweets*.

Ainda em relação às ocorrências dos dados fora da Galícia, aventou-se a hipótese de que o falante é usuário da língua galega em seus textos, mesmo morando fora da Espanha e da região geográfica da Galícia. Como língua de cultura, percebe-se que talvez essa seja uma opção do falante: escrever em galego, estando ele localizado na Galícia ou não. Tal hipótese justifica a ocorrência de *tweets* em galego documentados em outros países (base de registro a localidade apurada no *Twitter*).

Em relação à localização no *Twitter*, o usuário pode informar a localização que desejar. Neste caso, um galego pode estar vivendo em outros países fora da Galícia e usa a sua língua materna como principal meio de comunicação nas redes sociais. Além disso, muitos *tweets* não possuem localização, pois a localização do usuário é opcional no cadastro do *Twitter* e muitas vezes não é informada pelo usuário no *tweet*.

Outro fator importante a ser analisado é referente às poucas ocorrências de termos que aparecem no ALGa e que também aparecem nos *tweets*. O índice desses termos está muito baixo mesmo estando com 10 mil *tweets* somente em galego. A ocorrência de *tweets* originados de outras localidades fora da Galícia também diminuiu significativamente o número dos termos, visto que muitos foram descartados por não terem sido "tweetados" em localidades dentro da Galícia.

E, por fim, a grande quantidade de termos polissêmicos cujos valores semânticos são distintos entre ALGa e *Twitter* foram também descartados. Esses termos não foram utilizados neste trabalho.

CONSIDERAÇÕES FINAIS

A vitalidade dos termos presentes nos Atlas Linguísticos é uma atividade relevante visto que permite analisar se os termos continuam sendo empregados e em quais localidades. A ocorrência dos termos registrados no ALGa foi analisada no *Twitter*, permitindo uma observação da vitalidade dos termos mesmo no curso da língua escrita. Além disso, as redes sociais, e em especial o *Twitter* têm características específicas que diminuem as possibilidades de análise tão fidedigna como nos atlas, mas permitem que se obtenham indícios de empregabilidade dos termos, instigando novas pesquisas nas áreas estudadas.

Neste trabalho, a vitalidade no *Twitter* de termos registrados no Atlas Linguístico do Galego (ALGa) foi avaliada por meio de uma metodologia desenvolvida para tal análise semiautomatizada. Constatou-se que, para a continuidade do estudo, diversos fatores devem ser observados ou reconsiderados nas etapas de filtragem de dados, tais como: (i) a localidade indicada/registrada no *Twitter*; (ii) o idioma identificado como o Galego, que foi analisado por uma rede neural; e (iii) as divergências entre os sentidos dos termos

presentes no ALGa e no *Twitter*. Esses são alguns dos fatores que precisam ser analisados em relação às redes sociais.

Embora esses fatores dificultem a obtenção dos termos, essa análise permite apontar os primeiros indícios de como pode ser conduzida uma metodologia para esse tipo de pesquisa em relação aos atlas linguísticos publicados no mundo e neste caso aplicado ao Atlas Linguístico da Galícia.

REFERÊNCIAS

ATLAS LINGUARUM EUROPAE (ALE). Assen-Maastricht: Van Gorcum, 1983-1990. v. 1-4. Roma: Istituto Poligrafico e Zecca dello Stato, 1998. v. 5.

ATLAS LINGUISTIQUE ROMAN (ALiR). v. 1. Roma: Instituto Poligrafico e Zecca dello Stato. Libreria dello Stato, 1996.

CARDOSO, Suzana et al. *Atlas Linguístico do Brasil*. Londrina: Eduel, 2014a. v. 1.

CARDOSO, Suzana et al. *Atlas Linguístico do Brasil*. Londrina: Eduel, 2014b. v. 2.

DE PAIVA, Valeria; RADEMAKER, Alexandre; DE MELO, Gerard. OpenWordNet-PT: An Open {B}razilian {W}ordnet for Reasoning. In.: *Proceedings of COLING 2012: Demonstration Papers.* The COLING 2012 Organizing Committee, Mumbai, 2012, p. 353-360. Disponível em: [http://www.aclweb.org/anthology/C12-3044]. Acesso em: mar. 2021.

DRUDE, Sebastian et al. *Language vitality and endangerment*. 2003. Disponível em: [https://unesdoc.unesco.org/ark:/48223/pf0000183699]. Acesso em: 20 abr. 2019.

FLÓREZ, Luís et al. *Atlas Lingüístico-Etnográfico de Colombia.* Bogotá: Instituto Caro y Cuervo, 1981-1983. 6 v.

GARCÍA, Constantino; SANTAMARINA, Antón; ÁLVAREZ BLANCO, Rosario; FERNÁNDEZ REI, Francisco; GONZÁLEZ GONZÁLEZ, Manuel. *Atlas Linguístico Galego*. Volume 5: Léxico. O Ser Humano (I). Fundación Pedro Barrié de la Maza. Instituto de Língua Galega, 2005.

GILLIÉRON, Jules; EDMONT, Edmond. *Atlas Linguistique de la France*. 35 fasc. Paris: Honoré Champion, 1902-1910.

JABERG, Karl; JUD, Jakob. *Sprach- und Sachatlas Italiens und der Südschweiz*, v. 1-8.I. Zofingen: Rieger, 1928-1940.

JOULIN, Armand; RAVE, Edouard; BOJANOWSKI, Piotr; MIKOLOV, Tomas. *Bag of Tricks for Efficient Text Classification*. arXiv preprint arXiv:1607.01759, 2016.

LOPE-BLANCH, Juan M. *Atlas Lingüístico de México.* 3 v. México: El Colegio de México; Fondo de Cultura Económica, 1990.

MONTGOMERY, Chris; STOECKLE, Philipp. Geographic information systems and perceptual dialectology: a method for processing draw-a-map data. *Journal of Linguistic Geography*, v. 1, n.1, p. 52-85, 2013.

NUNES, Arley Prates; JESUS, Luis Emanuel N.; CLARO, Daniela Barreiro; MOTA, Jacyra; RIBEIRO, Silvana; PAIM, Marcela; OLIVEIRA, Josane. *Vitality Analysis of the Linguistic Atlas of Brazil on Twitter*, Computational Processing of the Portuguese Language - 14th International Conference, LNCS 12037, Évora, Portugal, 2020, p. 184-194.

PAVALANATHAN, Umashanthi; EISENSTEIN, Jacob. Confounds and Consequences in Geotagged Twitter Data. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, p. 2138-2148, 2015.

PERES, Edenize Ponzo. Análise da vitalidade do Vêneto em uma comunidade de imigrantes do Espírito Santo. *Revista (Con) textos Linguísticos*, v. 5, n. 5, 2011.

QUINN, Sterling. A geolinguistic approach for comprehending local influence in Open Street Map. *Cartographica: The International Journal for Geographic and Geovisualization,* v. 51, n. 2, p. 67-83, 2016.

Nota do editor:
Artigo submetido para avaliação em: 05 de abril de 2021.
Aprovado em sistema duplo cego em: 21 de julho de 2021.