LINGUASAGEM

A FORMAÇÃO EM LINGUÍSTICA COMPUTACIONAL[1]

Entrevista com Cláudia Freitas[2]

RESUMO
Nesta entrevista, em virtude do lançamento de seu livro *Linguística Computacional*, em novembro de 2022, pela Editora Parábola, a Professora nos apresenta sua inserção no ramo da Linguística Computacional, assim como discute o futuro da área e de seus profissionais frente aos rápidos avanços da Inteligência Artificial.

PALAVRAS-CHAVE: Linguística Computacional; Inteligência Artificial; Divulgação Científica.

ABSTRACT
In this interview, due to the launch of her book *Linguística Computacional*, in November 2022, by Editora Parábola, the Professor introduces us to her insertion in the field of Computational Linguistics, as well as discusses the future of the area and its professionals in the face of the rapid advances in Artificial Intelligence.

KEYWORDS: Computational Linguistics; Artificial Intelligence; Scientific Dissemination.

Formação e Pesquisa em Linguística Computacional

Entrevistadores(as): Inicialmente, poderia nos relatar sobre sua trajetória na área da Linguística Computacional?

Cláudia Freitas: Eu sou uma linguista computacional cuja trajetória nesta área não foi linear. Na graduação, fui apresentada à área de Processamento de Línguas Naturais (PLN)

[1] Entrevista concedida no dia 09 de janeiro de 2023, de forma remota, como atividade das disciplinas Laboratório 6 e 7 da *Ênfase II - Textos: Meios e Materiais Instrucionais*. A equipe responsável pela produção, transcrição, retextualização e revisão desta entrevista foi composta por Beatriz Habara Morgon, Bruna Roje Sanches, Clarissa Lenina Scandarolli, Helena Bonuccelli Stefani, João Pedro Gonçalves Munhoz e Stefanie Alves dos Santos, discentes do curso de Bacharelado em Linguística, e Luzmara Curcino, docente no Departamento de Letras e no Programa de Pós-graduação em Linguística da Universidade Federal de São Carlos (DL/PPGL/UFSCar).

[2] Graduada, mestre e doutora em Letras pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), instituição da qual é docente no Departamento de Letras. Coordena o Programa de Pós-Graduação em Estudos da Linguagem desde 2007, além de ser colaboradora da Linguateca desde 2007 (Fonte: Plataforma Lattes). E-mail: claudiafreitas@usp.br.

na Iniciação Científica. Trabalhei com a interface entre Psicolinguística e Sintaxe no Mestrado e como pesquisadora em Educação. No doutorado, retornei ao campo de estudos de PLN com ênfase em Semântica. Ao final do curso, pude fazer parte da equipe de organização do evento *Processing of the Portuguese Language* (PROPOR)[3] realizado no Rio de Janeiro (2006). Com minha atuação neste evento, tive a oportunidade de conhecer a professora Diana Santos[4], uma das conferencistas e uma figura de relevo do projeto *Linguateca*[5]. Por seu intermédio, fui apresentada ao projeto e pude fazer parte como pesquisadora da equipe multidisciplinar responsável pelas atividades da *Linguateca* ao longo de dois anos. Participar deste projeto foi, para mim, uma grande escola. Ao cuidar de aspectos de infraestrutura, tais como avaliações e recursos para PLN, entre as atividades que eu exercia no âmbito deste projeto, descobri o que eu gostava de fazer.

Entrevistadores(as): Quais eram as atividades que você exercia na *Linguateca*?

Cláudia Freitas: Na *Linguateca*, trabalhei com esses procedimentos de infraestrutura a que fiz menção. A avaliação, por exemplo, é um procedimento típico de uma área aplicada. É preciso avaliar a qualidade daquilo que foi feito, ou seja, pensar em critérios e formas de avaliação, entre outros. No que diz respeito aos recursos, por exemplo, de avaliação, sua produção é bastante trabalhosa. Eles são elaborados como se fossem um gabarito complexo, como um esquema de previsões a serem operadas pela máquina. É como se a máquina se perguntasse *Como é que os humanos avaliariam esse texto?* e ela mesma respondesse: *isso é um local, isso é uma organização, isso é uma pessoa*. Então, constrói-se um recurso de avaliação que nesse caso é chamado de *corpus* padrão *ouro* ou *gold* e que será utilizado para avaliar se e como um sistema funcionou.

