Anais

ELC 2026

EBRALC

ANAIS DO XVII ENCONTRO DE LINGUÍSTICA DE CORPUS E DA XIV ESCOLA BRASILEIRA DE LINGUÍSTICA COMPUTACIONAL

Salvador, de 09 a 12 de novembro de 2026

ORGANIZAÇÃO E APOIO

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Anais do XVII Encontro de Linguística de Corpus e
da XIV Escola Brasileira de Linguística
Computacional [livro eletrônico] / [organização
Jackson Wilke da Cruz Souza, Roana Rodrigues,
Lavínia Karolayne dos Santos]. -- Salvador, BA :
Ed. dos Autores, 2026.
eBook

Vários autores.
Vários colaboradores.
Bibliografia.
ISBN 978-65-02-50093-4

1. Inteligência artificial 2. Linguagem
3. Linguística 4. Linguística de corpus 5. Tecnologia
I. Souza, Jackson Wilke da Cruz. II. Rodrigues,
Roana. III. Santos, Lavínia Karolayne dos.

26-415172.0 CDD-410.285

Índices para catálogo sistemático:

1. Linguística de corpus : Linguística 410.285

Camila Aparecida Rodrigues – Bibliotecária CRB –
SP-010133/O

COMISSÃO ORGANIZADORA

Jackson Wilke da Cruz Souza (UFBA)
Elisa Duarte (UnB)
Gabriela Fortalesa (UEFS)
Guilherme de Araújo (UFBA)
Henrique A. Moraes Lima (UFMA)
João Cyrino (UFBA)
Julia Raiza (UFBA)
Lavínia Karolayne Santos (UFS)
Lilian Teixeira (UFBA)
Lucas Antonio (UFBA)
Lucas dos Santos (UFBA)
Mateus Pereira (UFBA)
Roana Rodrigues (UFS)
Roney Santos (UFBA)
Silvia Melo (UFBA)
Theciana Silveira (UFMA)

COMISSÃO CIENTÍFICA

Adriane Orenha Ottaiano (UNESP)
Angela Maria Tenório Zucchi (USP)
Annallena de Souza Guedes (IFBA)
Bianca Pasqualini (UFRGS)
Bruna Rodrigues da Silva (UFRGS)
Cristiane Krause Kilian (Faculdade Instituto Ivoti)
Deise Amaral (UFRGS)
Deise Prina Dutra (UFMG)
Diego Alves (Saarland University)
Eleonora Bambozzi Bottura (UFSCar)
Elisa Duarte Teixeira (UnB)
Gabriela Wick Pedro (Atento)
Gleiton Malta Magalhaes (UFBA)
Graziele Altino Frangiotti (UFSC)
Guilherme Fromm (UFU)
Jean Michel Pimentel Rocha (IFMS)
Jorge Baptista (Universidade do Algarve)
Josefa Mônica Almeida Silva Alves (UFS)
Luciana Carvalho Fonseca (USP)
Magali Sanches Duran (USP)

Malila Carvalho de Almeida Prado (Beijing Normal-Hong Kong Baptist University)
Marcos de Campos Carneiro (UnB)
Maria José Bocorny Finatto (UFRGS)
Marine Laísa Matte (IFSul)
Maucha Andrade Gamonal (UNICAMP)
Oto Vale (UFSCar)
Paula Tavares Pinto (UNESP)
Rafaela Araújo J. R. Peixoto (Universidade da Força Aérea)
Ryan Marçal Saldanha Magaña Martinez (UFRJ)
Roney Lira de Sales Santos (UFBA)
Rozane Rodrigues Rebechi (UFRGS)
Shirlene Bemfica de Oliveira (IFMG)
Simone Sarmento (UFRGS)
Stella E. O. Tagnin (USP)
Theciana Silva Silveira (UFMA)
Thiago Alexandre Salgueiro Pardo (USP)

ORGANIZAÇÃO DO VOLUME

Jackson Wick da Cruz Souza (UFBA)
Roana Rodrigues (UFS)
Lavínia Karolayne Santos (UFS)

CAPA

Gabriela Fortalesa
Lucas Antonio dos Santos

APRESENTAÇÃO

A **Linguística de Corpus (LC)** tem se consolidado, ao longo das últimas décadas, como um campo de investigação fundamental para o estudo da linguagem, contribuindo para a descrição e a análise de fenômenos linguísticos a partir de dados empíricos organizados em *corpora*. Desde os primeiros trabalhos desenvolvidos com o auxílio de computadores, na segunda metade do século XX, as abordagens teóricas e metodológicas da área têm acompanhado as transformações tecnológicas que foram acontecendo e, como resultado, ampliamos nosso escopo de investigação e robustez de análise de dados linguísticos. Nesse sentido, também presenciamos o fortalecimento de interfaces com outras áreas do conhecimento, especialmente com o Processamento de Linguagem Natural (PLN), área que, embora possua trajetória e objetivos próprios, compartilha o interesse pela representação, pelo processamento e pela descrição de aspectos próprios da linguagem humana e autêntica.

Como nada que conhecemos não conseguiu não ser atravessado pelas transformações provocadas pela Inteligência Artificial, particularmente com a difusão dos *Large Language Models* (LLMs), a LC também agregou para si novos desafios e contornos metodológicos com essas interfaces. A disponibilidade de modelos capazes de produzir, transformar, classificar e analisar textos em diferentes línguas, linguagens e contextos tem ampliado as possibilidades de pesquisa. Porém, ao mesmo tempo, suscitam-se questões sobre a natureza dos dados utilizados, a representação do conhecimento linguístico, a diversidade das línguas contempladas e os limites das análises produzidas automaticamente. Nesse cenário, a LC mantém posição especialmente relevante e privilegiada: os *corpora* não são apenas fontes de dados para o treinamento e a avaliação de sistemas computacionais, mas também espaços de investigação, descrição e sistematização de conhecimentos sobre o funcionamento das línguas. Assim, a aproximação entre LC, PLN e IA deve ser encarada como um convite à reflexão sobre o que as tecnologias podem realizar com os dados linguísticos, mas também sobre como o conhecimento produzido no âmbito de nossas pesquisas linguísticas pode contribuir para o desenvolvimento e a avaliação crítica dessas tecnologias.

É nesse contexto que organizamos e realizamos, entre os dias 9 e 12 de novembro de 2026, na Universidade Federal da Bahia (UFBA), em Salvador, o XVII Encontro de Linguística de Corpus (ELC) e a XIV Escola Brasileira de Linguística Computacional (EBRALC). Realizado pela primeira vez em 1999, na Universidade de São Paulo, o ELC tem se firmado como um espaço de interlocução científica entre pesquisadores interessados nas diferentes abordagens da LC. Já a EBRALC, criada em 2007, surgiu da necessidade de ampliar as oportunidades de formação em métodos e ferramentas computacionais para pesquisadores da linguagem, especialmente aqueles vinculados às áreas de Letras e Linguística. Ao longo de suas edições, os dois eventos têm contribuído para a formação de pesquisadores, a circulação de conhecimentos e o fortalecimento de uma comunidade científica caracterizada pela diversidade de perspectivas teóricas,

metodológicas e institucionais. Preservando essa trajetória, a edição de 2026 reafirmou o compromisso com o diálogo interdisciplinar e com a aproximação entre investigação linguística e desenvolvimento tecnológico.

A programação desta edição expressou a amplitude dessas interlocuções. As atividades formativas da EBRALC contemplaram desde a exploração e a anotação de *corpora* até a utilização de ferramentas e modelos computacionais em diferentes tarefas de análise linguística. Entre as temáticas abordadas estão a anotação morfossintática de línguas indígenas com base no modelo *Universal Dependencies*, o uso de *corpora* de referência para o ensino de línguas, a introdução aos LLMs, a mineração de textos com ferramentas de análise de dados e a aplicação de recursos de LC e IA à simplificação de bulas de medicamentos.

Já os trabalhos científicos reunidos nestes Anais, apresentados nas modalidades de comunicação oral e pôster, também revelam a diversidade de objetos e problemas de investigação que caracterizam a LC contemporânea, com novo fôlego. As pesquisas abrangem questões relacionadas à escrita e à comunicação acadêmica, à complexidade lexical e gramatical, à tradução e à fraseologia, à terminologia e à acessibilidade textual, à análise do discurso, à fonética e à prosódia, bem como à construção, à preparação e à anotação de corpora. De maneira geral, os trabalhos selecionados demonstram que a incorporação de novas tecnologias não elimina a necessidade de reflexão teórica e metodológica sobre os dados, mas amplia as possibilidades de formular perguntas, explorar fenômenos e produzir conhecimentos sobre a linguagem em suas diferentes manifestações.

Por fim, gostaria de agradecer a todes, todas e todos que contribuíram para a realização da edição ELC & EBRALC 2026. A vocês que emitiram seus pareceres científicos; àqueles que se desdobraram em tantas atividades e contribuíram em atender às (inúmeras) demandas de organização de um evento, a vocês que propuseram minicursos para compartilhar seus conhecimentos; a vocês, colegas, que atenderam nossos pedidos para serem palestrantes; a vocês que sobrepuseram papeis e ajudaram em tantas frentes de trabalho: muito obrigado! Foi maravilhoso trabalhar e contar com cada um de vocês colegas, parceiros, amigos, alunos e orientandos. Agradeço generosamente também pelo fato de termos provocado um deslocamento geográfico de discussões científicas do eixo sul-sudeste para cá, para o nordeste, e olharmos o que produzimos e pensamos a partir desta região tão rica e importante para nosso país.

Espero que estes Anais contribuam para ampliar as interlocuções entre pesquisadores, estimular novas parcerias e reafirmar a importância dos estudos empiricamente fundamentados para a compreensão da linguagem humana e para o desenvolvimento crítico, responsável e socialmente relevante de tecnologias linguísticas.

