A documentação de variação linguística em amostras de tweets por meio do software R: limites e potencialidades

Mariana Gonçalves da Costa (mariana.goncalves@letras.ufrj.br),
Pedro Giovani Duarte Poppolino (poppolino@ufrj.br),
Lais Lima de Souza (lais@letras.ufrj.br)
Orientador / Directeur de recherche: Marcia dos Santos Machado Vieira
Instituição / Institution: Universidade Federal do Rio de Janeiro (UFRJ)

Resumo / Résumé

Durante a pandemia de COVID-19, deparamo-nos com o desafio de coletar dados em um contexto completamente novo, que afetou diretamente as investigações desenvolvidas no âmbito do Projeto PREDICAR. Como já estávamos interessados no uso da linguagem R em nossas pesquisas, decidimos direcionar esforços para a coleta automatizada de dados linguísticos ao aliar esta ferramenta à rede social Twitter. O Twitter é hoje um espaço de interlocução online já consagrado como uma das redes sociais mais populares no Brasil e no mundo, constituindo um grande acervo de dados autênticos. Seu extenso uso faz com que se configure como um banco de dados de um gênero próprio (tweet/tuíte) que já se convencionalizou socialmente como um legítimo espaço de interlocução. Neste contexto, mostra-se de suma importância reconhecer esta rede como um banco de dados autênticos relevante (e promissor) às pesquisas linguísticas. A partir disso, pretendemos mapear algumas das características específicas desse gênero com as quais nos deparamos durante as pesquisas do Projeto PREDICAR, assim como questões abertas à reflexão em relação ao manejo desses dados. Dessa forma, nesta apresentação, buscamos relatar a nossa experiência com o programa R conjugado à plataforma Twitter, a fim de mapear as vantagens e desafios enfrentados durante o processo. Acreditamos que a linguagem R e o Twitter podem ser utilizados como ferramentas computacionais promissoras para a linguística de corpus por serem capazes de coletar dados de maneira extensa e automatizada. Há, entretanto, a necessidade de considerar os múltiplos fatores que norteiam as decisões relativas à triagem de enunciados, sendo eles: (i) a interferência da formulação da busca nos resultados da coleta e (ii) como a contextualidade (GOLDBERG, 2016) pode influenciar na definição da extensão e da abrangência da amostra, visto que, por o Twitter se tratar de uma plataforma com característica fortemente imediatista, acontecimentos sociais podem levar determinados tópicos a terem picos de frequência.

Introdução / Introduction

A proposta desta apresentação centra-se no relato de experiência sobre o que temos aprendido com a utilização da linguagem de software R para a coleta de dados linguísticos produzidos na rede social Twitter. Tendo em vista a necessidade de reunir um grande número de dados autênticos para as análises estatísticas que interessam às pesquisas desenvolvidas no projeto PREDICAR, aliamos o conhecimento de alunos da graduação de Letras e da graduação em Computação da UFRJ a fim de garantir a construção coletiva de metodologias úteis às pesquisas linguísticas.

- Por que o R e o Twitter?

Gentry (2014) observa que a possibilidade oferecida pelo Twitter aos seus usuários de postar um grande número de mensagens curtas (os "tweets"), fez dessa rede social uma ferramenta valiosa para a mineração de dados. Apesar de haver diversos programas para a análise de dados linguísticos, o R se diferencia por, conjugado ao Twitter, permitir a coleta, manipulação e análise de dados em um só ambiente.

Objetivos / Objectifs

- Relatar as vantagens e desafios identificados durante o processo de coleta de dados linguísticos com o programa R conjugado à plataforma Twitter;
- Destacar os benefícios da associação entre as duas ferramentas para as pesquisas linguísticas.

