REVISTA DA **ABRALIN**

REVIEW

Contributions of *corpus* linguistics in different domains

Marta Deysiane Alves Faria SOUSA
Federal University of Sergipe (UFS)

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 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.

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.

OPEN ACCESS

EDITED BY
Raquel Freitag

REVIEWED BY
Dany Thomaz Gonçalves

DATES
Received: 16/06/2020
Accepted: 12/07/2020
Published: 29/07/2020

HOW TO CITE
Sousa, M. D. A. F. (2020). Contributions of *corpus* linguistics in different domains. *Revista da Abralin*, v. 19, n. 2, p. 1-5, 2020.

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

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

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 Abralin Ao Vivo – Linguists Online event on June 8th, 2020 and mediated by Dr. Fernanda Canever. The purpose of the lecture was to offer an overview of how useful quantitative *corpus* linguistics can be in different domains. The lecturer based his argument on five case studies: the first two related to the *corpus* Spanish Internet Orthography (SIO), the third one on how speakers signal the end of turn for the hearers, and the last two ones on legal/forensic linguistics.

Gries starts his talk reporting his research on the deletion of "-d" in the segment "-ado" in the SIO. In this study, chi-squared tests were used to compare the frequency of deletion of "-d" in SIO to the *corpus* in the study by Llisterri (2002) which was based on chat conversations, and a correlation test between this deletion and the vulgarity of words was performed. In the first test, the researcher observed that there was no difference in the frequency of deletion in either corpora ($p=0.75$), and that there was no effect of the kind of interaction on the data (chat or comments and descriptions) either. The second test confirmed the correlation between the deletion and the vulgarity of words ($p_{\,binomial\ test} <0.001$), which was strong.

When comparing the most frequent 50 word tokens with the two realizations (with and without "-d") with the Standard Spanish *Corpus*, Gries made a scatterplot in order to visualize the effect of frequency in the occurrence of the deletion, indicating that the higher the word frequency is, the lower the deletion is. For Gries, there are three reasons for that effect: word frequency (the higher the overall degree of entrenchment of a word is, the higher the probability of the speaker use the standard variant is); lack of stress on the penultimate syllable; and the pragmatic effect of the deletion, which was focused on the explanation. For him, the deletion of "-d" in more frequent words has low social value. On the other hand, if the word has already been modified someway, the chances of the deletion appear are three times higher, which demonstrates that spelling modification and deletion cluster.

After talking about the deletion of "-d", Gries reports his research on the repetition of characters to indicate attitudes and emotions similarly to the prosody in speech. The lecturer argues that the explanation for using the repetition is the iconicity principle of quantity (according to which "the amount of phonetic material reflects quality/intensity or quantity/pluralization"). For Gries, this cognitive principle has already been studied in other languages and handles the repetition phenomenon better, since it does not try to justify the repetition through a sentence that tries to replicate it.

Gries's hypothesis is that there is a relation between the desired effect on the use of prosody and the repetition of characters. The lecturer found three kinds of repetition: at the beginning and at the end of words, and when the repetition is the whole word. In his explanation, the lecturer only focuses on the repetition in the beginning of words.

Through a correlation test, Gries discovered that there is a strong correlation between the length of the repetition and its frequency in the *corpus* (T=0.86***). Then, the lecturer demonstrated through a bar plot that there is also a phonological effect, since the most frequent repeated characters corresponded to vowels and glides, and the least frequent characters corresponded to consonants. Discourse markers, expressions of emotion, terms of address, I+verb phrases, and positive adjectives are the classes of words in which the phenomenon often occurred.

For Gries, the data found reveal tendencies that may be relevant for an exemplar theory. This happens due to the fact that there is an interaction of different features for the deletion such as pragmatic, sociolinguistic, semantic, phonological and frequency effects, and for the repetition, the iconicity and articulatory features. Gries believes that the computer mediated communication displays characteristics of linguistic change, and it is also a field that comprises innovation.