Prof. Dr. Jackson Wilke da Cruz Souza
Presidente da Comissão Organizadora – UFBA, 2026

PROGRAMAÇÃO

Escola Brasileira de Linguística de Corpus – EBRALC

DIA 1 **09 de novembro de 2026**		
Horário	**Atividade**	**Local**
07:30	Inscrição e retirada de material	Auditório Instituto de Matemática e Estatística – campus Ondina
08:00	*Solenidade de abertura* Prof. Dr. Jackson Wilke da Cruz Souza (UFBA), Profa. Dra. Stella Esther Ortweiler Tagnin (USP), Profa. Dra. Maria Cristina Vieira de Figueiredo Silva (PPGLinC-UFBA), Profa. Dra. Maiana Brito de Matos (ICTI-UFBA), Profa. Dra. Lílian Teixeira de Sousa (IL-UFBA), Profa. Dra. Olívia Maria Cordeiro de Oliveira (PROPICI-UFBA) *Palestra de abertura* "Diversidade e atrito: o que se perde quando tudo converge" (Profa. Dra. Cláudia Freitas – UFBA)	
10:00	*Coffee break*	
10:30	Introdução aos Grandes Modelos de Linguagem: história, funcionamento e desafios da Inteligência Artificial contemporânea (Livy Real, Cibele Simões de Oliveira Santos)	Sala 12 do Instituto de Matemática e Estatística – campus Ondina
	O uso de corpora de referência por profissionais de Letras para a aprendizagem de língua inglesa (Rachel Lourenço, Thiago Blanch Pires)	Smart class III Alpha do Instituto de Computação – campus Ondina
12:30	*Intervalo para almoço*	
14:00 - 16:00	Introdução aos Grandes Modelos de Linguagem: história, funcionamento e desafios da Inteligência Artificial contemporânea (Livy Real, Cibele Simões de Oliveira Santos)	Sala 12 do Instituto de Matemática e Estatística – campus Ondina
	O uso de corpora de referência por profissionais de Letras para a aprendizagem de língua inglesa (Rachel Lourenço, Thiago Blanch Pires)	Smart class III Alpha do Instituto de Computação – campus Ondina
16:00 - 16:30	*Intervalo*	
16:30 - 18:30	Anotação morfossintática de corpora de línguas indígenas: da Plataforma Tycho Brahe ao modelo Universal Dependencies (Leonel F. de Alencar, Maria Filomena S. Sandalo, Charlotte Marie C. Galves)	Smart class III Alpha do Instituto de Computação – campus Ondina
	Como os Grandes Modelos de Linguagem representam a linguagem? Uma introdução aos Transformers para linguistas com Orange Data Mining (Marcos Moisés Crisóstomo de Oliveira)	Smart class III Beta do Instituto de Computação – campus Ondina

DIA 2 **10 de novembro de 2026**		
Horário	**Atividade**	**Local**
08:00 - 10:00	Anotação morfossintática de corpora de línguas indígenas: da Plataforma Tycho Brahe ao modelo Universal Dependencies (Leonel F. de Alencar, Maria Filomena S. Sandalo, Charlotte Marie C. Galves)	Smart class III Alpha do Instituto de Computação – campus Ondina
	Como os Grandes Modelos de Linguagem representam a linguagem? Uma introdução aos Transformers para linguistas com Orange Data Mining (Marcos Moisés Crisóstomo de Oliveira)	Smart class III Beta do Instituto de Computação – campus Ondina
10:00	*Coffee break*	
10:30 - 12:30	Anotação morfossintática de corpora de línguas indígenas: da Plataforma Tycho Brahe ao modelo Universal Dependencies (Leonel F. de Alencar, Maria Filomena S. Sandalo, Charlotte Marie C. Galves)	Smart class III Alpha do Instituto de Computação – campus Ondina
	Como os Grandes Modelos de Linguagem representam a linguagem? Uma introdução aos Transformers para linguistas com Orange Data Mining (Marcos Moisés Crisóstomo de Oliveira)	Smart class III Beta do Instituto de Computação – campus Ondina
12:30	*Intervalo para o almoço*	
14:00 - 16:00	Linguística de Corpus, modelos de linguagem e Inteligência Artificial: técnicas e ferramentas para simplificar bulas de medicamentos em português do Brasil (Maria José Bocorny Finatto, Bianca Franco Pasqualini)	Smart class III Alpha do Instituto de Computação – campus Ondina
	Introdução à Mineração de Texto com Orange Data Mining: fundamentos e aplicações práticas (Crysttian Paixão, Eduardo Grizenti)	Smart class III Beta do Instituto de Computação – campus Ondina
16:00	*Intervalo*	
16:30 - 18:30	Linguística de Corpus, modelos de linguagem e Inteligência Artificial: técnicas e ferramentas para simplificar bulas de medicamentos em português do Brasil (Maria José Bocorny Finatto, Bianca Franco Pasqualini)	Smart class III Alpha do Instituto de Computação – campus Ondina
	Introdução à Mineração de texto com Orange Data Mining: fundamentos e aplicações práticas (Crysttian Paixão, Eduardo Grizenti)	Smart class III Beta do Instituto de Computação – campus Ondina

Encontro de Linguística de Corpus – ELC

DIA 3 **11 de novembro de 2026** **Local:** Auditório Milton Santos A do Instituto de Geociências – campus Ondina	
Horário	**Atividade**
08:30	Inscrição e retirada de material
09:00	*Solenidade de abertura* Prof. Dr. Jackson Wilke da Cruz Souza *Palestra de abertura* "Representação de Conhecimento em IA: delimitando fronteiras, desfazendo mitos" (Prof. Dr. Tiago Timponi Torrent - UFJF)
10:00	*Coffee break*
10:30 - 12:00	**Sessão de Comunicação A** Escrita acadêmica, complexidade linguística e ensino Mediadora: Profa. Dra. Stella Tagnin 1. A complexidade lexical e fraseológica na distinção entre níveis de proficiência em inglês: um estudo baseado em corpus (Deise Amaral) 2. Complexidade gramatical na escrita universitária: nominalizações e adjetivos atributivos em produções de graduandos de Letras Inglês (Marine Laisa Matte e Simone Sarmento) 3. A apresentação da pesquisa nas ciências agrárias: a descrição das introduções de artigos científicos como base para a criação de atividades (Deise Dutra e colaboradores) 4. O corpus C-ORAL-BRASIL II (Bruno Rocha e Tommaso Raso)
12:00	*Intervalo para o almoço*
14:00 - 15:30	**Sessão de Comunicação B** Inteligência artificial, discurso e análise de gêneros Mediador: Prof. Dr. Thiago Pardo 1. Mudanças linguísticas em abstracts acadêmicos após a difusão do ChatGPT: uma análise baseada em corpus (Renato de Vargas Matos e Simone Sarmento) 2. Automatização da Identificação de Funções Retóricas de Pacotes Lexicais em Artigos Científicos das Ciências Agrárias: uma análise exploratória do uso de Large Language Models (LLM) (Karol Traslatti Soares e Ana Eliza Pereira Bocorny) 3. Aplicações iniciais de modelos de língua na anotação Rhetorical Structure Theory em textos jurídicos do português (Winycius Alberto Nascimento do Brasil e Jackson Wilke da Cruz Souza) 4. Uma Metodologia de Extração de Dados Dialetais de Documentos Científicos do Atlas Linguístico do Brasil (Alfredo Victor do N. S. Sena; Amanda Sammer do N. Brandão; Felipe B. S. dos Reis; Laila P. M. Santos; Daniela B. Claro).

	Intervalo
15:30 - 16:30	Apresentação de Pôsteres
16:30	Reunião com a Comunidade e Prestação de contas

DIA 4 **12 de novembro de 2026** **Local:** Auditório Milton Santos A do Instituto de Geociências – campus Ondina	
Horário	**Atividade**
09:00 - 10:00	**Sessão de Comunicação C** Terminologia, acessibilidade e comunicação em saúde Mediadora: Profa. Dra. Ana Eliza Bocorny 1. Ressignificação terminológica e conceptualização infantil da experiência oncológica: um estudo com crianças em tratamento para leucemia linfoblástica aguda (Josefa Mônica Almeida Silva Alves e Soraya Carvalho Souza Biller Teixeira) 2. Acessibilidade textual e terminológica em ambiente digital: análise de corpus e impacto social na educação em saúde (Bruna Rodrigues da Silva) 3. Bulas de medicamentos para o novo CorPop Saúde: curadoria e avaliação linguística de simplificações geradas por modelos de inteligência artificial (Maria José Bocorny Finatto, Arthur Scalercio e Bianca Franco Pasqualini)
10:00	*Coffee break*
10:30 - 12:00	**Sessão de Comunicação D** Corpora anotados, ferramentas e recursos para PLN Mediadora: Profa. Dra. Lilian Teixeira 1. UD_Kadiweu-Unicamp: expandindo a diversidade tipológica da coleção Universal Dependencies com o primeiro treebank de uma língua da família Guaicuru (Leonel F. de Alencar, Maria Filomena S. Sandalo e Charlotte Marie C. Galves) 2. Córpus Alienista: uma obra machadiana anotada com Universal Dependencies (Magali S. Duran, Ursula Puello Sydio, Lucelene Lopes e Thiago A. S. Pardo) 3. TTWtools: uma ferramenta aberta e transversal para etiquetagem e análise em linguística de corpus (Luciana Aguiar de Oliveira e José Henrique Padovani) 4. FREIRE: Um Motor Computacional Adaptativo e Sensível a Logica para Alfabetização de Adultos em Comunidades Tradicionais (Lucas Parreias)
12:00	*Intervalo para o almoço*
14:00 - 15:30	**Sessão de Comunicação E** Tradução, fraseologia e estudos multilíngues Mediadora: Profa. Maria José Finatto

	1. Ampliação e proposta de sistema de remissivas para um glossário multilíngue direcionado por corpus (Maria Eduarda Magalhães Furtado Dutra e Elisa Duarte Teixeira) 2. Compilação e anotação de um corpus paralelo inglês–português de legendas para identificação de expressões idiomáticas (Rozane R. Rebechi, Diego Alves, Isabela Moreira de Oliveira e Christopher dos Santos Filho) 3. E se Pellegrino Artusi fosse blogueiro? Tradução funcional de uma receita histórica na era da IA generativa (Stella E. O. Tagnin e Rozane R. Rebechi) 4. Ecolinguística: uso do IRaMuTeQ na análise do nódulo "animal" em corpus das Ciências Agrárias (Shirlene Bemfica de Oliveira e colaboradoras)
15:30	*Intervalo*
16:00 - 17:30	*Mesa-redonda de encerramento* "Linguística de Corpus, discurso e transformação digital" Profa. Dra. Livy Real (UFAM), Prof. Dr. Oto Araujo Vale (UFSCar) e Profa. Dra. Raquel Meister Ko Freitag (UFS)

Sumário

TRABALHOS COMPLETOS ELC 2026 17

A complexidade lexical e fraseológica na distinção entre níveis de proficiência em inglês: um estudo baseado em corpus 18
Deise Amaral

Complexidade gramatical na escrita universitária: nominalizações e adjetivos atributivos em produções de graduandos de Letras Inglês 27
Marine Laisa Matte; Simone Sarmento

A apresentação da pesquisa nas ciências agrárias: a descrição das introduções de artigos científicos como base para a criação de atividades 36
Deise Dutra e colaboradores

O corpus C-ORAL-BRASIL II 46
Bruno Rocha; Tommaso Raso

Mudanças linguísticas em abstracts acadêmicos após a difusão do ChatGPT: uma análise baseada em corpus 56
Renato de Vargas Matos; Simone Sarmento

Automatização da Identificação de Funções Retóricas de Pacotes Lexicais em Artigos Científicos das Ciências Agrárias: uma análise exploratória do uso de Large Language Models (LLM) 63
Karol Traslatti Soares; Ana Eliza Pereira Bocorny

Aplicações iniciais de modelos de língua na anotação *Rhetorical Structure Theory* em textos jurídicos do português 69
Winycius Alberto Nascimento do Brasil; Jackson Wilke da Cruz Souza