Entrevistadores(as): Então sua experiência prática neste projeto foi decisiva no seu percurso de início um tanto errático?

[3] O evento PROPOR (*International Conference on Computational Processing of Portuguese Language*) é realizado desde 1993, sendo a principal conferência na área de Processamento Computacional do Português, destinada à apresentação de resultados de pesquisas acadêmicas e tecnológicas e à integração dos grupos de pesquisa da área. Atualmente, é realizado ora no Brasil, ora em Portugal.

[4] Profa. Dra. Diana Santos é uma das responsáveis por lançar as bases da Linguateca. Trabalhou e trabalha em áreas como: tradução automática, processamento de corpus, serviço de web, resposta automática a perguntas etc.

[5] Linguateca é um centro de recursos distribuído para o processamento da língua portuguesa, financiado pelo governo português desde maio de 2000, que foi criado após o projeto preparatório intitulado *Processamento computacional do português* (1998-2000).

Cláudia Freitas: De certa forma, este meu percurso relativamente variado me permitiu uma melhor atuação na *Linguateca* e minha participação neste projeto foi fundamental no meu percurso posterior na área de PLN e na formação de novos jovens linguistas.

O percurso para ingresso em PLN, para quem é oriundo dos cursos de Letras, como eu, ou mesmo de Linguística, não é muito fácil, tendo em vista as condições de inserção na área, se comparadas com as de outros pesquisadores, como as dos graduandos de Computação. Estes normalmente travam contato com a metalinguagem gramatical na escola, enquanto os estudantes de Letras ou Linguística não necessariamente tiveram algum contato com a linguagem computacional. Portanto, as condições de inserção destes diferentes estudantes na área são bastante assimétricas. Um exemplo claro dessa assimetria é de que eu, mesmo com a minha formação atual, ao ler um livro de Computação, não consigo compreender prontamente a linguagem ali mobilizada. O mesmo não se pode dizer de alguém com formação em Computação, que tem maior possibilidade de compreensão da metalinguagem gramatical da qual ele já ouviu falar na escola. Todos eles, por menos afeitos que sejam à gramática, já ouviram falar de objeto direto, indireto, sintagma nominal, ou seja, já conhecem parte da terminologia linguística desde o ensino básico.

Pesquisas em Processamento de Linguagem Natural

Entrevistadores(as): Você poderia nos falar mais sobre o projeto *Linguateca*?

Cláudia Freitas: A *Linguateca*, assim como o Núcleo Interinstitucional de Linguística Computacional (NILC)[6], é um espaço fundamental para o desenvolvimento do PLN em Língua Portuguesa. Em 1998, quando quase não se falava deste tema na área, ela foi criada por uma equipe de pesquisadores com a intenção de promover o estudo de processamento computacional da língua portuguesa com os mesmos recursos e ferramentas que o estudo e a descrição de outras línguas já dispunham. O objetivo era não deixar a descrição e análise da Língua Portuguesa em desvantagem nas pesquisas da área por falta de recursos de PLN.

[6] NILC - Núcleo Interinstitucional de Linguística Computacional, criado em 1993 para fomentar projetos de pesquisa e desenvolvimento em Linguística Computacional e Processamento de Linguagem Natural, contando com cientistas da computação, linguistas e bolsistas de diferentes universidades e centros de pesquisa, como a Universidade de São Paulo (USP) e a Universidade Federal de São Carlos (UFSCar).

A criação da *Linguateca* foi uma ideia encampada pelo Ministério da Ciência e Tecnologia de Portugal, que financiou este projeto internacional e interinstitucional durante muitos anos, e que foi conduzido por pesquisadores de Portugal e do Brasil de diferentes universidades. Outra característica fundamental do projeto é a disponibilização de todo o seu acervo de dados de forma totalmente pública, acessível e gratuita, o que é decisivo para o avanço dos estudos na área.

O objetivo dos pesquisadores envolvidos é produzir recursos e avaliações para o processamento de Língua Portuguesa, sendo útil tanto para a comunidade de linguistas que trabalham com *corpus* e com PLN, quanto para comunidade de Computação que trabalha com PLN e, eventualmente, para a academia e a indústria. Atualmente, a *Linguateca* dispõe de recursos diversos e de relevo, em paralelo com o NILC no Brasil, responsáveis por avanços significativos para a história da Linguística Computacional em Língua Portuguesa.

