DOI: https://doi.org/10.35520/diadorim.2021.v23n3aa54729
Received on: February 08, 2022 / Accepted on: August 19, 2022

ANÁLISE DE DADOS LINGUÍSTICOS
Entrevista com *Adriana Picoral*

LINGUISTIC DATA ANALYSIS
Interview with *Adriana Picoral*

Adriana Picoral[1]
Marcia dos Santos Machado Vieira[2],
Vanessa Meireles
Ravena Beatriz de Sousa Teixeira
Mariana Gonçalves da Costa

RESUMO

Adriana Picoral é Professora Assistente de prática no Departamento de Ciência da Computação e membro do corpo docente afiliado no Programa de Pós-Graduação Interdisciplinar de Aquisição e Ensino de Segunda Língua da Universidade do Arizona. Ela é a fundadora da R-Ladies Tucson, que faz parte de uma organização mundial para promover a diversidade de gênero na comunidade R. É doutora em Linguística Aplicada e Bacharel em Ciência da Computação (UFRGS). Sua pesquisa baseia-se em Linguística Computacional e Linguística de Corpus, no intuito de esclarecer o uso, a aquisição e o desenvolvimento multilinguísticos. A presente entrevista focaliza temas relativos à ciência da análise de dados linguísticos, ciência da informação e ciência computacional.

PALAVRAS-CHAVE: Sociolinguística; Linguística de Corpus; Linguística Computacional; Processamento e análise de dados linguísticos; Português e Inglês.

1 A pesquisadora entrevistada, Professora Doutora *Adriana Picoral* (adrianaps@arizona.edu), atua na Universidade do Arizona, na área de coleta, processamento, manutenção e análise de dados linguísticos em ambientes acadêmicos e além.

2 *Marcia dos Santos Machado Vieira* (marcia@letras.ufrj.br, docente da UFRJ, Pesquisadora do CNPq e Cientista do Nosso Estado/Faperj), *Vanessa Meireles* (vanessa.meireles@univ-montp3.fr, docente da UPVM), R*avena Beatriz de Sousa Teixeira* (ravena_beatriz@letras.ufrj.br , ravenabst@gmail.com, discente da UFRJ) e *Mariana Gonçalves da Costa* (marianag.costta@gmail.com, discente da UFRJ) – as entrevistadoras – são pesquisadoras vinculadas a projetos de pesquisa que têm articulação interinstitucional: Programa de Pós-Graduação *stricto sensu* em Letras Vernáculas (Língua Portuguesa) da Universidade Federal do Rio de Janeiro (UFRJ) e RESO (Recherches sur les suds e les orients) da Universidade Paul-Valéry/Montpellier 3. As três primeiras são membros do Projeto VariaR (Variação em línguas românicas). E Marcia Machado Vieira, Ravena Teixeira e Mariana Costa são pesquisadoras do Projeto PREDICAR (Formação e expressão de predicados complexos e predicações: estabilidade, variação e mudança construcional).

ABSTRACT

Adriana Picoral is an Assistant Professor of practice in the Computer Science Department, and an affiliated faculty member in the interdisciplinary Graduate Program of Second Language Acquisition and Teaching at the University of Arizona. She is the founder of the R-Ladies Tucson, which is part of a world-wide organization to promote gender diversity in the R community. She holds a PhD in applied linguistics and a bachelor's degree in computer science (UFRGS). Her research draws from Corpus Linguistics and Computational Linguistics to shed light on multilingual language use, acquisition, and development. This interview focuses on topics related to the science of linguistic data analysis, information science and computational science.

KEYWORDS: Sociolinguistics; Corpus Linguistics; Computational Linguistics; Processing and analysing language data; Portuguese and English.

Questão 1

Iniciamos agradecendo à Profa. Dra. Adriana Picoral por aceitar nosso convite para esta entrevista e pedindo-lhe que nos conte um pouco sobre sua trajetória acadêmica e sobre como se iniciou seu interesse por análise linguística e por ciência da informação. O que destacaria no seu percurso acadêmico-científico? Quais são seus projetos atuais? Qual é sua área e sua atuação na Universidade do Arizona?