Uma Metodologia de Extração de Dados Dialetais de Documentos Científicos do Atlas Linguístico do Brasil 79
Alfredo Victor do N. S. Sena; Amanda Sammer do N. Brandão; Felipe B. S. dos Reis; Laila P. M. Santos; Daniela B. Claro

Ressignificação terminológica e conceptualização infantil da experiência oncológica: um estudo com crianças em tratamento para leucemia linfoblástica aguda 92
Josefa Mônica Almeida Silva Alves; Soraya Carvalho Souza Biller Teixeira

Acessibilidade textual e terminológica em ambiente digital: análise de *corpus* e impacto social na educação em saúde 102
Bruna Rodrigues da Silva

Bulas de medicamentos para o novo Corpop Saúde: curadoria e avaliação linguística de simplificações geradas por modelos de inteligência artificial 110
Maria José Bocorny Finatto; Arthur Scalercio; Bianca Franco Pasqualini

UD_Kadiweu-Unicamp: expandindo a diversidade tipológica da coleção Universal Dependencies com o primeiro *treebank* de uma língua da família Guaicuru 122
Leonel F. de Alencar; Maria Filomena S. Sandalo; Charlotte Marie C. Galves

Córpus Alienista: uma obra machadiana anotada com Universal Dependencies .133
Magali S. Duran; Ursula Puello Sydio; Lucelene Lopes; Thiago A. S. Pardo

TTWtools: uma ferramenta aberta e transversal para etiquetagem e análise em linguística de corpus 143
Luciana Aguiar de Oliveira; José Henrique Padovani

FREIRE: Um Motor Computacional Adaptativo e Sensível à Lógica para Alfabetização de Adultos em Comunidades Tradicionais 152
Lucas Parreias

Ampliação e proposta de sistema de remissivas para um glossário multilíngue direcionado por corpus 160
Maria Eduarda Magalhães Furtado Dutra; Elisa Duarte Teixeira

Compilação e anotação de um corpus paralelo inglês–português de legendas para identificação de expressões idiomáticas.....169
Rozane R. Rebechi; Diego Alves; Isabela Moreira de Oliveira; Christopher dos Santos Filho

E se Pellegrino Artusi fosse blogueiro? Tradução funcional de uma receita histórica na era da IA generativa.....179
Stella E. O. Tagnin; Rozane R. Rebechi

Ecolinguística: uso do IRaMuTeQ na análise do nódulo "animal" em corpus das Ciências Agrárias.....190
Shirlene Bemfica de Oliveira e colaboradoras

ARTIGOS CURTOS ELC 2026 202

A inteligência artificial como parceira da escrita acadêmica em inglês: um assistente de feedback baseado em linguística de corpus.....203
Vivian Meurer Lameira; Ana Eliza Pereira Bocorny; Walcir Cardoso

Um MOOC informado por corpus para a Escrita Acadêmica em inglês na área das Humanidades: Design, Implementação e Avaliação Preliminar.....209
Ana Eliza Pereira Bocorny; Juliana Monteiro dos Santos Almeida

A produção oral acadêmica em inglês: análise de dados extraídos de *corpora* para subsidiar a elaboração de materiais didáticos para o *webinar* 216
Rafael Leandro Götz; Ana Eliza Pereira Bocorny

Redução do espaço acústico em pessoas com esquizofrenia 225
Vitor Marçal de Oliveira; Bruno Neves Rati de Melo Rocha

Detecção e diagnóstico de erros segmentais de pronúncia em corpus oral de aprendizes de Inglês com modelos auto-supervisionados sem ajuste de pesos 232
Marcos Antonio R. Pinto Junior; Michaël Grégoire; Jackson Wilke da C. Souza

Prosódia e função discursiva em palavrões no português brasileiro: uma análise preliminar de "pô" na fala espontânea baseada em corpora.....238
Victor Hugo Medina; Átila Soares Vital

A anotação morfossintática de um corpus pedagógico do kubeo em universal dependencies: desafios tipológicos e decisões linguísticas 245
Pedro Paulo Freire Oliveira; Thiago Blanch Pires

Preparação auditável de corpus histórico: um pipeline em camadas para análise linguística baseada em corpus 252
Luiz Henrique Santos Torres

Investigando marcas de autoria em contos atribuídos a Machado de Assis 257
Ursula Puello Sydio

Reflexões preliminares sobre a anotação das relações de coerência em um corpus de relatos sobre saúde mental do português brasileiro 263
Graziele Thainá Maciel Lima; Roana Rodrigues; Paula Christina Figueira Cardoso; Ewerson Dantas

Capacitismo e Ética em Processamento de Linguagem Natural: Reflexões Semânticas do Léxico Especializado na FrameNet Brasil 270
Iasmin Valéria Miranda Rabelo; Maucha Andrade Gamonal

CoRpMIA: Proposta de um Corpus de Representações em Mídias Culturais da Inteligência Artificial 277
Cibele Santos; Eduardo de Moura; Livy Real

Uma proposta de agenda data-driven para o ensino de italiano em contexto universitário 283
Graziele Altino Frangiotti

Análise de convenções disciplinares em *abstracts* gerados por inteligência artificial 290
Gabriela Carvalho Escobar; Ana Eliza Pereira Bocorny

A prosódia semântica na linguagem das Ciências Agrárias: contribuições da Linguística de *Corpus* para o desenvolvimento de materiais didáticos de inglês para fins acadêmicos 296
Ana Carolina Martins Párcio e colaboradoras

OFICINAS EBRALC 2026 304

A. Introdução aos Grandes Modelos de Linguagem: história, funcionamento e desafios da Inteligência Artificial contemporânea 305
Livy Real; Cibele Simões de Oliveira Santos

B. O uso de corpora de referência por profissionais de Letras para a aprendizagem de língua inglesa 307
Rachel Lourenço; Thiago Blanch Pires

C. Anotação morfossintática de corpora de línguas indígenas: da Plataforma Tycho Brahe ao modelo Universal Dependencies 309
Leonel F. de Alencar; Maria Filomena S. Sandalo; Charlotte Marie C. Galves

D. Como os Grandes Modelos de Linguagem representam a linguagem? Uma introdução aos Transformers para linguistas com Orange Data Mining 311
Marcos Moisés Crisóstomo de Oliveira

E. Linguística de Corpus, modelos de linguagem e Inteligência Artificial: técnicas e ferramentas para simplificar bulas de medicamentos em português do Brasil.....314
Maria José Bocorny Finatto; Bianca Franco Pasqualini

F. Introdução à Mineração de Texto com Orange Data Mining: fundamentos e aplicações práticas 316
Crysttian Paixão; Eduardo Grizenti

TRABALHOS COMPLETOS
ELC 2026

A complexidade lexical e fraseológica na distinção entre níveis de proficiência em inglês: um estudo baseado em corpus

Deise Amaral

Universidade Federal do Rio Grande do Sul (UFRGS)

amaraldeise@gmail.com

Resumo. *Este estudo investiga a sofisticação lexical e fraseológica em produções escritas de exames de proficiência em inglês, com foco na distinção entre os níveis B2 e C1 do QCER. Partindo da premissa de que níveis mais avançados de proficiência se caracterizam pelo uso mais preciso, variado e apropriado de recursos lexicais, a pesquisa analisa 327 textos produzidos em condições operacionais de um exame de larga escala por meio de uma abordagem baseada em corpus. A análise contempla diferentes dimensões da complexidade lexical, incluindo sofisticação, diversidade e densidade lexical, além de medidas fraseológicas relacionadas à frequência, força de associação e diversidade de bigramas. Os resultados indicam que textos C1 apresentam maior densidade, diversidade e sofisticação lexical e fraseológica em comparação aos textos B2. As três dimensões de complexidade também contribuem para a distinção entre os níveis de proficiência. Análises das linhas de concordância de combinações de palavras consideradas sofisticadas, identificadas por altos valores de MI, evidenciam o uso predominante de colocações relacionadas ao desenvolvimento temático, além de expressões voltadas à construção do gênero textual em ambos os níveis. O estudo contribui para a compreensão da sofisticação linguística em contextos de avaliação de proficiência e para o ensino e a preparação para exames.*

Abstract. *This study investigates lexical and phraseological sophistication in written productions from English proficiency examinations, focusing on the distinction between the B2 and C1 levels of the CEFR. Based on the assumption that higher proficiency levels are characterized by more precise, varied, and appropriate use of lexical resources, the study analyses 327 texts produced under operational conditions of a large-scale examination through a corpus-based approach. The analysis encompasses different dimensions of lexical complexity, including lexical sophistication, diversity, and density, as well as phraseological measures related to bigram frequency, strength of association, and diversity. Statistical results indicate that C1 texts display greater lexical and phraseological density, diversity, and sophistication than B2 texts. The three dimensions of complexity also contribute to the distinction between proficiency levels. Concordance-line analyses of word combinations considered sophisticated, identified through high MI scores, revealed the predominant use of collocations related to thematic development, as well as expressions associated with the construction of textual genre in both levels. The study contributes to the understanding of linguistic sophistication in proficiency assessment contexts and to language teaching and test preparation.*

Palavras-chave: complexidade lexical; sofisticação fraseológica; avaliação de proficiência

1. Introdução

A complexidade lexical ocupa um papel central nos critérios de avaliação de produção escrita em exames de proficiência de larga escala. Nos níveis mais avançados do Quadro Comum Europeu de Referência para Línguas (QCER), particularmente nos níveis C1 e C2, espera-se que os textos demonstrem maior precisão, variedade e sofisticação no uso de recursos lexicais, incluindo combinações idiomáticas e itens menos frequentes (Brown, 2006; Hawkey & Barker, 2004). Nesse contexto, importantes descritores de avaliação incluem uso idiomático, sofisticação lexical e precisão colocacional entre seus critérios (Biber & Gray, 2013; Council of Europe, 2020; IELTS, 2023). A sofisticação lexical é comumente medida através da frequência de uso de palavras individuais, enquanto a sofisticação fraseológica pode ser compreendida a partir da frequência e da intensidade de associação entre palavras que tendem a ocorrer conjuntamente de maneira estatisticamente significativa. Pesquisas sobre linguagem de aprendizes (Bestgen & Granger, 2014; Kyle et al., 2018) indicam que textos avaliados em níveis mais altos apresentam combinações lexicais mais fortemente associadas e compostas por vocabulário menos frequente.

Com base nesse pressuposto, o presente estudo de corpus investiga características lexicais e fraseológicas que podem estar associadas à distinção entre níveis intermediários e avançados de proficiência em língua inglesa (B2 e C1), em textos produzidos na parte escrita de um exame de proficiência de alto impacto. O estudo baseia-se na análise de um corpus de textos produzidos em contexto de avaliação de proficiência e busca responder a duas questões relacionadas: (i) Quais dimensões de complexidade lexical e fraseológica contribuem para distinguir textos em níveis B2 e C1? E (ii) Em que medida padrões fraseológicos se manifestam no uso de combinações sofisticadas de palavras em cada nível?