Entrevistadores(as): Até agora, utilizamos bastante as terminologias *PLN* e *Linguística Computacional*. Há alguma diferença entre essas designações? E como você definiria esses estudos?

Cláudia Freitas: Diferentemente de outros colegas, eu não costumo fazer essa distinção, uma vez que tanto a Linguística Computacional quanto o PLN se ocupam do processamento de línguas naturais.

Quando tenho de explicar o que fazemos nessa área, começo dizendo que, apesar de toda a sua produção teórica, a Linguística Computacional ou PLN é uma área prioritariamente aplicada cujos pesquisadores se valem de computadores com a finalidade de realizar tarefas de linguagem. O grande diferencial na Linguística Computacional ou PLN é ter computadores resolvendo demandas que têm, especificamente, a linguagem como objeto.

Entrevistadores(as): Considerando que a Linguística Computacional é uma área prioritariamente aplicada, em que consistiria essa sua especificidade?

Cláudia Freitas: Ao me referir a esse traço da área, o de ser uma abordagem *aplicada*, é preciso cautela no uso desse termo, uma vez que se poderia entender equivocadamente que a nossa tarefa é uma mera aplicação daquilo que pessoas "mais inteligentes" fizeram

ou pensaram em outras áreas. Pelo contrário. Em qualquer área aplicada, a aplicação produz conhecimento enquanto aplicação, não sendo, portanto, apenas uma implementação de um conhecimento que foi produzido antes e por outros pesquisadores de outros campos. O próprio ato de aplicar demanda reflexão e contemplação de alternativas e soluções para um problema, o que por si só é produção de conhecimento.

As contribuições da área de PLN, dado o seu caráter prioritariamente aplicado, são inúmeras e têm variado com grande velocidade em decorrência das demandas do mundo atual. Entre as funcionalidades produzidas e disponibilizadas pela área encontram-se a produção automatizada de resumos ou sumarização de textos, a correção ortográfica automática, a criação de ferramentas de auxílio à escrita ou de tradução automática de textos e também procedimentos de extração de informação. Penso nesta última de forma mais ampla, porque essa funcionalidade pode envolver desde a identificação de informações até a de posições ideológicas distintas inscritas nos textos. Fazem parte ainda do rol dessas funcionalidades produzidas pela área, os programas de reconhecimento de voz e os agentes conversacionais, ou seja, *chatbots*.

Sendo assim, qualquer ação que fazemos com a linguagem e que podemos contar com a ajuda do computador interessa aos pesquisadores da área de PLN. Um ótimo exemplo disso é o *ChatGPT*, ainda que esse sistema, no estado atual, não seja tão assertivo.

Linguística, Matemática e Computação em intersecção

Entrevistadores(as): O aprendizado de máquina é uma das ações fundamentais do PLN hoje. O *ChatGPT* é uma aplicação bem concreta dessa contribuição da área de PLN. Ele é construído sobretudo por métodos baseados numa lógica matemático-computacional do que por conhecimento linguístico? Poderíamos falar em termos de competição entre as áreas?

Cláudia Freitas: Eu não acredito que devamos falar em termos de uma competição, mas antes de uma cooperação. Existem recursos de aprendizado de máquina que não envolvem o conhecimento linguístico específico, assim como existem muitos outros que demandam esse conhecimento especializado.

A anotação a partir da metalinguagem linguística, por exemplo, permite-nos organizar quaisquer dados linguísticos. Quanto mais informação linguística um texto possui - mesmo que não seja nosso objetivo final -, mais eficazmente conseguimos

produzir materiais. Afinal, nós linguistas sabemos como lidar com esse tipo de informação, porque está relacionada à linguagem. A presença de informações linguísticas em materiais como *datasets* ou materiais anotados permite que possamos extrair outros dados linguísticos de natureza distinta.

Por exemplo, se meu objetivo é anotar relações retóricas, fazer uso de certas informações sintáticas me ajuda a identificar mais facilmente os parágrafos que exemplificam ou os parágrafos que apresentam índices de contradição. É razoável imaginarmos que alguém que passou por uma formação de anos em um curso de Letras ou Linguística saiba lidar com dados de natureza linguística. O papel de linguistas de PLN em processo de aprendizagem de máquina será, todo ele, voltado para contribuições de ordem linguística, isso porque, eventualmente, será preciso fazer anotações de informações linguísticas clássicas como classe, função, formação retórica etc, mas também porque será preciso pensar outros tipos de anotação para solucionar novas tarefas que possam surgir. Além disso, a mobilização de metalinguagem linguística pode agilizar não somente o nosso trabalho de análise, mas também o processo de aprendizado da máquina, a partir de um *corpus* anotado por um linguista.

