RESENHA

Contribuições da linguística de *corpus* em diferentes domínios

Marta Deysiane Alves Faria SOUSA
Universidade Federal de Sergipe (UFS)

RESUMO

Resenha-se, neste texto a conferência A *versatilidade da linguística de corpus quantitativa: exemplos de ortografia, fonologia e linguística forense* proferida pelo Professor Stefan Th. Gries, no evento Abralin Ao Vivo – *Linguists Online* no dia 08 de junho de 2020, e mediada pela Doutora Fernanda Canever. O objetivo principal da conferência foi o de demonstrar como a linguística de *corpus* quantitativa pode ser utilizada em diferentes domínios do conhecimento. Para tanto, empregando diversos recursos estatísticos e de visualização gráfica, o conferencista faz uma explanação de cinco estudos de caso: os dois primeiros relacionados ao *corpus* Ortografia da Internet Espanhola (OIE), o terceiro sobre como os falantes sinalizam o final de turnos para os ouvintes e, os dois últimos, na área de linguística forense, como fundação de sua argumentação.

ABSTRACT

This text is a review of the lecture *The Versatility of Quantitative Corpus Linguistics: examples from orthography, phonology, and legal/forensic linguistics* delivered by Professor Stefan Th. Gries at the Abralin Ao Vivo – Linguists Online event on June 8th 2020 and mediated by Dr. Fernanda Canever. The main objective of this lecture was to demonstrate how quantitative *corpus* linguistics may be used in different domains. In order to do so, the lecturer used a wide range of statistical resources and graphics, he also explained five case studies: the first two related to the *corpus* Spanish Internet Orthography, the third on how speakers signals the end of turn for the hearers, and the last two ones on legal/forensic linguistics. These cases were the basis for his claim.

OPEN ACCESS

EDITADO POR
Raquel Freitag

AVALIADO POR
Dany Thomaz Gonçalves

DATAS
Recebido: 16/06/2020
Aceito: 12/07/2020
Publicado: 29/07/2020

COMO CITAR
Sousa, M. D. A. F. (2020). Contribuições da linguística de *corpus* em diferentes domínios. *Revista da Abralin*, v. 19, n. 2, p. 1-5, 2020.

PALAVRAS-CHAVE
Linguística de *corpus* quantitativa. Análise Estatística. Linguística.

KEYWORDS
Quantitative *corpus* linguistics. Statistical analysis. Linguistics.

Pretende-se com este texto resenhar a conferência A *versatilidade da linguística de corpus quantitativa: exemplos de ortografia, fonologia e linguística forense* proferida pelo Professor Stefan Th. Gries, no evento Abralin Ao Vivo – *Linguists Online* no dia 08 de junho de 2020, e mediada pela Doutora Fernanda Canever. O propósito da conferência foi oferecer à audiência uma visão de como a linguística de *corpus* quantitativa pode ser útil para diferentes domínios. O conferencista faz, então, sua argumentação baseada em cinco estudos de caso, os dois primeiros relacionados ao *corpus* Ortografia da Internet Espanhola (OIE), o terceiro sobre como os falantes sinalizam o final de turnos para os ouvintes, e os dois últimos, na área de linguística forense.

Gries inicia sua fala relatando a pesquisa sobre o apagamento de "-d" no seguimento "-ado" no *corpus* OIE. Neste estudo, foram utilizados testes de qui-quadrado, para comparar a frequência de apagamento de "-d" no OIE e no *corpus* baseado em conversas de *chat* do estudo de Llisterri (2002), e um teste de correlação entre esse apagamento e a vulgaridade das palavras. No primeiro teste, o conferencista observou que não houve diferença na frequência do apagamento nos dois *corpora* (p=0,75), não havendo efeito do tipo de interação (*chat* ou comentários e descrição). O segundo teste confirmou a correlação entre o apagamento e a vulgaridade das palavras ($p_{\ binomial\ test}$ <0,001), sendo uma forte correlação.