2. Referencial teórico

Nas últimas décadas, pesquisas voltadas ao desenvolvimento e à validação de exames de proficiência têm recorrido de forma crescente a dados empíricos gerados por ferramentas da Linguística de Corpus (LC). Ao possibilitar a análise de usos reais da língua, esses dados permitem a elaboração de critérios de avaliação e descritores mais confiáveis e fundamentados (Fulcher, 1996), além de favorecerem a compreensão de como avaliadores interpretam e aplicam esses critérios. Estudos que investigam a percepção de avaliadores sobre critérios de avaliação (Brown, 2006; Hawkey; Barker, 2004), bem como análises de respostas produzidas em exames de proficiência (Iwashita et al., 2008; Read; Nation, 2006), apontam de forma recorrente o uso de palavras menos frequentes, amplitude lexical, sofisticação lexical e uso de linguagem idiomática como característica associada a usuários mais proficientes da língua. A sofisticação lexical é geralmente compreendida como uma das dimensões da Complexidade Lexical (Paquot, 2018) ou da Riqueza Lexical (Read, 2000). Outras dimensões incluem a Diversidade Lexical, entendida como a variedade de palavras empregadas, e a Densidade Lexical, operacionalizada como a proporção de palavras lexicais em um texto (Kyle, 2019; Paquot, 2018). Tradicionalmente, a Sofisticação Lexical é entendida como a proporção

de palavras pouco frequentes ou incomuns em relação a palavras frequentes em um texto (Paquot, 2018, 2019), tomando como base sua frequência de ocorrência em corpora de referência.

Recentemente, pesquisas têm ampliado essa concepção ao incorporar a Sofisticação Fraseológica, voltada à análise de combinações de palavras, e não apenas de itens lexicais isolados. Nessa perspectiva, passam a ser considerados índices como média da frequência de n-gramas, força de associação entre palavras em combinações recorrentes e proporção de combinações sofisticadas nos textos. Paquot (2018) propõe que, de modo semelhante à sofisticação lexical, a sofisticação fraseológica refere-se ao uso de combinações de palavras "incomuns que permitem aos escritores expressar significados de maneira precisa e sofisticada" (Read, 2000, p. 200).

Neste estudo, o conceito de combinação de palavras, ou item fraseológico, fundamenta-se na abordagem distribucional, baseada em frequência, adotada em estudos de colocações baseados em corpus (Gablasova et al., 2017; Granger, 2021). Nessa perspectiva, colocações são definidas pela coocorrência de palavras que aparecem juntas com frequência superior àquela esperada a partir de suas ocorrências individuais e do tamanho do texto em que ocorrem (Jones; Sinclair, 1974). Assim, as categorias fraseológicas não são previamente estabelecidas, mas identificadas por meio de medidas estatísticas de frequência e distribuição. Esta concepção é operacionalizada pelo método *collgrams* (Bestgen; Granger, 2014), que extrai sequências adjacentes de palavras (bigrams, neste estudo) com forte atração entre si, posteriormente classificadas em categorias linguísticas reconhecíveis por professores e alunos, como *phrasal verbs* e expressões idiomáticas. A medida de força de atração utilizada, Informação Mútua (MI, do inglês *Mutual Information*), é empregada como idicador de sofisticação fraseológica por identificar combinações colocacionalmente convencionais, precisas, semanticamente adequadas e apropriadas ao gênero (Paquot, 2018, 2019, p.138).

Essa concepção mostra-se particularmente adequada aos objetivos desta investigação, uma vez que permite analisar combinações sofisticadas no contexto da avaliação de recursos lexicais em diferentes níveis de proficiência.

3. Metodologia

O corpus de estudo é composto por 327 textos produzidos em condições operacionais de aplicação do exame de proficiência em inglês IELTS, em resposta à tarefa 2 da parte escrita. Nessa tarefa, os examinados devem produzir um *essay* acadêmico semiformal de, no mínimo, 250 palavras, no qual são apresentadas soluções para um problema ou a discussão de diferentes pontos de vista, com justificativas e exemplos quando relevante (IELTS, 2026). Os textos, previamente avaliados por examinadores certificados, receberam notas 5,5, 6, 7 ou 7,5 e foram agrupados nos níveis B2 (5,5 e 6) e C1 (7 e 7,5), conforme a classificação do exame (IELTS, 2025). Após a limpeza (remoção de símbolos, números e pontuação), o corpus totalizou 94.164 *tokens* e 50.067 *types*.

Anteriormente às análises, o corpus foi pré-processado mediante a padronização de variantes ortográficas típicas de corpora de aprendizes, mantendo-se o registro das alterações para análises posteriores. O procedimento buscou reduzir possíveis distorções nas medidas baseadas em corpus, evitando a inflação artificial de índices fundamentados

em *types* lexicais. A compilação, limpeza, tokenização, padronização e análise dos dados foram realizadas com Excel, AntConc v. 3.5.9, TAALES (Kyle et al., 2018) e scripts em R.

As análises foram conduzidas a partir de um modelo multidimensional de complexidade lexical e fraseológica, com foco principal na sofisticação fraseológica, mas considerando também diversidade lexical e fraseológica, além de sofisticação e densidade lexicais. Foram computadas 11 medidas baseadas no corpus de estudo e no de referência, composto por 240 artigos publicados (467.884 *tokens*) de diferentes áreas provenientes dos subcorpora de Prosa Acadêmica dos corpora BE06, BE21 e AmE06 (Baker, 2009, 2023; Potts; Baker, 2012). Essa escolha está alinhada às necessidades dos candidatos ao IELTS Academic, cuja certificação se destina principalmente a fins acadêmicos, e segue Durrant e Brenchley (2019), para quem os corpora de referência devem representar os objetivos de produção textual. Além disso, o uso de corpora de referência mais amplos é prática recorrente em estudos de sofisticação pois permite interpretações mais significativas e comparáveis da raridade lexical e evita a circularidade de frequências internas ao próprio corpus (Kyle, 2019; Paquot, 2019).

Os índices de sofisticação lexical e fraseológica (Tabela 1) incluíram frequência de palavras e bigramas e escores de MI, complementados por diferentes índices de diversidade — *Type/Token Ratio* (TTR), MATTR e MTLD para diversidade lexical, e RTTR (*Root Type-Token Ratio*) para diversidade fraseológica — além da medida de densidade lexical (a razão entre as palavras de conteúdo e as gramaticais). Os dados foram primeiramente explorados por meio de estatísticas descritivas, incluindo medidas de tendência central (média, mediana e desvio padrão), variabilidade (teste de Levene) e distribuição (teste de Shapiro-Wilk). Com base nos resultados das análises de normalidade, foram selecionados testes inferenciais: testes *t* para as medidas com distribuição normal na comparação entre textos B2 e C1, e testes não paramétricos (Mann-Whitney U) para as demais medidas.

Em seguida, foram aplicados modelos de regressão logística binária para identificar quais dimensões de complexidade contribuíam de forma mais significativa para a predição do nível de proficiência, com nível (B2 e C1) como variável dependente e todas as medidas de diversidade, densidade e sofisticação como preditoras. Esses resultados foram complementados por análises qualitativas das linhas de concordância de bigramas selecionados com altos escores de MI (>7) buscando identificar os padrões lexicais das combinações sofisticadas em cada nível.

4. Resultados

Os resultados das análises estatísticas indicaram que os textos classificados como C1 apresentaram maior densidade e diversidade lexical em todas as medidas testadas, como pode ser visto na Tabela 1. Os índices de sofisticação, por sua vez, revelaram resultados divergentes: a expectativa de redução significativa da frequência média de palavras no nível mais avançado foi confirmada apenas na medida baseada em *types* (Freq. *Types* na Tabela 1), enquanto as medidas de frequência de bigramas apontaram aumento de B2 para C1 (Freq. Bigr. *tokens* e *types*). Esses resultados sugerem que os textos classificados como C1 apresentam maior uso de palavras menos frequentes quando as repetições são desconsideradas, convergindo, assim, com os resultados obtidos pelos

índices de diversidade. Em contrapartida, esse padrão não foi observado nas medidas baseadas em bigramas. A força de associação entre os bigramas, medida pelo MI, foi significativamente maior em C1 nos dois índices (MI *tokens* e MI *types*), corroborando estudos que vinculam maior sofisticação fraseológica a níveis mais elevados de proficiência (Bestgen; Granger, 2014; Paquot, 2018).

Tabela 1. Resultados das análises estatísticas descritivas e inferenciais entre B2 e C1

Medidas	B2		C1			Entre grupos	
	Média (DP)	Mediana	Média (DP)	Mediana	Compara-ção	*p*	Efeito
TTR	0,52 (0,06)	0,53	0,55 (0,05)	0,54	t = 3,63	< 0,001	d = 0,41
MATTR	0,80 (0,03)	0,81	0,82 (0,002)	0,82	U = 31984	< 0,001	P = 0,65
MTLD	70,8 (16,3)	69,2	81,5 (18,2)	79,9	U = 32651	< 0,001	P = 0,67
Densidade	0,59 (0,03)	0,59	0,61 (0,02)	0,61	t = 4,73	< 0,001	d = 0,06
Freq. tokens	7937,25 (1633,24)	7789,73	8134,67 (1527,96)	1090,66	U = 29248	0,15	P = 0,54
Freq. types	1242,16 (145,76)	1225,13	1090,66 (123,12)	1089,03	U = 20444	< 0,001	P = 0,21
RTTR	14 (1,07)	14,1	15 (1,07)	14,1	U = 7450	< 0,001	P = 0,72
Freq. Bigr. tokens	127 (46,6)	121	141 (47,2)	132	U = 9088	< 0,001	P = 0,58
Freq. Bigr. types	94,1 (22.3)	98,1	95,7 (19,5)	95,9	t = 0,68	0,5	-
MI tokens	2,57 (0,22)	2,57	2,64 (0,21)	2,63	t = 2.97	0,003	d = 0,33
MI types	2,50 (0,24)	2,49	2,59 (0,21)	2,59	t = 3,48	< 0,001	d = 0,39

Nota: Nível de significância (alfa) definido em α = 0,05. Freq. = frequência; Bigr. = bigramas; *t* é o resultado do test t para dados com distribuição normal enquanto U é o resultado do Mann-Whitney para dados não paramétricos; o tamanho do efeito é indicado pelo *d* de Cohen (0,2 = pequeno, 0,5 = médio, 0,8 grande) e pela Probabilidade de Superioridade (*P*) - quanto mais próximo de 0,50 (50%), menos diferença entre os grupos.