Existem determinadas tarefas que a máquina vai desempenhar e que nos pouparão alguns esforços e funcionarão como atalho, permitindo-nos criar e inovar. Devemos valorizar não apenas nossa competência técnica como linguistas, mas também nossa capacidade crítica, analítica e de fazer boas perguntas, competências que nenhuma máquina dispõe.

Podem as máquinas pensar e falar?

Entrevistadores(as): Em um dos capítulos de seu livro publicado em 2022, *Linguística Computacional*, pela editora Parábola, há uma discussão relacionada ao Empirismo e ao Racionalismo, e relacionado à polêmica acerca da capacidade de *pensar* das máquinas. Essa polêmica fez com que Turing[7] evitasse o termo quando se referia às máquinas, enquanto Searle[8] trouxe novamente essa ideia. Qual sua posição nessa questão?

[7] Matemático, lógico, criptoanalista e cientista da computação britânico. Turing é muitas vezes considerado o pai da ciência da computação moderna. Durante a Segunda Guerra Mundial, Turing foi chefe da Hut 8, a seção responsável pela criptoanálise naval alemã. (*Cf*. Turing, 1954).

[8] Filósofo analítico e escritor norte-americano, dedicado à análise dos atos da fala, relacionados à consciência da realidade social e institucional, da racionalidade, da intencionalidade individual e coletiva nos usos da linguagem.

Cláudia Freitas: Foi prazeroso escrever esse capítulo. A ciência não tem pontas soltas. Elas sempre têm uma história, uma origem e visam responder determinadas perguntas. Sendo o conhecimento uma produção humana, em todo o livro, procuro abordar essa dimensão fundamental, a da historicidade do conhecimento. Na linguística, por exemplo, sabemos que as ideias têm um ponto de partida histórico, embora dificilmente nos lembremos de situá-las nestes termos e como um produto humano.

Neste caso em específico, de um lado, encontramos Turing que evita empregar o termo *compreensão* quando se refere à aprendizagem de máquina. Para ele, as máquinas não *compreendem*, elas decodificam. Do outro lado, temos Searle, que emprega especificamente este termo. Eu tendo a me situar entre os dois a esse respeito. Ter uma visão menos dogmática permite-me evitar uma definição presunçosa de um padrão de funcionamento tanto da língua como do PLN. Para nós que trabalhamos com a linguagem, somos constantemente expostos às instabilidades, o que nos exige alguma flexibilidade.

Toda essa discussão resume-se ao modo como definimos *compreensão*. Afinal, estamos tratando de palavras. Wittgenstein, um reconhecido filósofo dedicado a temas de interesse linguístico, que mudou de opinião com relação a seus estudos da linguagem ao longo de sua vida, afirmou em um de seus aforismos que *Compreende uma ordem aquele que age de acordo com ela*[9]. Assim, como podemos saber se alguém compreendeu algo? É possível ter certeza disso quando esse alguém está fazendo exatamente o que lhe foi pedido.

Quando vamos para o mundo das máquinas, tudo depende de como definimos *compreensão*, porque muitas vezes há uma definição já pronta de que *só humanos compreendem*. Isso nos impede de aceitar que uma máquina pode *compreender*. No entanto, se ela produz um resultado a partir de um dado comando ou de uma dada programação, logo ela *compreende*.

Por exemplo, quando nós, professores, elaboramos uma prova e pedimos para os alunos responderem, quem pode garantir que o aluno que respondeu certo uma questão compreendeu bem a matéria? Talvez, ele conheça apenas aquilo que o professor gostaria de ler como resposta ou tenha decorado as respostas do livro ou simplesmente "chutado" e acertado a questão. É difícil comprovar isso. Por outro lado, poderíamos dizer que um aluno que não acertou uma questão não a compreendeu? Ele deixa de ser uma pessoa porque não forneceu uma resposta correta?

[9] *Cf.* Wittgenstein (1953).

Portanto, essa discussão acerca do termo *compreender*, quando relacionado a humanos e máquinas, está muito alinhada à mera demarcação de territórios. Máquinas compreendem? Precisamos entender primeiro o que chamamos de *compreensão*.

