A conversation about digital meaning-making, research methods and open access

Uma conversa sobre a construção de sentido no espaço digital, métodos digitais de pesquisa e acesso aberto

Camila Lívio[1],* iD

Marcia dos Santos Machado Vieira[2],** iD

[1] University of Georgia, Athens, Georgia, United States
[2] Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
E-mail: camila.emidio25@uga.edu
E-mail: marcia@letras.ufrj.br

* Dr. Camila Lívio Emidio (camila.emidio25@uga.edu) is a research professional at the Center for International Trade and Security at the University of Georgia. Much of her research explores language use and change in digital communities, utilizing digital scholarship techniques such as corpus linguistics and text mining, while also engaging with issues of open research and open access.

** The interviewer, Doctor Marcia dos Santos Machado Vieira (marcia@letras.ufrj.br), is the editor-in-chief of this journal, she works at Federal University of Rio de Janeiro in the area of (socio)linguistic and socioconstructionist studies of Portuguese as well as in the area of open access repositories and datasets of Romance languages. She is researcher of the funding agencies CNPq and Faperj. She coordinates research projects in these areas (Predicar, VariaR and inCorpora projects) as well as the Capes PrInt project of Vernacular Letters Postgraduate Program/UFRJ – Vozes e escritas nos diferentes espaços da língua portuguesa.

Abstract

Camila Lívio holds a PhD in Romance Languages - Linguistics from the University of Georgia in the United States, and is currently a research professional at the Center for International Trade and Security at the same institution. Previously, she served as the Scholarly Communications Librarian in the Department of Research and Computational Data Management at the UGA Libraries. Her interests and experiences lie at the intersection of open research, computational social science, and linguistics. In her research, she has published on topics such as digital communication (e.g., Twitter) and the construction of meaning in digital environments, particularly through the analysis of product reviews on platforms like Amazon and Mercado Livre. She employs methods grounded in corpus linguistics, computational text analysis, and Digital Humanities to deepen her understanding of linguistic and cultural phenomena. In this interview, she shares insights into her professional journey and explores the intersections between linguistics, open access, and digital methods.

Keywords

Digital communication, Text mining, Open search, Linguistics

Resumo

Camila Lívio é doutora em Línguas Românicas — Linguística pela Universidade da Geórgia, nos Estados Unidos, onde ocupa o cargo de pesquisadora no Centro de Comércio Internacional e Segurança Global. Anteriormente, atuou como Coordenadora de Comunicações Acadêmicas no Departamento de Pesquisa e Gestão de Dados Computacionais da biblioteca central da mesma instituição. Seus interesses e experiências se concentram na interseção entre pesquisa aberta, ciência social computacional e linguística. Em sua pesquisa, publicou sobre temas como comunicação digital em plataformas como o antigo Twitter, e a construção de significado em ambientes digitais, com ênfase na análise de avaliações de produtos em plataformas como Amazon e Mercado Livre. Ela utiliza métodos baseados em linguística de corpus, análise computacional de texto e Humanidades Digitais para aprofundar sua compreensão dos fenômenos culturais, sociais e linguísticos. Nesta entrevista, ela compartilha insights sobre sua trajetória profissional e explora as interseções entre linguística, acesso aberto e métodos digitais.

Palavras-chave

Comunicação digital, Mineração de texto, Pesquisa aberta, Linguística

Question 1

First of all, many thanks for this interview! You once told me that your work focuses on data and research infrastructure with open access, and that your research interests include language variation of Portuguese and other languages with a 'digital/computational flavor.' Could you tell us a little bit about it, your endeavors at the University of Georgia, and your experience and plans in terms of working in and/or for the linguistic area? Also, what stands out in your academic-scientific trajectory that led you to become interested in computational analysis of data samples and also in affective language?