Os modelos de regressão logística demonstraram que a diversidade fraseológica (RTTR) exerceu o maior efeito sobre o nível de proficiência (OR = 2,47, p < ,001), seguida pela densidade lexical (OR = 1,57, p < ,001) e pela força de atração entre os bigramas, considerando *types* (OR = 1,52, p < ,001). Os *odds ratios* (OR) superiores a 1 indicam que, quanto maior os valores das medidas, maiores as chances de o texto ser classificado no nível mais avançado, embora a maioria seja relativamente próxima de 50% (por exemplo, OR = 1,57 corresponde a chances 57% maiores). Esses resultados evidenciam que as duas dimensões fraseológicas analisadas (diversidade - RTTR e sofisticação, medida pelo MI), assim como a densidade lexical, contribuem para a probabilidade de os textos serem avaliados no nível mais avançado (C1).

As análises qualitativas das linhas de concordância dos bigramas sofisticados (MI > 7) revelaram a prevalência de combinações de palavras lexicais relacionadas ao desenvolvimento do tópico. Expressões convencionais oriundas de termos técnicos, como "*global warming*" e "*climate change*", cujos significados resultam da combinação de seus componentes e constituem unidades semânticas distintas e reconhecidas, representaram a maior parte dos bigramas selecionados pelo MI nos dois níveis analisados. Esses achados corroboram a literatura sobre a tendência de o MI selecionar combinações exclusivas, especializadas ou técnicas (Bestgen; Granger, 2014;

Gablasova et al., 2017).

Embora o MI selecione preferencialmente combinações de palavras lexicais, os dados também revelaram contribuição significativa de bigramas com uma palavra gramatical entre as combinações consideradas sofisticadas. Análises mais granulares identificaram que essas expressões contribuem para a realização do gênero textual, que, no IELTS, requer que os examinandos se posicionem em relação à discussão proposta (IELTS, 2026). Essa característica pode explicar a presença de expressões como "[*I*] *am convinced*", "*I think*" e "[*as far as I*] *am concerned*", usadas para marcar explicitamente a posição dos autores.

Expressões idiomáticas figuraram entre os bigramas mais sofisticados em ambos os níveis de proficiência, sem diferenças relevantes entre eles. Tanto *phrasal verbs,* como "*slow down*" e "*point out*", quanto expressões caracterizadas por opacidade semântica e fixidez, como "[*at a*] *first glance*" e "*keep track* [*of*]", ocorreram nos dois níveis, com frequência ligeiramente maior no B2 (cerca de 10% dos bigramas sofisticados, em comparação a 8% em textos C1).

5. Conclusão

As diferenças significativas entre B2 e C1 em todas as dimensões da Complexidade Lexical e Fraseológica, assim como os resultados dos modelos de regressão indicando a contribuição da maioria dos índices para a previsão dos escores dos textos, corroboram estudos que apontam o efeito dos recursos lexicais na avaliação da proficiência escrita (Bestgen; Granger, 2014; Hawkey; Barker, 2004; Paquot, 2018, 2019). No entanto, diferentemente do sugerido por Paquot (2018), não foi observada maior contribuição dos índices fraseológicos, uma vez que os índices lexicais também apresentaram efeito na distinção de níveis.

Se, por um lado as análises qualitativas dos bigramas sofisticados corroboram estudos que utilizam o MI como medida de sofisticação fraseológica ao revelarem a predominância de combinações de palavras de conteúdo relacionadas ao tópico do texto, por outro, a ocorrência de combinações com elementos gramaticais voltadas à expressão do gênero textual aponta para uma questão que será investigada em estudos subsequentes: o papel do gênero na avaliação de textos de proficiência. Expressões que indicam posicionamento autoral respondem à proposta de escrita do IELTS; contudo, ainda não está claro em que medida formas mais pessoais, como "*I think*", ou marcadores discursivos, como "*let's start*", influenciam a avaliação dos textos. A análise dessa relação, portanto, será objeto de investigações futuras.

Embora expressões idiomáticas tenham sido encontradas em todos os níveis do CEFR analisados, elas não parecem contribuir significativamente para a distinção entre níveis quando definidas como expressões fixas e semanticamente não transparentes. Entretanto, sob uma concepção mais ampla de idiomaticidade, os bigramas com MI superior a 7 podem ser considerados combinações convencionais e, portanto, idiomáticas. Assim, em resposta à segunda pergunta de pesquisa, os padrões identificados pelo MI indicam que as combinações sofisticadas incluem tanto expressões idiomáticas quanto colocações cujos elementos apresentam forte associação e desempenham funções comunicativas relacionadas ao gênero textual e ao desenvolvimento do tópico.

A explicitação dos recursos lexicais associados a avaliações de textos mais

avançados em contextos de proficiência, evidenciados neste estudo, busca contribuir para a formação de professores e aprendizes, não apenas na preparação para exames, mas também no ensino de inglês como língua adicional de modo mais amplo. Da mesma forma, os achados têm o potencial de contribuir para a elaboração de descritores de níveis de proficiência mais precisos, fundamentados em dados empíricos de uso da língua.

Referências

BAKER, P. The BE06 Corpus of British English and recent language change. **International Journal of Corpus Linguistics,** *v.* 14, n. 3, p. 312–337, 2009. Doi: 10.1075/ijcl.14.3.02bakissn 1384–6655.

BAKER, P. A year to remember? Introducing the BE21 corpus and exploring recent part of speech tag change in British English. **International Journal of Corpus Linguistics,** v. 28, n. 3, p. 407–429, 23 maio 2023. Disponível em:

https://doi.org/10.1075/ijcl.22007.bak. Acesso em: 5 maio 2026.

BESTGEN, Y.; GRANGER, S. Quantifying the development of phraseological competence in L2 English writing: an automated approach. **Journal of Second Language Writing**, v. 26, p. 28-41, 2014. Disponível em: https://doi.org/10.1016/j.jslw.2014.09.004 . Acesso em: 5 fev. 2026.

BIBER, D.; GRAY, B. Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: a lexico-grammatical analysis. **TOEFL iBT® Research reports**. ETS, 2013.

BROWN, A. Candidate discourse in the revised IELTS Speaking Test. *In*: MCGOVERN, P.; WALSH, S. (ed.). **IELTS Research Reports**, v. 6. IELTS Australia e British Council, p. 71-89, 2006. Disponível em: https://ielts.org/cdn/Research/candidate-discourse-in-revised-ielts-speaking-test-brown-2006.pdf. Acesso em: 5 maio 2026.

COUNCIL OF EUROPE. **Common European framework of reference for languages:** learning, teaching, assessment – Companion volume. Council of Europe Publishing, 2020. Disponível em: www.coe.int/lang-cefr. Acesso em: 3 abr. 2026.

DURRANT, P.; BRENCHLEY, M. Development of vocabulary sophistication across genres in English children's writing. **Reading and Writing**, v. 32, p. 1927–1953, 2019. Disponível em: https://doi.org/10.1007/s11145-018-9932-8. Acesso em: 03 set. 2026.

FULCHER, G. Does thick description lead to smart tests? A data-based approach to rating scale construction. **Language Testing**, v. 13, n. 2, 1996. Disponível em: https://doi.org/10.1177/026553229601300205. Acesso em: 5 maio 2026.

GABLASOVA, D.; BREZINA, V.; MCENERY, T. Collocations in corpus-cased language learning research: identifying, comparing, and interpreting the evidence. **Language Learning**, v. 67, n. S1, p. 155-179, 2017. Disponível em: https://doi.org/10.1111/lang.12225. Acesso em: 3 maio 2026.

GRANGER, S. Phraseology, corpora and L2 research. *In*: GRANGER, S.

(ed.), **Perspectives on the L2 phrasicon:** the view from learner corpora (pp. 3-21). Multilingual Matters, 2021.

GRIES, S. Phraseology and linguistic theory: a brief survey. *In*: GRANGER, S.; MEUNIER, F. (ed), **Phraseology:** an interdisciplinary perspective. John Benjamins, p. 3-25, 2008.

HAWKEY, R.; BARKER, F. Developing a common scale for the assessment of writing. **Assessing Writing**, v. *9, n.* 2, p. 122-159, 2004. Disponível em: https://doi.org/10.1016/j.asw.2004.06.001. Acesso em: 5 maio 2026.

IELTS™. Writing Band Descriptors. **IELTS.org**. British Council, IDP: IELTS Australia and Cambridge University Press and Assessment, 2023. Disponível em: https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Guides/ielts-writing-band-descriptors.pdf. Acesso em 5 maio 2026.

IELTS™. IELTS and the CEFR. **IELTS.org**. British Council, IDP: IELTS Australia and Cambridge University Press and Assessment, 2025. Disponível em: https://ielts.org/organisations/ielts-for-organisations/compare-ielts/ielts-and-the-cefr . Acesso em 3 set. 2026.

IELTS™. **IELTS.org**. British Council, IDP: IELTS Australia and Cambridge English Language Assessment, 2026. Disponível em: https://ielts.org/. Acesso em: 3 abr. 2026.

IWASHITA, N.; BROWN, A.; MCNAMARA, T.; O'HAGAN, S. Assessed levels of second language speaking proficiency: how distinct? **Applied Linguistics**, v. 29, n. 1, p. 24–49, 2008. Disponível em: https://doi.org/10.1093/applin/amm017. Acesso em: 2 jan 2026.

KYLE, K. Measuring Lexical Richness. *In*: WEBB, S. (ed.), **The Routledge handbook of vocabulary studies,** p. 454-476, 2019. Routledge.

KYLE, K.; CROSSLEY, S.; BERGER, C. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. **Behavior Research Methods**, v.50, p.1030–1046.Disponível em: https://doi.org/10.3758/s13428-017-0924-4. Acesso em 5 abr. 2026.

PAQUOT, M. Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. **Language Assessment Quarterly**, v. 15, n. 1, p. 29-43, 2018. Disponível em: https://doi.org/10.1080/15434303.2017.1405421. Acesso em: 2 abr. 2026.

PAQUOT, M. The phraseological dimension in interlanguage complexity research. **Second Language Research**, v. 35, n.1, p. 121-145, 2019. Disponível em: https://doi.org/10.1177/0267658317694221. Acesso em: 5 maio 2026.

POTTS, A.; BAKER, P. Does semantic tagging identify cultural change in British and American English? **International Journal of Corpus Linguistics**, v.17, n.3, p. 295–324, 2012. Disponível em: https://doi.org/10.1075/ijcl.17.3.01pot. Acesso em: 2 mar. 2026.

READ, J. **Assessing Vocabulary**. Cambridge: Cambridge University Press, 2000. Doi: https://doi.org/10.1017/CBO9780511732942.

READ, J.; NATION, P. An investigation of the lexical dimensions of the IELTS

Speaking Test. *In:* McGovern, P.; Walsh, S. (ed.), **IELTS Research Reports**, 6. UCLES, The British Council, IDP Education Australia: IELTS Australia, p. 207-231, 2006. Disponível em: https://ielts.org/cdn/Research/investigation-of-lexical-dimension-of-ielts-speaking-test-read-et-al-2006.pdf. Acesso em: 5 maio 2026.