The third case reported by the lecturer observed if the speakers give cues indicating turn-finality. The researcher points to two hypotheses to explain how speakers know the turn-finality: the lexical and speech-rate. However, Gries excluded the first hypothesis based on dialectal studies, which the conclusion was that listeners adjust to speaker-specific speech characteristics. For his study, Gries extracted a random sample (800 10-word turns) from the British National *Corpus* (BNC) and controlled for the following variables: word duration (dependent variable); position in turn (main predictor); nucleus (position of the nucleus); frequency of word type; phonetic length of the word type; surprisal (informativity of the word in the turn); difference of mean previous (change of the mean in the duration of the word in the turn). The lecturer highlighted the existence of other random effects such as file/speaker, word, and word class.

Gries decided to perform a mixed linear regression model with backward elimination. In this model, all variables are inserted and, then, eliminated step by step, until a final model is reached. The results revealed that even with messy observational data, confounds, and idiosyncrasies, the correlation between the position in turn and duration was found and was mediated by the nucleus, confirming the speech-rate hypothesis. In other words, the position of the nucleus may accelerate or decrease the speed of the speech. If it is in the beginning, the speed linearly decreases, if it is in the middle, the speed decreases, and later the speed starts to increase again, displaying a curvilinear effect which was showed during the lecture.

The penultimate case, Smith vs. USA, refers to a sentence enhancement due to the fact that the defendant was "using a firearm" in exchange for drugs, and the judge sentenced him based on the dictionary meaning of the word "use" which would also comprise the meaning of "trade". Gries clarifies that if the meaning of a word is not described in the statue, the Supreme Court of the United States (SCOTUS) states that the "common" or "ordinary" meaning of such word must be considered.

However, words are not opaque, their meanings belong to a continuum. Gries believes that when the SCOTUS mentions "ordinary" meaning, they must be referring to the relative frequency of such meaning inside the spectrum "possible->common-> most frequent-> prototype-> exclusive".

From this case, Gries and colleagues used an R script to search the lemma "use" followed by weapon related words in the *Corpus* of Contemporary American English (COCA). The results showed that "use" did not have the meaning of exchange in any of the valid cases, leading to a conclusion that the SCOTUS was not really considering the ordinary meaning of the word, they were considering the possible one.

For the lecturer, the legal/forensic area may be a wide field for the application of *corpus* linguistics. Besides, he demonstrates that there is a need for a linguistic expert to deal with the multidimensional space of the ordinary meanings of a word in this context.

The last case concerns the attribution of authorship, which a costumer sued a restaurant for having a discriminatory treatment. One of the strategies of the restaurant's attorney was to prove that the plaintiff had a history of suing people and business and also use fake accounts to accuse them of discrimination and bad service on the internet.

Gries explained that in order to verify the authorship, the texts to which the authorship would be attributed and a training material (comprising the texts of the customer and texts known to be written by other authors) were necessary. Lexical, sub-lexical, morphological, syntactic features, and also a combination of factors (frequency of certain words in certain contexts and multiple metrics from the mentioned features) were observed. From all the material, the relative frequency of each item was calculated through an R script in each text file.

Next, a random forest was trained with the material in order to distinguish the suspect from all the other authors, and, then, it was applied to the texts to be attributed the authorship. The results of the model evidenced that all the texts that did not have authorship were probably written by the customer (more than 90% of accuracy).

The two last cases, according to Gries, explored how *corpus* linguistics tools may be used to make the interpretation of texts more objective and fairer. In addition, he demonstrates how *corpus* linguistics methods could offer an expert analysis to attribute authorship in one case.

Gries successfully accomplished his goal of demonstrating the contributions of quantitative *corpus* linguistics in different domains. Even though, the lecturer was not able to describe the last two cases in deep, throughout whole lecture the relevance of statistical analysis (from very simple ones, such as correlation tests, to more complex ones, such as mixed-model regression) in order to justify the rejection of hypotheses, the exclusion of variables, and the model itself was observed. Gries also presented different means of visualizing results, even when there were different and many variables involved (bar plot, scatter plot, among others). This variety contributes and motivates sociolinguists to reflect on the use of different statistical resources to found the data analysis, mainly the ones that have a wide range of types of variables.

REFERENCES

THE Versatility of Quantitative *Corpus* Linguistics: examples from orthography, phonology, and legal/forensic linguistics. Lecture delivered by Stefan Th. Gries. [s.l., s.n], 2020. 1 video (1h 47min 24s). Published by the channel of the Brazilian Association of Linguistics. Available at: https://www.youtube.com/watch?v=HuOc6AzQ4Ig&t=757s. Access: June 8th, 2020.