Ao comparar 50 *tokens* mais frequentes de palavras com as duas realizações, uma sem e a outra com o apagamento, com o *Corpus* do Espanhol Padrão, Gries fez um gráfico de dispersão para visualizar o efeito de frequência na ocorrência do apagamento, indicando que quanto maior a frequência de uma palavra, menor é o apagamento. Para Gries, existem três justificativas para esse efeito: frequência da palavra (quanto maior o grau de entrincheiramento, maior a probabilidade de o falante recorrer à variante padrão); ausência de tonicidade na penúltima sílaba e, efeito pragmático da deleção, explicado com maior ênfase. Para ele, o apagamento de "-d" em palavras que são mais frequentes demonstraria certo desprestígio social. Por outro lado, se a palavra já tiver sido modificada de alguma forma, a chance de o apagamento ocorrer é três vezes maior, demonstrando que a modificação prévia e o apagamento se agrupam.

Após discorrer sobre o apagamento de "-d", Gries relata sua pesquisa com a repetição de caracteres para indicar atitudes e emoções de forma similar à prosódia na fala. O conferencista aponta como explicação para a utilização de repetição o princípio de iconicidade da quantidade (no qual "a quantidade de material fonético reflete a qualidade/intensidade ou a quantidade/pluralização").

Para Gries, este princípio cognitivo já foi estudado em outras línguas e lida melhor com o fenômeno da repetição por não tentar justificá-la por meio de uma frase que tenta remontá-la.

A hipótese de Gries é de que existe relação entre o efeito desejado no uso da prosódia e a repetição de caracteres. O conferencista encontrou três formas de repetição: no início e no fim das palavras, e quando a repetição é a palavra toda. Em sua explanação, ele se atém somente à repetição no início das palavras.

Fazendo um teste de correlação, Gries descobriu que existe uma correlação forte entre o tamanho da repetição e sua frequência no *corpus* (T=-0,86***). Em seguida, utilizando um diagrama de barras agrupadas, o conferencista demonstrou que também há efeito fonológico, pois os caracteres mais frequentemente repetidos correspondiam a vogais e glides e os menos, a consoantes. Marcadores discursivos, palavras que expressam emoções, palavras usadas para se referir as pessoas, construções como "Eu + sintagma verbal", adjetivos positivos são as classes de palavras nas quais o fenômeno ocorreu com maior frequência.

Para Gries, os dados encontrados revelam tendências que podem ser relevantes para uma representação de exemplares. Isso porque, no apagamento, há diferentes fatores interagindo para que ele ocorra como efeitos pragmáticos, sociolinguísticos, semânticos, fonológicos e de frequência, e, na repetição, a iconicidade e fatores articulatórios. Gries acredita que a comunicação mediada por computadores apresenta características motivadoras de mudança, além de ser uma área que comporta bastante inovação.

O terceiro caso citado pelo conferencista observa se os falantes dão pistas de que terminaram o turno de fala. O pesquisador levanta duas hipóteses para explicar como os falantes sabem que o turno de fala terminou: a lexical e a taxa de elocução. Contudo, Gries excluiu a primeira hipótese baseado em estudos dialetais, nos quais foi concluído que os ouvintes se ajustam às características específicas dos falantes. Para seu estudo, Gries, extraiu uma amostra aleatória (800 turnos de fala com dez palavras) do *Corpus* Nacional Britânico (CNB) e controlou as seguintes variáveis: duração da palavra (variável dependente); posição no turno (preditor principal); posição da palavra nuclear; frequência de *type*; tamanho fonético do *type*; *surprisal* (grau de informatividade da palavra no turno); diferença da média anterior (mudança na média da duração da palavra no turno). O conferencista ressaltou a existência de efeito de outros fatores aleatórios como: arquivo/ falante, palavra e classe de palavra.

Gries optou por fazer um modelo de regressão linear misto, com eliminação passo atrás (*backwards*). Nesse modelo, incorporam-se todas as variáveis e depois, por etapas, cada uma pode ser eliminada, até chegar ao modelo final. Os resultados revelaram que, mesmo havendo um grande volume de dados confusos e idiossincrasias, a correlação entre a posição no turno e a duração foi encontrada, sendo ela mediada pelo núcleo, corroborando a hipótese da taxa de elocução. Em outras palavras, a posição do núcleo pode acelerar ou diminuir a velocidade na fala. Se ele está no início, a velocidade diminui linearmente, se está no meio, diminui-se a aceleração e depois ela volta a subir, fazendo um efeito curvilíneo que pode ser visualizado durante a conferência.