Complexidade gramatical na escrita universitária: nominalizações e adjetivos atributivos em produções de graduandos de Letras Inglês

Marine Laisa Matte[1], Simone Sarmento[2]

[1]Instituto Federal Sul-rio-grandense (IFSul), Brasil; [2]Universidade Federal do Rio Grande do Sul (UFRGS), Brasil

marinematte@ifsul.edu.br; simone.sarmento@ufrgs.br

Resumo. *Este estudo analisa o uso de nominalizações e adjetivos atributivos em 272 textos do Corpus of Academic Written English Students' Assignments (CAWESA), produzidos por graduandos de Letras Inglês em oito disciplinas. Os textos foram classificados segundo o nível de estudo e o propósito comunicativo. Após a etiquetagem linguística, análises descritivas e modelos fatoriais de ANOVA examinaram os efeitos dessas variáveis sobre as características investigadas. Os resultados apontam aumento não linear no uso de nominalizações e adjetivos atributivos ao longo dos níveis de estudo. O propósito comunicativo apresentou maior impacto sobre as escolhas linguísticas, sobretudo em tarefas de explicação, proposta e avaliação. Os achados apoiam práticas pedagógicas que articulem formas linguísticas e funções comunicativas no ensino da escrita acadêmica.*

Abstract. *This study examines the use of nominalizations and attributive adjectives in 272 texts from the Corpus of Academic Written English Students' Assignments (CAWESA), produced by English Language and Literature undergraduates across eight courses. The texts were classified by level of study and communicative purpose. Following linguistic tagging, descriptive analyses and factorial ANOVA models examined the effects of these variables on the features investigated. The results indicate a non-linear increase in nominalizations and attributive adjectives across levels of study. Communicative purpose had a greater impact on linguistic choices, particularly in explanation, proposal, and evaluation tasks. The findings support pedagogical practices that integrate linguistic forms and communicative functions in academic writing instruction.*

Palavras-chave: escrita universitária; complexidade gramatical; elementos lexicogramaticais.

1. Introdução

A escrita acadêmica constitui uma exigência central da trajetória universitária, demandando dos estudantes não apenas domínio de conteúdos específicos, mas também familiaridade com convenções discursivas e recursos linguísticos característicos desse contexto. Entre tais recursos, estudos sobre complexidade gramatical destacam a recorrência de elementos lexicogramaticais associados à densidade informacional e à

elaboração frasal, como nominalizações e adjetivos atributivos, frequentemente empregados em textos acadêmicos (Biber; Gray; Poonpon, 2011; Staples et al., 2016, 2023). Em comparação a outros registros, a escrita acadêmica tende a apresentar maior complexidade frasal, marcada pela expansão do sintagma nominal e pela condensação de significados.

Nas últimas décadas, pesquisas baseadas em corpus têm ampliado a compreensão sobre padrões linguísticos da escrita universitária, incluindo investigações sobre gêneros acadêmicos, vocabulário especializado e desenvolvimento da complexidade linguística (Gardner; Nesi, 2013; Staples et al., 2016). Entretanto, permanece em debate a forma mais adequada de categorizar produções estudantis, especialmente devido à diversidade de tarefas desenvolvidas em contextos universitários. Nesse sentido, estudos recentes sugerem que a análise da escrita acadêmica pode se beneficiar de uma abordagem centrada nos propósitos comunicativos dos textos, em vez de classificações baseadas em gêneros que circulam na sociedade (Goulart, 2024).

Partindo dessa perspectiva, o presente estudo investiga o uso de nominalizações e adjetivos atributivos em textos produzidos por graduandos de Letras Inglês, considerando dois fatores: (i) o nível de estudo dos estudantes e (ii) os propósitos comunicativos das produções. Para isso, utiliza-se o *Corpus of Academic Written English Students' Assignments* (CAWESA), composto por textos produzidos em diferentes disciplinas de língua inglesa. Busca-se responder em que medida essas características léxicogramaticais variam ao longo da formação acadêmica e conforme as demandas comunicativas das tarefas propostas, bem como verificar o impacto dessas variáveis por meio de análises fatoriais de variância (ANOVA).

2. Referencial teórico

Estudos sobre escrita acadêmica têm demonstrado que diferentes registros mobilizam padrões linguísticos específicos, sendo a linguagem acadêmica caracterizada por maior densidade informacional e complexidade sintática (Biber et al., 1999; Biber, 2006). Na perspectiva da Linguística de Corpus (LC), a análise de produções autênticas permite identificar regularidades linguísticas e compreender como recursos léxicogramaticais se distribuem em diferentes contextos de escrita (McEnery; Wilson, 1996). Nesse sentido, pesquisas baseadas em corpus têm contribuído para o entendimento de padrões típicos da escrita acadêmica, incluindo investigações sobre vocabulário acadêmico, linguagem disciplinar e pacotes lexicais (Coxhead; Nation, 2001; Cortes, 2004; Gardner; Nesi, 2013; Goulart, 2016; Bocorny et al., 2021).

Entre os estudos seminais da área, Biber et al. (1999) demonstram que a linguagem acadêmica apresenta traços distintos de registros não acadêmicos, sobretudo pela maior complexidade estrutural. Enquanto a linguagem falada tende a refletir produção imediata, sendo frequentemente "*produced on the spur of the moment*" (Biber et al., 1999, p. 964), a escrita acadêmica favorece estruturas mais densas e planejadas. Pesquisas posteriores reforçam que a complexidade da escrita acadêmica ocorre predominantemente no nível frasal, diferentemente da oralidade, que tende a mobilizar estruturas clausais (Biber; Gray; Poonpon, 2011; Biber; Gray, 2016). Assim, em contextos conversacionais, a complexidade pode emergir em exemplos como *"they are supposed to check the following items that appeared in their research"* (CAWESA), em

que a oração modifica o substantivo *items*. Já na escrita acadêmica, a complexidade tende a concentrar-se no sintagma nominal, como em *"the federal system will not make a difference in reducing crime"* (CAWESA), em que a construção nominal contribui para maior condensação informacional.

Nesse contexto, nominalizações e adjetivos atributivos destacam-se como importantes indicadores de complexidade gramatical, uma vez que favorecem a compactação de significados, o aumento da densidade lexical e a especificação de informações em textos acadêmicos (Staples et al., 2016; Staples et al., 2023). As nominalizações transformam processos e ações em entidades nominais, promovendo maior abstração discursiva, enquanto os adjetivos atributivos exercem função modificadora sobre substantivos, ampliando a precisão descritiva e, frequentemente, contribuindo para avaliações e posicionamentos argumentativos. Estudos prévios indicam que tais recursos tendem a ocorrer com maior frequência em níveis mais avançados de escrita acadêmica, embora sua distribuição possa variar conforme o tipo de tarefa e o contexto disciplinar (Staples; Reppen, 2016; Ansarifar; Shahriari; Pishghadam, 2018).

Além do nível de estudo, pesquisas recentes sugerem que os propósitos comunicativos dos textos podem influenciar significativamente as escolhas linguísticas realizadas pelos estudantes (Goulart, 2024). Sob essa perspectiva, analisar produções acadêmicas a partir das funções desempenhadas pelos textos, como explicar, argumentar, propor ou revisar, pode oferecer evidências mais robustas sobre a relação entre forma linguística e contexto de uso, especialmente no que diz respeito à mobilização de recursos associados à complexidade gramatical.

3. Metodologia

Esta pesquisa caracteriza-se como um estudo baseado em corpus, centrado na análise de duas características lexicogramaticais associadas à complexidade gramatical: nominalizações e adjetivos atributivos. Os dados provêm do *Corpus of Academic Written English Students' Assignments* (CAWESA), composto por 272 textos produzidos por graduandos de Letras Inglês, totalizando 279.990 palavras. Os textos foram produzidos em oito disciplinas de língua inglesa (English 1 a English 8), representando diferentes momentos da trajetória formativa dos estudantes.

Tabela 1. Distribuição dos textos e do número de palavras por nível de estudo

	Número de textos	**Número de palavras**
Inglês 1	15	15.680
Inglês 2	17	9.471
Inglês 3	23	14.256
Inglês 4	30	22.597
Inglês 5	50	37.414
Inglês 6	83	120.341

	Número de textos	**Número de palavras**
Inglês 7	20	17.608
Inglês 8	34	42.623
TOTAL	*272*	*279.990*

Considerando a heterogeneidade dos gêneros acadêmicos produzidos no contexto universitário, os textos foram organizados segundo seus propósitos comunicativos, seguindo adaptações do modelo de Goulart (2024). As categorias incluíram: analisar, argumentar, explicar, aconselhar, narrar, narrar experiência pessoal, propor e revisar. Essa classificação buscou contemplar a diversidade de tarefas acadêmicas sem depender exclusivamente de categorizações tradicionais de gênero. A categorização dos textos foi realizada de forma independente por duas pessoas: uma das autoras deste trabalho, à época de seu doutoramento, e um bolsista de iniciação científica. Nos casos em que houve discrepância entre as classificações, as ocorrências foram analisadas conjuntamente pelos dois avaliadores, até que se chegasse a um consenso.

Após a compilação do corpus, os textos foram submetidos à etiquetagem morfossintática (*part-of-speech tagging*) com o *Biber Tagger* (Biber, 2006). Concluídas a verificação e a correção de possíveis erros de etiquetagem, seguindo o procedimento de Goulart (2024), todas as características apresentaram índices de precisão e *recall* superiores a 0,90. Em seguida, as características foram contabilizadas com o TagCount (Biber, 1988) e normalizadas para 1.000 palavras. A contagem das características em cada texto permitiu examinar variações em função do nível de estudo e do propósito comunicativo. Em seguida, realizaram-se análises estatísticas descritivas para observar padrões de distribuição das características investigadas nos diferentes níveis de estudo e propósitos comunicativos. Para verificar o impacto dessas variáveis, foram conduzidas análises fatoriais de variância (ANOVA), considerando como variáveis dependentes a frequência de nominalizações e os adjetivos atributivos e, como variáveis independentes, o nível de estudo, o propósito comunicativo e a interação entre ambos. Complementarmente, foi calculado o tamanho de efeito (η^2) para estimar a magnitude das diferenças observadas.

4. Resultados

As análises descritivas revelam que o uso de nominalizações e adjetivos atributivos não apresenta crescimento linear ao longo dos níveis de estudo, embora seja possível identificar tendências gerais de aumento entre os níveis iniciais e finais da formação. No caso das nominalizações, observam-se oscilações nos primeiros níveis (Inglês 1 a Inglês 4), seguidas de maior frequência nos níveis avançados, especialmente em Inglês 7 e Inglês 8.