A minha trajetória acadêmica é interdisciplinar e variada, como minha atual área de atuação. Minha graduação foi em Ciência da Computação na UFRGS, e minha bolsa de iniciação científica na época foi com a Dr. Margarete Schlatter, Professora de Prática de Ensino de Línguas Adicionais e de Linguística Aplicada. No meu doutorado em linguística aplicada eu continuei esse processo de combinar minha formação em métodos computacionais com análise linguística no meu trabalho com as professoras Dras. Ana Carvalho, em Sociolinguística, e Shelley Staples, em Linguística de Corpus. Meus projetos atuais incluem o estudo de variação linguística do Português Brasileiro no Twitter e do Inglês acadêmico por aprendizes. Minha área de atuação na Universidade do Arizona é Ciência de Dados em termos amplos, e, em termos mais específicos, é análise quantitativa de dados de dados linguísticos.

Questão 2

Poderia nos apresentar um pouco das metas e dos desafios perspectivados pela equipe de pesquisa que coordena? Quais são os perfis dos subprojetos de pesquisa, das áreas de contribuições destes? Em que medida e de que maneira a ciência de dados, informações, textos se relaciona à ciência linguística?

O maior desafio atual relacionado à análise quantitativa de dados linguísticos é estabelecer quais métodos estatísticos são mais úteis para a generalização de resultados de análises de variação linguística em larga escala. Os subprojetos de pesquisa que coordeno incorporam métodos originários da Sociolinguística de Variação e da Linguística de Corpus, dando atenção maior a como minerar e analisar grandes quantidades de produção natural de linguagem. A

Sociolinguística de Variação historicamente faz uso de métodos estatísticos detalhados para a análise do que, hoje em dia, é considerado uma quantidade de dados menor e mais selecionados. Por outro lado, a Linguística de Corpus foca na coleta de muitos dados (em milhões e até bilhões de palavras), com métodos de análise estatísticas desses dados ainda não estabelecida pelos pesquisadores na área. Não há um conjunto de testes estatísticos usados na Linguística de Corpus, alguns pesquisadores reportam só porcentagens, ou frequências normalizadas, enquanto outros aplicam regressão para diferentes propósitos. Para dar um exemplo mais concreto, o estudo de variação do sujeito pronominal no Português Brasileiro é muito produtivo na área de Sociolinguística, com uma série de estudos pequenos (com dezenas de participantes e centenas de tokens) com dados coletados em diversas regiões do Brasil. Nosso grupo de pesquisa construiu um corpus de tweets em Português (produzidos por centenas de usuários, e milhares de tokens) que permitiu o estudo da variação do sujeito pronominal no Português Brasileiro em diversas regiões do Brasil no mesmo meio de comunicação, no mesmo período de tempo (Picoral et. al, 2021). Implementamos diversos métodos e ferramentas computacionais para minerar dados para análise estatística confiável e generalizável, a partir dos dados crus extraídos do Twitter. O próximo passo para nosso grupo é organizar e compartilhar essas ferramentas e métodos com outros pesquisadores com o propósito de expandir esse tipo de pesquisa em larga escala.

Questão 3

Desde 2019, você atua como embaixadora da iniciativa global *Women in Data Science*. Como se configura o projeto? O que a motivou a participar da proposta? Segundo sua perspectiva, como este contribui para o posicionamento da figura feminina nos espaços acadêmico-científicos?

Representação feminina em específico, e diversidade em termos mais gerais, são problemas em espaços relacionados à informática, tanto na esfera acadêmica quanto na indústria. Na minha graduação na UFRGS nós mulheres tínhamos menos de 10% de representatividade. Ser embaixadora da iniciativa global *Women in Data Science* faz parte do meu trabalho de inclusão de mulheres em um ambiente técnico que não recebe mulheres de braços abertos. O treinamento em métodos computacionais exige uma comunidade de prática que permite o processo de falhar e aprender com erros, e essa comunidade de prática não existe para mulheres em ambientes dominados por homens. Então se faz preciso criar ambientes dominados por mulheres, para dar espaço para mulheres, e outras pessoas excluídas dessas áreas técnicas, receberem esse treinamento que é tão essencial para avançar o conhecimento e a pesquisa em áreas que estão começando a utilizar mais e mais ferramentas computacionais, como a ciência linguística.