C.L: Thank you for the opportunity to talk a little bit about my professional trajectory so far. In 2016, I came to the University of Georgia to pursue a Ph.D. in the Department of Romance Languages with Professor Chad Howe. Back then, I was interested in comparative approaches to Romance Linguistics but wanted to pursue it through a more quantitative approach, which differed from my previous research experiences during my master's and undergraduate studies in Brazil. In my first year, Chad suggested that I take a class on statistics for linguists, and, looking back, I believe that class was pivotal in shaping everything that followed, including my dissertation and my first job after obtaining my Ph.D. The class, superbly taught by Dr. Peggy Renwick, involved learning how to use the statistical software R. Knowing nothing about programming, I found it incredibly enjoyable to learn a new way to analyze languages and data. These experiences led me to co-found a chapter of R-Ladies Global in Athens, along with a colleague who was a postdoc in infectious diseases at the time, Dr. Ana Bento. I enjoyed this data-driven approach to linguistics so much that I ended up taking other quantitative courses, such as Finite State Linguistics, which proved extremely beneficial in delving deeper into formal languages, finite state morphology, regular expressions, and corpus linguistics. These and other esteemed professionals I met along the way, such as Dr. John Hale and Dr. Adriana Picoral, were instrumental in helping me explore new perspectives on language and linguistic theories. As far as affective language goes, around the same time that I was taking that quantitative class, I was also studying sociolinguistics and remember reading a paper by Esther Brown and Mayra Cortés-Torres in which the authors showed how speakers of Puerto Rican Spanish use the intensifiers *muy/mucho* and *bien*. I quickly became interested in learning about the distribution of *muito* and *bem* in Brazilian Portuguese. In the process of writing this class paper, which was later published Deve see (Lívio and Howe 2020), I became fascinated by intensifiers and observing the way they encode the speaker's perspective in the utterance. It is also intriguing to me how the Romance languages, compared to English, for instance, have such a wide variety of strategies to encode intensification, such as through affixation, as in *elegantérrimo*, and the nuances in their meaning. Relatedly, I am also interested in the

speakers' perceptions around the use of intensifiers. Modeling after an outstanding piece of research by Vaughn *et al.* (2018), in my dissertation, I designed a survey that asked native speakers of Brazil and Mexico to choose from a list of attributes to label the sentences they heard in the stimuli according to how friendly, annoying, modern, conservative, feminine, or intelligent the speakers sounded. The sentences in the stimuli contained several intensifying forms. One interesting piece of data that I discovered was that female and male native speakers from Brazil chose attributes at an inversely proportional rate: while female participants chose positive attributes, like intelligent and friendly, more frequently, negative attributes like immature and annoying were chosen more frequently by those who identify as male when listening to the audio. I concluded that gender is a sociodemographic factor that is not only important for predicting the frequency of use of intensifiers (cf. Fuchs, 2017), but also for predicting social perceptions regarding their use.

Methodologically speaking, I typically work under a Digital (or computational) Humanities, Digital Linguistics, and Corpus Linguistics frame, or a combination thereof. I like to use linguistic data from social media and digital communication in general because it offers a whole new dimension to language through the use of symbols, emojis, caps lock etc. My research interests intersect well with my position in that my responsibilities include assisting students with their projects and designing research that lends itself to open research practices, such as gathering online and publicly available data, as well as being part of a large research team. These and other esteemed professionals I met along the way, such as Dr. John Hale and Dr. Adriana Picoral, were instrumental in helping me explore new perspectives on language and linguistic theories.

Question 2

Could you briefly address computational text analysis and its relevance to the study of language?

C.L.: Text mining, understood as the set of computational techniques used to retrieve information from a text or a collection of texts, can be as an extremely useful approach to make sense of language data. Text mining has been around for a long time (the article "A vector space for automatic indexing" by Salton, Wong and Yang was published in 1975!) but has somewhat recently gained more traction with the increasing volume of text data that is generated digitally, as well as with research in natural language processing (Salton; Wong; Yang, 1975). The ability to gather information from texts in a systematic way, such as through the extraction of bigrams, grammatical patterns, geographic location, and frequencies is essential to the formulation of hypotheses about how and where certain phrases are used, which, in turn, can help the researcher generate insights about how languages and

communication work. My favorite part of working with corpus linguistics methods, text mining and computational techniques is the idea of repurposing. For example, sentiment analysis, or opinion mining, was first designed to help businesses and researchers understand public opinion but has been used by linguists to explore language variation, discourse analysis, and language attitudes, serving as an interesting tool to explore the relationship between language, emotion, identity, and culture.

Question 3

Considering the technological turn we are experiencing in linguistics due to the advancements in Information and Computing Technology and Open Research initiatives, which field(s) of linguistics do you consider to be the most methodologically pluralistic in order to harness the benefits of these advancements and address the associated challenges?