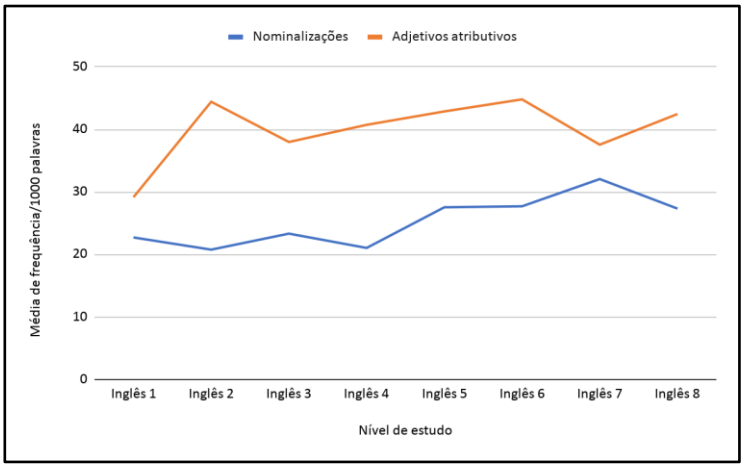


Figura 1. Frequência média de nominalizações e adjetivos atributivos por nível de estudo

Esse comportamento sugere uma progressiva incorporação de recursos associados à densidade informacional e à compactação discursiva, características frequentemente atribuídas à escrita acadêmica (Biber; Gray, 2016).

Os adjetivos atributivos, por sua vez, apresentam maior frequência geral ao longo do corpus, com destaque para Inglês 2 e Inglês 6. Apesar das oscilações entre os níveis, verifica-se crescimento global no uso dessa característica léxicogramatical, sobretudo quando comparados os níveis inicial e final da trajetória formativa. Esses resultados corroboram estudos anteriores que associam a expansão do sintagma nominal ao desenvolvimento da escrita acadêmica e ao aumento da complexidade frasal (Staples et al., 2016; Staples et al., 2023).

Quando observadas as diferenças entre propósitos comunicativos, os resultados tornam-se mais expressivos. As nominalizações apresentaram maiores frequências em tarefas com os propósitos de "propor", "revisar" e "analisar", enquanto a menor ocorrência foi observada em textos voltados ao aconselhamento pessoal. Esse padrão pode ser explicado pela natureza dos textos acadêmicos mais analíticos e propositivos, nos quais processos e ações tendem a ser reformulados como entidades nominais, favorecendo maior abstração e condensação de informações.

Os adjetivos atributivos apresentaram maior frequência em textos com os propósitos de "revisar" e "explicar", contextos em que a descrição detalhada e a avaliação de objetos, conceitos ou argumentos desempenham papel central. Em textos de revisão, por exemplo, a recorrência de adjetivos permite caracterizar e avaliar teorias, estudos e posicionamentos, enquanto, em textos explicativos, esses recursos auxiliam na especificação de conceitos e fenômenos. Em contrapartida, tarefas voltadas ao aconselhamento pessoal apresentaram menor uso tanto de nominalizações quanto de adjetivos atributivos, sugerindo maior proximidade com estruturas clausais típicas de registros menos densos informacionalmente.

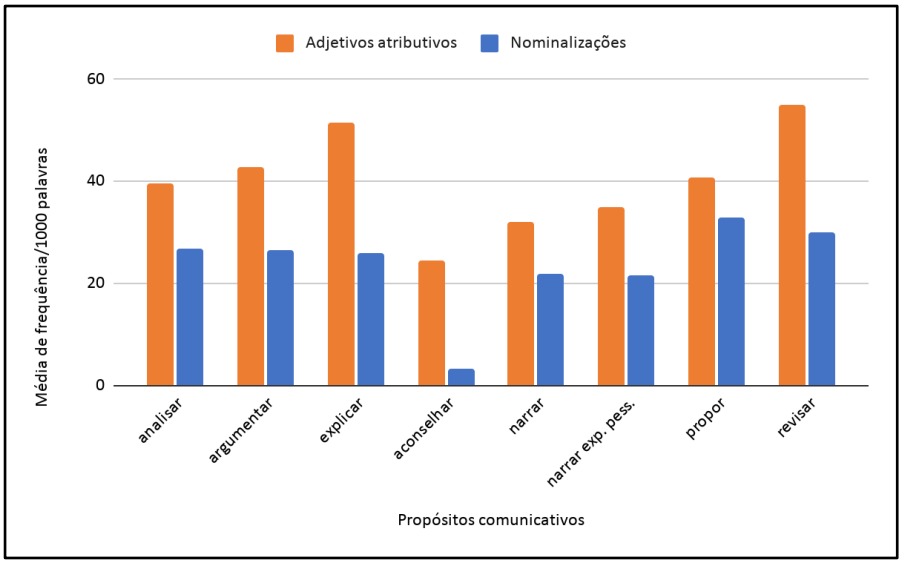


Figura 2. Frequência média de nominalizações e adjetivos atributivos por propósito comunicativo

Os resultados das análises fatoriais de variância (ANOVA) reforçam essas tendências.

Tabela 2. Resultados das análises fatoriais de variância

Elemento léxico gramatical	**Modelo**	**R^2**	**Nível de estudo**	***η² [H]***	**Propósito comunicativo**	***η² [H]***	**Nível x propósito**	***η² [H]***
Attributive adjectives	<.001	.322	<.05	0.02	<.001	0.11	ns	0.16
Nominalizations	<.001	.294	ns	0.06	<.001	0.12	<.05	0.18

Para os *adjetivos atributivos*, o modelo mostrou-se estatisticamente significativo, com um R^2 igual a 0,322, indicando que explica 32,2% da variância dessa característica. Os resultados apontam para uma influência significativa tanto do nível de estudo quanto do propósito comunicativo sobre o uso dessa característica, com tamanhos de efeito pequeno e grande, respectivamente (0,02 e 0,11). A maior magnitude de efeito observada para o propósito comunicativo sugere que a natureza da tarefa exerce um papel mais decisivo sobre as escolhas linguísticas do que o estágio formativo do estudante. A interação entre as duas variáveis, contudo, não se mostrou estatisticamente significativa, indicando que seus efeitos operam de forma independente. Esses resultados referentes aos *adjetivos atributivos* estão parcialmente alinhados com pesquisas anteriores (Staples; Reppen, 2016; Staples et al., 2016). Em ambos os estudos, os modelos, os fatores investigados (L1 e gênero, no primeiro caso; nível e gênero, no segundo) e a interação entre eles mostraram-se estatisticamente significativos.

Em relação às *nominalizações*, o modelo mostrou-se capaz de explicar 29,4% da variância dessa característica no corpus, com o propósito comunicativo destacando-se como fator explicativo mais relevante, superando o impacto do nível de estudo. Individualmente, o propósito apresentou um tamanho de efeito de 0,12, enquanto a interação entre nível de estudo e propósito comunicativo apresentou um tamanho de efeito grande, de 0,18. Diferentemente do observado para os *adjetivos atributivos*, portanto, a interação entre as duas variáveis mostrou-se estatisticamente significativa, indicando que o emprego das *nominalizações* pode variar conforme a combinação entre o estágio formativo do estudante e as demandas comunicativas da tarefa. Esses resultados reforçam a hipótese de que a escrita acadêmica não se desenvolve apenas em função da progressão linguística do estudante, mas também em resposta às exigências discursivas específicas de cada prática de escrita.

Esses achados apresentam convergências e diferenças em relação a pesquisas anteriores. Em Staples et al. (2016), as *nominalizações* mostraram-se significativas para o modelo como um todo, bem como para o nível de estudo, a área disciplinar e a interação entre ambos, embora com tamanho de efeito pequeno. Biber et al. (2014), por sua vez, também incluíram essa característica em seu modelo fatorial, e seus resultados indicam que as *nominalizações* podem constituir uma medida relevante de complexidade, além de se configurarem como um traço distintivo entre diferentes tarefas, tanto em termos de modalidade quanto de tipo.

5. Conclusão

Este estudo investigou o uso de nominalizações e adjetivos atributivos em produções escritas de graduandos de Letras Inglês, buscando compreender em que medida essas características lexicogramaticais variam em função do nível de estudo e dos propósitos comunicativos dos textos. Os resultados evidenciaram tendências gerais de aumento no uso desses recursos ao longo da trajetória acadêmica, embora de forma não linear, o que sugere que o desenvolvimento da complexidade gramatical ocorre de maneira gradual e sensível às demandas de escrita vivenciadas pelos estudantes.

Entre os principais achados, destaca-se o papel dos propósitos comunicativos como variável mais influente do que o nível de estudo na determinação das escolhas linguísticas dos estudantes. Textos orientados à explicação, proposta e revisão apresentaram maior frequência de recursos associados à complexidade frasal, enquanto tarefas de aconselhamento pessoal exibiram padrões menos densos lexicalmente. Esses resultados reforçam a importância de compreender a escrita universitária como reflexo do avanço linguístico dos estudantes, e, ao mesmo tempo, como prática situada, moldada pelas funções comunicativas das tarefas acadêmicas.

Do ponto de vista pedagógico, os resultados indicam a necessidade de práticas de ensino de escrita acadêmica que articulem formas linguísticas e funções comunicativas. Nesse sentido, é importante propor atividades que levem os estudantes a mobilizar diferentes características lexicogramaticais em tarefas com propósitos variados, ampliando sua consciência sobre os usos da linguagem em contextos acadêmicos.

Por fim, ressalta-se que o estudo se baseia em um corpus relativamente restrito e circunscrito ao contexto de um curso de Letras Inglês, o que limita generalizações para outras áreas do conhecimento. Pesquisas futuras podem ampliar o escopo disciplinar e incluir outras características associadas à complexidade gramatical, contribuindo para uma compreensão mais abrangente do desenvolvimento da escrita universitária.

Questões éticas

O estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal do Rio Grande do Sul (UFRGS). Ferramentas de inteligência artificial foram utilizadas como apoio à revisão linguística e à formatação do trabalho. As duas autoras participaram do planejamento da pesquisa e da definição dos instrumentos de coleta. Marine Laisa Matte realizou a coleta dos dados e conduziu sua análise, com a colaboração de Simone Sarmento. Ambas participaram da interpretação dos resultados, da escrita e da revisão do artigo.

Referências

Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes, 31*, 58–71. https://doi.org/10.1016/j.jeap.2017.12.008

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education.

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Bocorny, A. E. P., Rebechi, R., Reppen, R., Delfino, M. C. N., & Lameira, V. M. (2021). A produção de artigos da área das ciências da saúde com o auxílio de key lexical bundles: Um estudo direcionado por corpus. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, 37*, 2021370101.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423.

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252–267). Cambridge University Press.

Gardner, S., & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics, 34*(1), 25–52.

Goulart, L. (2016). Compilation of a Brazilian academic written English corpus. *Revista e-Scrita: Revista do Curso de Letras da UNIABEU, 8*(2), 32–47.

Goulart, L. (2024). *Variation in university student writing*. John Benjamins.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh University Press.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing, 32*, 17–35. https://doi.org/10.1016/j.jslw.2016.02.002

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication, 33*, 149–183. https://doi.org/10.1177/0741088316631527

Staples, S., Gray, B., Biber, D., & Egbert, J. (2023). Writing trajectories of grammatical complexity at the university: Comparing L1 and L2 English writers in BAWE. *Applied Linguistics, 44*(1), 46–71.