Questão 4

A partir de seu conhecimento empírico, poderia comentar brevemente sobre a relevância de saberes referentes ao tratamento, processamento e análise quantitativa de dados no contexto atual dos estudos linguísticos? Quais são os pontos que carecem de aprimoramento na área?

Acredito que o ponto mais importante que carece aprimoramento na área dos estudos linguísticos é reprodutibilidade, que possibilita a reprodução de todos os passos de tratamento, processamento e análise quantitativa de dados, a partir dos dados crus. Nosso forte nos estudos linguísticos é coletar dados, mas não temos métodos estabelecidos de o que fazer a partir desse dados, até mesmo pela ampla variedade de subcampos com diferentes perspectivas de análise linguística. Falta também a prática de ciência aberta nessa área no Brasil, seguindo inciativas como a da Open Science Framework (https://osf.io/) que possibilitam o compartilhamento aberto de dados e métodos. A estatística não é uma ciência que é aplicada descasada dos outros métodos e conhecimentos de uma área. Pelo contrário, é preciso entender bem quais as perguntas e perspectivas que são exploradas em estudos linguísticos para entender como a estatística pode responder essas perguntas em termos quantitativos. A partir do compartilhamento aberto dos métodos que estão sendo atualmente explorados e utilizados na linguística é que podemos construir um corpo de conhecimento sobre os métodos estatísticos mais eficazes, com reprodução desses métodos de maneira sistemática em diferentes estudos.

Questão 5

Em sua opinião, o tópico (tratamento de dados na análise linguística) é abordado de forma satisfatória no que concerne à formação de linguistas em território brasileiro? Considerando sua formação (inclusive, no Brasil), em que falta aos currículos dos cursos de graduação (Bacharelado e Licenciatura) na área de Letras e Linguística investir, para o enriquecimento das análises linguísticas que consubstanciam o que é descrito?

O tratamento de dados na análise linguística não é abordado de forma satisfatória no que concerne à formação de linguistas nem em território brasileiro (nem nos Estados Unidos). Como aluna de graduação na UFRGS há duas décadas, eu tive que lutar muito para ter uma educação interdisciplinar. Apesar de algumas colaborações existirem, a conversa entre áreas como Ciência da Computação e Letras e Linguística em universidades Brasileiras, por exemplo, é muito limitada. Sem se falar na inclusão de cursos de estatística que oferecem uma perspectiva aplicada de conceitos estatísticos em práticas de pesquisa que envolvem análises linguísticas. Ainda existe muita resistência discente e docente em atravessar essa fronteira fictícia entre ciências humanas e ciências exatas. Mas cultura não muda sem exemplos, e eu acredito que, a partir de mais grupos e projetos de pesquisa que incluem métodos quantitativos de análise

linguísticas, maior nossa comunidade de prática e mais pessoas qualificadas para ministrar cursos aplicados de estatística e métodos computacionais em cursos de graduação na área de Letras e Linguística no Brasil. Essa lacuna vai demorar a ser preenchida, por isso a necessidade de fomentar ainda mais e com mais emergência esse tipo de pesquisa.

Questão 6

Durante seu minicurso ministrado no I Congresso PREDICAR[3], houve uma discussão quanto à falta de fundamentação teórica na aplicação de métodos de análise estatística nas pesquisas linguísticas. Quais seriam as possíveis repercussões desse problema na leitura e adequação da análise? Quais propostas podem ser (ou estão sendo) feitas para solucionar essa questão na área da linguística?