C.L.: Fields like Computational Sociolinguistics and Second Language Acquisition, as well as researchers dealing with experimental methods, have contributed enormously to the Open Research Movement. It is now an established practice to see researchers making their datasets and scripts available on open platforms like GitHub, as well as in generalist repositories like the Open Science Framework (OSF) or Zenodo. The recent publication of The Open Handbook of Linguistic Data Management (Kroeker *et al.*, 2018) is symptomatic of this shift in our field and serves as an excellent resource. Furthermore, a recent publication by Coretta *et al.* (2023) invites us to consider the necessity of pedagogy in open research practices to ensure replicable research and minimize biases in linguistics. In Brazil, I follow the initiatives, workshops, and publications led by Dr. Raquel Freitag, Dr. Miguel Oliveira Jr., Dr. Marcia Machado Vieira, and their research groups, which focus on promoting open research practices, such as reproducibility in the Humanities and Social Sciences.

Generally speaking, I believe that any subfield of linguistics can benefit from technological advances and contribute to them as well. For example, focusing on language and learner corpora such as the MACAWS learner corpus (Sommer *et al.*, 2022), and other language corpora like Corpus do Português (Davies, 2006), fake.BR (Monteiro *et al.*, 2018), FakeRecogna (Afonso *et al.*, 2022), and the VariaR project (Machado Vieira; Meireles, 2023), these tools serve a variety of purposes, including language learning, historical linguistics, semantics and pragmatics, and language technologies. I also think that researchers don't necessarily need to approach their analyses from a quantitative standpoint to benefit from understanding data-driven approaches to linguistic analysis, digital technologies, and open research practices. Ultimately, I believe these skills can provide valuable insights into how to conduct an analysis, recognize potential biases, and collaborate effectively as part of a team. To cite a more concrete example, in 2020, I joined a group of early-career linguists and

Ph.D. students to think and write about computational approaches to Portuguese linguistics as part of the GIRL project (Group of Inter-institutional Research on Languages). Our levels of coding skills and interests in linguistics varied, but having a large and diverse group of researchers with a solid understanding of the project was essential for devising a plan to interpret the data and its output effectively. Together, we developed a protocol for coding the data and manually annotated a corpus of tweets from Brazil that provided us with excellent material for discussing the variation between *tu* and *você* in different parts of the country. Through our meetings and collaborative learning, I gained a much better understanding of language variables, coding, statistics, and open research practices.

Question 4

Could you briefly address your research on "Digital Technology meets Romance Linguistics: The Use of Computer-Based Methods for the Study of Language", considering cyber social spaces sources, data-gathering, methods, and main results? Could you tell us about affective language, intensifiers, and innovative high-frequency intensifiers in Portuguese?

C.L.: This research ended up being one of the chapters of my dissertation titled "Digital Approaches to Intensification in Portuguese and Spanish". I was interested in learning about the role of intensifiers in the very particular setting of online customer reviews. In this chapter, I argue that intensifiers are essential elements in customer reviews since reviews are acts of evaluation, they express whether something is good or bad. The choice to study intensifiers in customer reviews stems from research that proposes that these reviews are highly subjective and evaluative texts. Intensifiers like *totalmente* and the suffix *-íssimo* in phrases such as *totalmente desapontado* and *satisfeitíssima* overtly mark the speakers' commitment to the utterance. One of the most interesting aspects of studying intensifiers in highly subjective texts like online reviews is the opportunity to reflect on what these texts achieve beyond the evaluation of a product. I argue that the almost 'compulsory' use of intensifiers in online reviews is connected to the user's need to assert relevance. Furthermore, the use of intensifiers in this online environment contributes to the construction of a reliable online shopping persona, as digital self-expression has become more accessible with advancements in technology and the internet. In that sense, demonstrating authority about a product is a way to exercise one's "entrepreneurial agency," and the use of intensifiers can be seen as a linguistic strategy to influence behavior.

Question 5

May intensifiers be considered a locus for both linguistic innovation and tradition? Which contribution research like your doctoral dissertation may have to the development of language teaching strategies and tools for engaging with language students' (cyber) social lives?