O penúltimo caso, Smith contra os Estados Unidos, refere-se a uma pena judicial agravada porque o condenado estava "usando uma arma de fogo" como moeda de troca e o juiz, baseando-se no significado dicionarizado da palavra "usar", entendeu que "usar" também abarcaria o sentido de "troca". Gries esclarece que, se o significado de uma palavra não estiver descrito na lei, a Suprema Corte Norte-Americana prescreve que deve ser considerado o significado "comum" ou "ordinário" de tal palavra.

No entanto, as palavras não são opacas, seus significados fazem parte de um *continuum*. Gries acredita que a Suprema Corte, quando menciona significado ordinário, deve querer dizer respeito à frequência relativa de tal significado dentro do espectro "possível -> comum -> mais frequente -> prototípico -> exclusivo".

A partir desse caso, Gries e colaboradores usaram um *script* do R para pesquisar o lema "usar" com contexto seguinte composto de palavras relacionadas a armas no *Corpus* Contemporâneo do Inglês Americano. Os resultados evidenciaram que "usar" não foi empregado como "trocar" em nenhum dos casos válidos, levando à conclusão de que a corte norte-americana não estava realmente considerando o significado "comum" da palavra, mas o possível.

Para o conferencista, a área jurídica pode ser um vasto campo para aplicação da linguística de *corpus*. Além disso, demonstra a necessidade de especialistas da linguística para lidar com o espaço multidimensional dos significados ditos comuns de uma palavra nesse contexto.

O último caso explorado é de atribuição de autoria, no qual um cliente processou um restaurante por alegar que este lhe deu tratamento discriminatório. Uma das linhas de argumentação do advogado do restaurante era provar que o cliente tinha histórico de processar pessoas e negócios e utilizar contas falsas na internet para acusá-los de discriminação ou de má prestação de serviço.

Gries explicou que para verificar a autoria foram necessários os textos aos quais seriam atribuídos a autoria e um material de treinamento (textos do cliente e textos reconhecidamente de outras pessoas). Foram observadas características lexicais, sublexicais, morfológicas, sintáticas e combinações de fatores (frequência de certas palavras em certos contextos e múltiplas medidas das características citadas). A partir de todo o material, a frequência relativa de cada item foi calculada por meio de um *script* do R em cada arquivo de texto.

Em seguida, o modelo de Floresta Aleatória foi treinado com o material para distinguir o suspeito de todos os outros autores e posteriormente aplicado aos textos os quais a autoria seria atribuída. Os resultados do modelo evidenciaram que todos os textos que não tinham autoria eram provavelmente de autoria do cliente, com um grau de precisão de mais de 90%.

Esses dois últimos casos, conforme Gries, exploram como ferramentas da linguística de *corpus* podem ser usadas para tornar a interpretação de textos jurídicos mais objetiva e justa. Ademais, demonstra como métodos da análise da linguística de *corpus* puderam oferecer uma análise de especialista para atribuir autoria em um caso.

Gries foi bem-sucedido em realizar seu objetivo de demonstrar as contribuições da linguística de *corpus* quantitativa em diversos domínios. Embora não tenha conseguido pormenorizar os dois últimos casos, durante toda a conferência foi vista a relevância de análises estatísticas (desde as mais

simples, como testes de correlação, às mais complexas como o modelo de regressão linear misto) para justificar a refutação de hipóteses, exclusão de variáveis, a modelagem dos dados per si. O conferencista apresentou também diversos meios de visualização gráfica dos resultados, mesmo quando várias e diferentes variáveis estavam envolvidas (diagrama de barras agrupadas, gráfico de dispersão, entre outros). Essa variedade contribui e motiva sociolinguistas a refletir sobre a utilização de recursos estatísticos diferentes para embasar as análises de dados, principalmente, aqueles que possuem diversidade de tipos de variáveis.

REFERÊNCIAS

THE Versatility of Quantitative *Corpus* Linguistics: examples from orthography, phonology, and legal/forensic linguistics. Conferência apresentada por Stefan Th. Gries. [s.l., s.n], 2020. 1 vídeo (1h 47min 24s). Publicado pelo canal da Associação Brasileira de Linguística. Disponível em: https://www.youtube.com/watch?v=HuOc6AzQ4Ig&t=757s. Acesso em: 08 jun 2020.