Formação em PLN, trajetos e desafios

Entrevistadores(as): Qual é o seu conselho para um estudante que queira se inserir na área de PLN. Existem outros caminhos além do acadêmico? Como funciona o mercado de trabalho?

Cláudia Freitas: Eu não tenho experiência de atuação no mercado, de modo geral. Sempre atuei na área acadêmica com pesquisa. Trabalhei ativamente no projeto da *Linguateca*, e, mesmo nele, meu contrato era de investigadora, como eles dizem em Portugal, o que equivale a pesquisadora no Brasil. Na academia, na universidade, muitos projetos de PLN têm algum vínculo com a indústria. Desta forma, o trabalho de pesquisa que realizamos nesta área contempla os interesses da indústria e do mercado.

Meu conselho é: estudem! Tentem se engajar o mais cedo possível em projetos interdisciplinares, como o da *Linguateca*, nos quais se possa *colocar a mão na massa*, dada a tradição aplicada da área de PLN. Procurem professores que atuem na área e se mostrem dispostos a trabalhar em seus projetos. Nesse ramo há demanda, mas ainda há carência de profissionais, sejam eles linguistas de formação ou da área de computação.

Outra recomendação, que eu pude abordar detidamente no meu livro, é a da importância que nossa área de PLN adota de olhar a língua não apenas do ponto de vista das teorias linguísticas, mas também do ponto de vista da computação. Eu me dediquei no livro a fomentar a importância do diálogo simétrico entre aqueles que se propõem a fazer PLN e vêm da computação e nós da linguística, com o objetivo comum de visualizar e enfrentar os desafios da área, de entender quais são os problemas linguísticos que nós precisaríamos resolver atualmente.

Cada vez mais, a tendência será essa: a de dispor de uma formação mais completa e complexa, ou seja, holística, na qual se tenha pessoas da computação tendo formação em linguística e pessoas da linguística tendo formação computacional.

Por fim, quanto mais cedo buscarem se inserir em projetos da área de PLN, interdisciplinares com orientação aplicada a problemas de linguagem contemporâneos, melhor. Caso tenham formação inicial na área de Letras ou Linguística, devem se

matricular em cursos relacionados ao PLN, como por exemplo, um curso básico de Programação. Caso tenham formação inicial na área de Computação, devem estudar a linguagem, de um ponto de vista científico, para compreender a diferença entre os dados linguísticos e os outros tipos de dados.

Será um grande aprendizado para os profissionais em formação de ambas as áreas esse precoce e contínuo diálogo teórico-aplicado. E como última recomendação, é muito importante, e isso para a formação de todo e qualquer cientista, não ter medo de fazer perguntas e não ter medo de procurar as respostas!

REFERÊNCIAS

FREITAS, Cláudia. **Linguística Computacional**. 1ª edição. São Paulo: Editora Parábola. 2022.

LINGUATECA. **Linguateca**, 2015. Disponível em: https://www.linguateca.pt/. Acesso em: 07 ago. 2023.

NILC. **Interinstitutional Center for Computational Linguistics**. Disponível em: https://sites.google.com/view/nilc-usp/. Acesso em: 07 ago. 2023.

PROPOR INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE LANGUAGE. **Comissão Especial de Processamento de Linguagem Natural**, 2023. Disponível em: https://sites.google.com/view/ce-pln/eventos/propor. Acesso em: 07 ago. 2023.

SANTOS, Diana. **Linguateca,** 2021. Disponível em: https://www.linguateca.pt/Diana/. Acesso em: 07 ago. 2023.

TURING, Alan Mathison. **Solvable and Unsolvable Problems**. Science News 31, 1954.

WITTGENSTEIN, Ludwig. **Philosophische Untersuchungen**. 1. ed. Oxford: Editora Basil Blackwell, 1953.

Como referenciar esta entrevista:

FREITAS, Cláudia. A formação em Linguística Computacional. [Entrevista concedida a] Beatriz Habara Morgon, Bruna Roje Sanches, Clarissa Lenina Scandarolli, Helena Bonuccelli Stefani, João Pedro Gonçalves Munhoz, Luzmara Curcino e Stefanie Alves dos Santos. **revista Linguasagem**, São Carlos, v.47, n.1, p. 14-22, 2024.