C.L.: In one of my favorite studies on Canadian English intensifiers, Sali Tagliamonte demonstrates that competition, change, and recycling among intensifying forms are the norms in intensification. She argues that intensifiers "cannot have staying power since their impact is only as good as their novelty" (Tagliamonte, 2008, p. 391). I found this to be true in an experimental task designed for my dissertation, wherein newer intensifying forms in Brazilian Portuguese, such as *super* and *totalmente*, were rated as stronger or more intense than more canonical forms like *muito* and *bem* among speakers of all ages. This survey was designed to quantify speakers' perceptions regarding the use of intensifiers in specific scenarios; however, it can only provide limited insights into language. The value of an intensifier, operationalized here as its strength or illocutionary force in discourse and represented by numbers on a scale, is updated in real-time during an interaction. This updating occurs because the meaning of an intensifier is context dependent. This type of research can be beneficial to foreign language teaching as it explores linguistic patterns of authentic language data and promotes the use of language corpora in the classroom as a tool to expand vocabulary acquisition, model different types of texts, and provide cultural context and references, which are key aspects in the development of intercultural competence.

Question 6

To end this interview, what technical, statistical, and technological recommendations to conduct linguistic research would you recommend? Could you suggest any bibliographical references for those who want to know a little more about the subjects addressed here? Would you like to make any final highlights to the readers?

C.L.: Nowadays, there is an incredible array of tools available that can greatly assist researchers in various ways. Personally, I rely heavily on the statistical software program R for almost everything I do, whether it's a new research project or handling data analysis requests in my current job. Learning how to use the command line has also been immensely beneficial, as it has provided me with a deeper understanding of the hierarchy of files and directories on a computer. My best advice is to seek out a community; both R-Ladies and The Carpentries have played crucial roles in my technical growth, and their materials are readily available online.

In terms of linguistic research, being part of a research group, or collaborating on a publication with your advisor (if you're a graduate student), can be a fantastic

way to get started. To have a trusted, more senior researcher to guide you in using these tools and accurately discussing results, while also pointing out biases, issues, and shortcomings, is essential.

Below, I added a reference list that provides examples of insightful research that combines linguistic theories with a digital element, whether in their methods and data or in their discussions about online communication.

References

AFONSO L.; GARCIA, G.; PAPA, J. Fakerecogna: A new brazilian corpus for fake news detection. *In: International Conference on Computational Processing of the Portuguese Language.* Cham: Springer International Publishing, 2022. p. 57-67.

ANDROUTSOPOULOS, J. Introduction: Sociolinguistics and computer-mediated communication. *Journal of sociolinguistics*, v. 10, n .4, p. 419-438, 2006.

ANDROUTSOPOULOS, J. Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism*, v. 19, n. 2, p. 185-205, 2015.

CARVALHO, A.; PICORAL, A. The acquisition of preposition+ article contractions in L3 Portuguese among different L1-speaking learners: A variationist approach. *Languages*, v. 5, n. 4, e-45, 2020.

CORETTA, S. *et al.* Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, v. 6, n. 3, e-25152459231162567, 2023.

FLANNERY, M. Aggression as impoliteness in a Facebook discussion about class discrimination in a Brazilian university. *Revista de Estudos da Linguagem*, v. 29, n. 1, p. 363-386, 2021.

FUCHS, R. Do women (still) use more intensifiers than men? Recent change in the sociolinguistics of intensifiers in British English. *International Journal of Corpus Linguistics,* v. 22, n. 3, p. 345-374, 2017.

HOFF, M. Cerca mío/a or cerca de mí? A variationist analysis of Spanish locative+ possessive on Twitter. *Studies in Hispanic and Lusophone linguistics*, v. 13, n. 1, p. 51-78, 2020.

HOWE, C. Employing Corpora in Usage-Based Linguistics. *The Handbook of Usage-Based Linguistics*, p. 247-268, 2023.

HUANG, Y. *et al.* Understanding US regional linguistic variation with Twitter data analysis. *Computers, environment and urban systems*, v. 59, p. 244-255, 2016.

GUO, D.; JACK, G.; NINI, A. Mapping lexical innovation on American social media. *Journal of English Linguistics*, v. 46, n. 4, p. 293-319, 2018.

JACK, G.; WOODFIELD, H. *The language of fake news*. Cambridge: Cambridge University Press, 2023.

JENSET, G.; McGILLIVRAY, B. *Quantitative historical linguistics:* A corpus framework. United Kingdom: Oxford University Press, 2017. 26 v.

KROEKER, B. *et a*l. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, v. 56, n. 1, p. 1-18, 2018. Disponível em: https://doi.org/10.1515/ling-2017-0032. Acesso em: 18 abr. 2024.

LÍVIO, C. *Digital Approaches to Intensification in Portuguese and Spanish*. Georgia: University of Georgia, 2023.