As ferramentas computacionais que usamos hoje em dia para análise estatística e a quantidade de tutoriais e informação gratuita na internet facilitam muito a aplicação desses métodos. No entendo, não existe transparência em como esses métodos são aplicados na maior parte dos meios de publicação cientifica na área da linguística. Então, alunos, professores, e pesquisadores que estão na posição de começar a aprender e utilizar estatística não tem referencias de como interpretar os resultados estatísticos, só existe referência de como rodar certos testes na sintaxe em linguagens de programação como Python e R. Porém, usar um certo tipo de função em uma dessas linguagens de programação não é suficiente. A análise de dados de maneira quantitativa se faz confiável a partir do tratamento de dados até a interpretação de resultados. Rodar uma função para um teste estatístico é uma pequena parte do que é necessário para produzir resultados de uma análise linguística. A maior repercussão dessa falta de entendimento sobre métodos estatísticos é a carência de reprodutibilidade. Temos projetos com alta qualidade de coleta de dados, mas que falham na análise estatística, não produzindo os resultados necessários que adicionam à literatura da área. Acredito que grupos como PREDICAR e o programa LAEAL da PUCSP são os protagonistas atuais que procuram avançar a formação de alunos nesses métodos computacionais e estatísticos na área da linguística no Brasil. Ainda mais grupos, mais projetos de pesquisa, que se preocupem com a qualidade das interpretações estatísticas de resultados são necessários para acelerar essa questão.

Questão 7

Considerando o seu trabalho com dados de uso da rede social *Twitter*, quais desafios e potencialidades foram encontrados por você e pela sua equipe quanto aos dados oriundos dessa plataforma? Quais especificidades estariam envolvidas na coleta e no tratamento de dados virtuais?

3 Disponível em: https://www.youtube.com/watch?v=fe1DIiIlLxE&list=PLYlbzUR7D2so-Xo9yP0yOhbPuu2el2YwtF. Acesso em: 31 de maio, 2022.

As redes sociais têm relevância para a nossa comunicação diária, muita linguagem ocorre exclusivamente através desses meios. No entanto, pouca pesquisa e entendimento existe sobre esse tipo de linguagem, especialmente em diferentes variedades do Português e outras línguas menos pesquisadas. Justamente por esse carência de estudos é que as potencialidades são muitas. O maior desafio com dados coletados do Twitter é a falta de informação sobre a maioria dos usuários. Não temos acesso a metadados demográficos como faixa etária, etnia, e classe social. Claro, se a coleta se faz a partir de perfis de celebridades e outras figuras públicas, mais se sabe sobre o suposto interlocutor. Outro desafio é a variedade ortográfica. Por exemplo, no nosso estudo pronominal, o sujeito *nós* produz grafias como *nóis, nos, nóix,* e assim por diante. Poucas ferramentas existem para a normalização ortográfica em Português, então o tratamento e anotação de dados exige ainda muito processamento manual, o que é dispendioso em termos de tempo e esforço. O maior potencial é com certeza a larga escala de coletas de dados, que apesar de terem que ser parcialmente processados manualmente, ainda demonstra vantagens em termos de custo comparado à transcrição de comunicação oral.

Questão 8

Quais são geralmente os requisitos de uma análise linguística de perfil científico considerada rigorosa e confiável? Em que as tecnologias e ciências da informação e da computação podem auxiliar concretamente? Que recomendações ressaltaria nesse sentido?

Uma análise linguística de perfil científico rigoroso e confiável é aquela que é reproduzível. Nossa capacidade tecnológica de armazenamento de dados é extensa, o que permite que todos os estágios de processamento dos dados sejam armazenados, dos dados crus, até a versão final processada dos dados. Tutoriais e recursos de como organizar projetos colaborativos que são reproduzíveis existem de graça e fácil acesso na internet. Plataformas como a Open Science Framework (osf.io) contém exemplos e modelos a partir do primeiro passo em qualquer processo, o pré-registro de estudos, até a comunicação final de resultados. A minha maior recomendação é usar outros projetos abertos como base de organização e gerenciamento de projetos, e ser transparente em relação a todos os passos em qualquer projeto de pesquisa.

Questão 9

Em que medida esses requisitos repercutem em ações desenvolvidas em prol da constituição ou reunião de acervos de dados linguísticos (grandes bancos de dados), a fim de potencializar diversificadas análises linguísticas? Tendo em vista movimentos como, por exemplo, o de uma equipe de comissões científicas e estratégicas da Abralin e do GT de Sociolinguística da ANPOLL trabalhando em prol de um repositório digital intitulado *Projeto Plataforma da Diversidade Linguística Brasileira*, o que, a partir de sua ótica de formação em ciência da informação, seria estratégico considerar no planejamento das ações?