A variação linguística e a coleta de dados do Twitter com o R / La Variation linguistique et la collecte de données de Twitter avec R

Como mencionado por Blommaert (2018), atualmente vivenciamos um cenário em que as fronteiras entre o mundo virtual e o mundo real são cada vez mais dispersas. Levando isso em consideração, como sociolinguistas, o mundo virtual nos abre mais uma porta de possíveis investigações por meio da análise de variantes no contexto digital. Para tal, novas formas de coleta e análise de dados são necessárias.

Em relação a isso, a combinação do software R com a plataforma Twitter pode proporcionar:

- A coleta de um corpus de grande porte de variadas línguas e países;
- Contemplação das variações encontradas no meio virtual;
- Verificação do comportamento dos falantes jovens em uma plataforma com maior privacidade em comparação com as demais redes sociais (ROCHA, 2020).

Considerações / Considérations

Até o momento pudemos disponibilizar uma vídeo-aula no evento Festival do Conhecimento da UFRJ (2020) que obteve mais de 550 acessos no Youtube, chegando a mais de 60 compartilhamentos em nosso post de divulgação no Facebook. Por conta desse vídeo, diversos linguistas de todo o Brasil nos contactaram trazendo suas dúvidas e agradecendo pelo conteúdo. Notamos que o processo de preenchimento da *Twitter Application* exigida pela plataforma era de especial dificuldade para a maioria dos usuários. A fim de resolver esse problema, disponibilizamos um manual detalhado em português que pode ser acessado na descrição do vídeo.

Em relação à nossa pesquisa, conseguimos aumentar nosso corpus em cerca de dez vezes, passando a contemplar duas construções ao invés de uma. Além disso, tivemos a possibilidade de aplicar a análise colostrucional como uma melhor alternativa de análise de colocação, o que não havia sido possível anteriormente pois não possuíamos dados o suficiente para esse tipo de análise.

Entretanto, encontramos também desvantagens, sendo elas: as restrições da ferramenta de busca do Twitter, o limite de dados coletados em um período de tempo, coleta limitada a dados de até 7-9 dias para trás, dificuldade de contemplação de abreviaturas ou diferentes ortografias de uma mesma palavra ou construção, contexto limitado com maior prevalência de usuários homens e jovens, e, por fim, boa parte das informações disponíveis sobre os programas estão em inglês e direcionadas a usuários com conhecimento prévio de programação.

Conclusão / Conclusion

Podemos perceber que há uma **grande demanda na área** pelo tipo de material que buscamos disponibilizar.

Há ainda a necessidade de verificar que a forma como uma consulta é feita através da linguagem R pode influenciar no conjunto de dados obtidos.

Além disso, a **prolificidade** e a diversidade acadêmica da comunidade dessa linguagem, em conjunto ao **crescente engajamento de linguistas** a essa ferramenta, impulsionará a criação de novos pacotes de funcionalidades, trazendo **novas possibilidades** de pesquisa.

Portanto, apesar das limitações atreladas à utilização da **linguagem R** conjugada ao **Twitter**, os benefícios dessas ferramentas mostram um **enorme potencial** para a linguística de corpus.

Referências / Références

GENTRY, J. **Twitter client for R**. 2014. Disponível em: http://geoffjentry.hexdump.org/twitteR.pdf. Acesso em 10 de março de 2021.

GOLDBERG, A. Compositionality. In: RIEMER, Nick. **The Routledge Handbook of Semantics**. London and New York: Routledge, p. 415-433, 2016.

ROCHA, Luciana Lins. **O Twitter como lócus de performances dissidentes de feminilidade**. Revista Indisciplina em Linguística Aplicada, Rio de Janeiro, v. 1, n. 1, 2020. Disponível em: https://revistas.ufrj.br/index.php/rila/article/view/39186. Acesso em: 10 mar. 2021.

Jan Blommaert on the online offline nexus. Jan Blommaert, 19 jun. 2018. Vídeo (6min). Disponível em: https://www.youtube.com/watch?v=z323kQgLCxE. Acesso em: 20 fev. 2021.