MACHADO VIEIRA, M. dos S.; MEIRELES, V. Variação em línguas românicas: ações do projeto VariaR como contributos de ciência aberta e cidadã. *Reflexos*, v. 6, 2023. Disponível em: https://interfas.univ-tlse2.fr/reflexos/1325. Acesso em: 18 abr. 2024.

McCULLOCH, G. *Because internet:* Understanding the new rules of language. United Kingdom: Penguin, 2020.

MONTEIRO, R. *et al.* Contributions to the study of fake news in portuguese: New corpus and automatic detection results. *In: Computational Processing of the Portuguese Language:* 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13. New York: Springer International Publishing, 2018.

MOTA, N.; VIEIRA, M. A construção de intensificação com lexemas de cor no português brasileiro. *Linguística*, v. 16, n. 2, p. 50-68, 2020. Disponível em: https://doi.org/10.31513/linguistica.2020.v16n2a33904. Acesso em: 18 abr. 2024.

NAPOLI, M.; RAVETTO, M. (ed.). *Exploring intensification:* synchronic, diachronic and cross-linguistic perspectives. Netherlands: John Benjamins Publishing Company, 2017. 189 v.

NGUYEN, D. *et al.* How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, v. 3, e-62, 2020.

NGUYEN, D.; NOAH, A.; ROSÉ, C. Author age prediction from text using linear regression. *Proceedings of the 5th ACL workshop on language technology for cultural heritage, social sciences, and humanities, LATECH@ ACL 2011, 24 June, 2011, Portland, Oregon, USA*. United States of America: Association for Computational Linguistics, 2011.

OUSHIRO, L. A computational approach for modeling the indexical field/Uma abordagem computacional para a modelagem de campos indexicais. *Revista de Estudos da Linguagem*, v. 27, n. 4, p. 1737-1786, 2019a.

OUSHIRO, L. Linguistic uniformity in the speech of Brazilian internal migrants in a dialect contact situation. *In: Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Campinas: University of Campinas, 2019b. p. 686-690.

PICORAL, A.; REPPEN, R.; SHELLEY, S. Automated annotation of learner English: An evaluation of software tools. *International Journal of Learner Corpus Research*, v. 7, n .1, p. 17-52, 2021.

RENWICK, M. *et al.* Boomer Peak or Gen X Cliff? From SVS to LBMS in Georgia English. *Language Variation and Change*, v. 35, n. 2, p. 175-197, 2023.

RICHTER, M.; SCHEFFLER, T.; VAN HOUT, R. Tracing and classifying German intensifiers via information theory. *Language Sciences*, v. 96, e-101535, 2023.

RUBOW, L,; SCHOFIELD, B.; SHEN, R. *Understanding Open Access:* When, Why & How to Make Your Work Openly Accessible. Berkeley: Authors Alliance, 2016.

SALTON, G.; WONG, A.; YANG, C. A vector space model for automatic indexing. *Communications of the ACM*, v. 18, n. 11, p. 613-620, 1975.

SCHEFFLER, T. *et al.* The processing of emoji-word substitutions: A self-paced-reading study. *Computers in Human Behavior*, v. 127, e-107076, 2022.

SCHWEINBERGER, M. Using intensifier-adjective collocations to investigate mechanisms of change. *Variation in Time and Space*, p. 231-256, 2020.

SCHWENTER, S. A., Miranda, L., Pérez, I., & Cataloni, V. (2024). Reanalyzing Variable Agreement with tu Using an Online Megacorpus of Brazilian Portuguese. Languages, 9(6), 197. https://doi.org/10.3390/languages9060197

SILGE, J.; ROBINSON, D. *Text mining with R:* A tidy approach. Sebastopol: O'Reilly Media, 2017.

SOMMER, B. *et al.* A multilingual learner corpus for less commonly taught languages. *International Journal of Learner Corpus Research*, v. 8, n. 2, p. 261-282, 2022.

TAGLIAMONTE, A. So different and pretty cool! Recycling intensifiers in Toronto, Canada1. *English Language & Linguistics*, v. 12, n. 2, p. 361-394, 2008.

VAUGHN, C.; KENDALL, T.; GUNTER, K. Probing the social meaning of english adjective intensifiers as a class lab project. *American Speech: A Quarterly of Linguistic Usage*, v. 93, n. 2, p. 298-311, 2018.

WENINGER, C.; LI, D. Performing microcelebrity: Analyzing Papi Jiang's online persona through stance and style. *Language in Society*, v. 52, n. 2, p. 263-294, 2023.