Grandes bancos de dados que são representativos da fala de comunidades bem definidas, como é o caso do Projeto Plataforma da Diversidade Linguística Brasileira, são essenciais para a diversificação de análises linguísticas. A estratégia de usar times grandes e diversos, como é o caso nesse projeto, é a melhor forma de garantir que o acervo seja de utilidade para outros grupos de pesquisa diversificados no Brasil e no exterior. Com os dados disponíveis a pesquisadores, um sistema de pré-registro de estudos se faz útil para não só a comunicação de projetos que estão em andamento, mas também para a abertura de colaborações interdisciplinares.

Questão 10

Que possibilidades de análises, recursos e perfis têm o potencial de enriquecer o tratamento de dados de variação linguística? Em que metanálises agregam poder exploratório a fenômenos em variação? Que articulação hoje em dia é possível perspectivar entre Sociolinguística e Ciência da Informação, ou a partir de Sociolinguística?

A Sociolinguística, especialmente a área de variação e mudança, tem seus métodos de análise quantitativo bem estabelecidos, com o uso de regressão logística com soma de contrastes desde os anos 70. Há muitos estudos em Sociolinguística que analisam as mesmas variáveis. Esses resultados agregados nos dão uma base de exploração para novos estudos mais inovativos, como os que fazem uso de dados de redes sociais. A Ciência da Informação entra com esse aspecto inovativo, que nos permite minerar e processar uma larga escala de dados. Esse aspecto inovativo casado com os métodos de análise quantitativa usados na Sociolinguística ajuda a expandir o que entendemos sobre essas variáveis sociolinguísticas.

Questão 11

Tendo em vista sua expertise em diversas ações formativas (minicursos, entre outras), o que geralmente tem tido maior acolhida entre pesquisadores (docentes e discentes) e o que normalmente lhes tem causado algum desconforto ou até os tem silenciado? Que (inter)ações, leituras e/ou atividades de formação lhes indicaria, no intuito de começarem a ver com outros olhos e possivelmente encarar com mais segurança o espaço (tecnológico) de análise de dados?

Existe um desejo e empenho entre os pesquisadores brasileiros por aprender novos métodos e ferramentas computacionais. Durante minhas ações formativas eu encontrei muito mais acolhimento do que o contrário. Os pesquisadores brasileiros, na minha experiência, já vão atrás de conhecimento e estão sempre se aprimorando, nunca param de estudar. O maior desconforto é a ideia de ciência aberta em todos os estágios de tratamento e análise de dados. É desconfortável saber que outros pesquisadores vão poder inspecionar e criticar o nosso trabalho por inteiro, incluindo o processo, e não só o produto final. Mas volto a iterar que grupos grandes, com colaborações interdisciplinares, que quebram as barreiras entre ciência humanas e exatas é

como vencemos esse desafios e criamos uma comunidade de prática capaz de treinar pesquisadores novatos e aprimorar pesquisadores experientes. A criação de grupos de leitura que abordam títulos disponíveis gratuitamente como R for Data Science (https://r4ds.had.co.nz/) e Introduction to Modern Statistics (https://openintro-ims.netlify.app/) são as minhas recomendações atuais.

Referências

PICORAL, A. MINICURSO – Analisando construções intensificadoras encontradas em tweets por meio do R. Youtube, 2022. Disponível em: https://www.youtube.com/watch?v=fe1DIiIlLxE&list=PLYlbzUR7D2soXo9yP0yOhbPuu2el2YwtF. Acesso em 31 de maio, 2022.

PROJETO PLATAFORMA DA DIVERSIDADE LINGUÍSTICA BRASILEIRA. Museu da Língua Portuguesa, 2022. Disponível em: https://www.museudalinguaportuguesa.org.br/projeto-plataforma-diversidade-linguistica-brasileira/. Acesso em: 31 de maio, 2022.

PICORAL, A.; Stumpf, E.; GOULART, L.; DE BARROS, I.C.; SOMMER-FARIAS, B.; Matte, M.L.; GARCIA, M.C.; BERTHO, M.C.. Parallels Between Spoken and CMC Language: Do Tweets Reflect Spoken Language Choices?. In: *Proceedings of the 8th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-Corpora 2021),* 28-29 October, 2021 (p. 84). https://cmc-corpora.org/publications/cmc-corpora-2021/