A apresentação da pesquisa nas ciências agrárias: a descrição das introduções de artigos científicos como base para a criação de atividades

Deise Dutra[1], Carolina Godoi de Faria Marques[1], Danilo Duarte Costa[2], Gustavo Teixeira[1], Ana Bocorny[3], Shirlene Bemfica de Oliveira[4], Carolina Grondona[1], Juliana Gomes Costa[1], Jasper Vilan Braga[1], Anna Clara Taborda[1], Mariana Gonzaga Silva[1], Marjorie Almeida[1], Katherine Oliva Ortolani[5].

[1] Universidade Federal de Minas Gerais, [2] Universidade Federal dos Vales do Jequitinhonha e Mucuri, [3] Universidade Federal do Rio Grande do Sul, [4] Instituto Federal de Minas Gerais, [5] Universidade Estadual de Montes Claros.

Autor correspondente: deisepdutra@gmail.com

Resumo. *As Ciências Agrárias têm contribuído substancialmente para o desenvolvimento econômico do Brasil, mas as publicações científicas em inglês de pesquisadores brasileiros ainda são limitadas. Com base no CorAgrarian, um corpus de 2,5 milhões de palavras composto por artigos científicos de Ciências Agrárias, este trabalho apresenta resultados de uma análise de atributos-chave das seções de Introdução. A análise identificou características léxico-gramaticais estatisticamente significativas, utilizando valores de d de Cohen, (e.g. perfect aspect d=1,08; modals of possibility d=0,30). O trabalho também apresenta atividades pedagógicas baseadas em corpus desenvolvidas a partir desses resultados, abordando o uso da linguagem acadêmica com vistas a apoiar futuras produções textuais.*

Abstract. *Agrarian Sciences has substantially contributed to Brazil's economic development, but Brazilian scholars' research publications in English remain limited. Based on CorAgrarian, a 2.5-million-word corpus of Agrarian Sciences research articles, this paper reports findings from a key-feature analysis of Introduction sections. The analysis identified statistically significant lexico-grammatical features, using Cohen's d values (e.g. perfect aspect d=1,08; modals of possibility d=0.30). The paper also presents corpus-based pedagogical activities developed from these findings. These activities aim to raise students' awareness of academic language use and support future text production.*

1. Introdução

O Brasil é uma potência global na produção e exportação de alimentos, na pecuária e no desenvolvimento de tecnologias aplicadas ao agronegócio. Entretanto, o número de publicações científicas em inglês produzidas por pesquisadores brasileiros da área de ciências agrárias não faz jus a sua relevância para o país e para o mundo [Teixeira et al. 2022; Dutra et al. 2024b]. Com efeito, essa área do conhecimento tem enorme importância no Brasil por apoiar diretamente o desenvolvimento da agricultura e pecuária há mais de cem anos. A primeira escola de Agronomia no país foi criada em 1875, na

Bahia, e hoje o país conta com quase 400 cursos de graduação na área[1], além de programas de pós-graduação em 85 universidades brasileiras [CAPES 2025].

Há pesquisas publicadas em inglês na área de ciências agrárias por pesquisadores brasileiros, mas o número de publicações não é proporcional à importância da área para o país e para o mundo. Foi pensando nesse contexto e nas necessidades da área que iniciamos o projeto "*Escrita acadêmica em língua inglesa nas Ciências Agrárias: necessidades e insumos para aplicações pedagógicas*" (FAPEMIG nº: APQ-01173-22) para o estudo da escrita acadêmica nas ciências agrárias. Neste sentido, decidimos investigar a linguagem utilizada em artigos científicos (AC), escritos em inglês, publicados em revistas de alto impacto das ciências agrárias, criando o corpus especializado *CorAgrarian* (*Corpus of Agrarian Science Research Articles*).

Há, ainda, poucos estudos que se propõem a investigar a escrita acadêmica das ciências agrárias [Melnikov 2020; Shi e Wannaruk 2014], sendo o principal deles Shi e Wannaruk (2014). Os autores analisaram a estrutura retórica dos ACs dessa área, identificando 16 movimentos retóricos (*moves*) distribuídos ao longo das seções de Introdução, Metodologia, Resultados e Discussão. Os resultados desse estudo apontam para a existência de uma organização retórica específica dessa área que a difere das demais áreas do conhecimento.

Nos últimos anos, tem-se observado uma ampliação nas ferramentas e metodologias de análise de corpora, o que tem viabilizado diferentes tipos de investigação sobre dados linguísticos e exercido impacto direto no planejamento e na elaboração de materiais didáticos voltados ao Inglês para Fins Acadêmicos (IFA) [Thompson 2022]. Abordagens que abrangem desde análises de gêneros textuais [Swales 1990] até análises multidimensionais [Gray 2015] fornecem subsídios para que professores e desenvolvedores de materiais para o ensino de IFA compreendam as estruturas retóricas e os padrões léxico-gramaticais, orientados pelos objetivos comunicativos subjacentes aos textos.

Recentemente, foi proposta uma nova metodologia de análise de corpus, denominada Análise de Atributos-Chave (AAC), ou *Key Feature Analysis* (KFA), a qual permite a detecção de características léxico-gramaticais distintivas de textos com base na medida estatística *d de Cohen* [Egbert e Biber 2023]. Em nosso trabalho, adotou-se a AAC para identificar os itens léxico-gramaticais distintivos dos ACs das subáreas que compõem as ciências agrárias, a saber: Agronomia, Engenharia Agrícola, Engenharia de Alimentos, Engenharia Florestal e Zootecnia. Baseados nos resultados da AAC, criamos atividades que poderão ser usadas em cursos presenciais de inglês da área ou online autonomamente pelos interessados, através de um site em processo de criação.

Neste trabalho, são descritos os resultados referentes à análise da seção Introdução dos ACs, escolhida por sua importância enquanto seção de abertura dos artigos científicos e por sua linguagem apresentar padrões específicos que variam segundo a área de conhecimento ou disciplina [Cortes 2013; Gray, Cotos e Smith 2020; Berber Sardinha et al. 2024]. Assim, estabelecemos os seguintes objetivos: a) apresentar os itens léxico-gramaticais característicos da seção introdução e b) apresentar um exemplo de atividade criada com base nos dados oriundos de a).

---

[1] Dados obtidos do Cadastro Nacional de Cursos e Instituições de Educação Superior (Cadastro e-MEC), disponível em https://emec.mec.gov.br/emec/nova. Acesso em 29/5/2026.

2. Metodologia

Para esta pesquisa, compilamos o *Corpus of Agrarian Sciences Research Articles (CorAgrarian*), um corpus de aproximadamente 2,5 milhões de palavras de ACs da área de ciências agrárias publicados em periódicos de alto impacto em língua inglesa. O corpus foi submetido a uma AAC, que identificou características linguísticas específicas de cada seção dos ACs e subáreas das ciências agrárias. Para este trabalho, utilizamos os resultados que descrevem a escrita da seção de introdução.

A seguir destacamos a composição do *CorAgrarian* (ver Tabela 1). Buscando-se o balanceamento, foram coletados 90 artigos para cada subárea, exceto Agronomia, com um número bastante aproximado. O número de palavras de cada subárea varia entre 459535 (Engenharia de Alimentos) e 526639 (Agronomia). A média de palavras varia entre 6053,32 (Agronomia) e 5105,94 (Engenharia de Alimentos). O corpus como um todo tem cerca de 2,5 milhões de palavras.

Tabela 1. Composição do *CorAgrarian* por subárea

Subcorpus	Nº de artigos	Nº de palavras	Média
Agronomia	87	526639	6053,32
Engenharia Agrícola	90	523245	5813,83
Engenharia de Alimentos	90	459535	5105,94
Engenharia Florestal	90	505454	5616,16
Zootecnia	90	487159	5412,88
TOTAL	447	2502032	5597,39

Fonte: Marques (2024)

Para analisar o corpus, fizemos a AAC [Egbert e Biber, 2023] utilizando o SAS OnDemand for Academics [SAS INSTITUTE INC. 2026]. A AAC funciona bem para detectar diferenças entre registros muito próximos e de textos dentro de um mesmo registro [Berber Sardinha et al. 2024]. A metodologia da AAC, que utiliza o d de Cohen, é capaz de identificar itens linguísticos relevantes e é classificada como uma técnica analítico-exploratória [Egbert e Biber 2023]. A seguir, apresenta-se uma breve explicação sobre o d de Cohen:

> O d de Cohen é uma medida estatística usada para quantificar a diferença padronizada [...] entre dois corpora. Ao invés de simplesmente indicar se a diferença existe, o d de Cohen nos diz quão substancial a diferença é. Ao comparar o Corpus 1 e o Corpus 2 em termos da frequência de um item léxico-gramatical específico (tal como, voz passiva, verbos modais, ou densidade lexical), a pergunta chave é: Este item é mais frequente em um corpus do que no outro? Se sim, a diferença é considerada pequena, moderada, ou grande […] ?" [Grondona e Oliveira, 2026, tradução nossa][2]

[2] No original: "Cohen's d is a statistical measure used to quantify the standardized difference between two groups—in this case, two corpora. Rather than simply indicating whether a difference exists, Cohen's d tells us how substantial that difference is. When comparing Corpus 1 and Corpus 2 in terms of the frequency of a particular lexicogrammatical feature (such as passive voice, modal verbs, or noun density), the key questions are: Is this feature more frequent in one corpus than the other? If so, is the difference considered small, moderate, or large [...]" [Bohórquez e Oliveira, 2026].

Para identificarmos as características específicas das seções dos artigos das ciências agrárias, as seções de introdução constituíram o nosso Corpus 1 e todas as outras seções fizeram parte do Corpus 2.

Para criarmos as atividades, o grupo de pesquisa foi dividido em grupos de 3 membros que seguiram um template para design de atividades para estudos autônomos. Esse template, criado pelo próprio grupo de pesquisa, contém itens, tais como: tipo de exercício (múltipla escolha, *cloze*, preenchimento de lacuna, correção de erros, ligar colunas, etc.), qual(is) elemento(s) léxico-gramaticais foram escolhidos como foco, função(ões) comunicativa(s) envolvidas, título da tarefa e fontes usadas.

3. Resultados

3.1 Resultados da análise do corpus

Os resultados da AAC demonstram que há variação significativa nas características léxico-gramaticais entre as seções dos artigos científicos do CorAgrarian, corroborando estudos anteriores [Gray 2016; Dutra e Berber Sardinha 2018, 2021; Berber Sardinha et al. 2024; Dutra et al. 2024a, Bocorny et al. 2026]. Devido à restrição de espaço, neste trabalho apresentamos uma síntese dos resultados da AAC referente à seção de Introdução.

Como valores de referência do d de Cohen, considera-se d = 0,2, d = 0,5 e d = 0,8 para o tamanho de efeito pequeno, médio e alto, respectivamente [Cohen 1988]. Os resultados da AAC nos permitiu verificar que na seção de Introdução dos artigos do CorAgrarian se destacam os seguintes itens: *perfect aspect* (1,08), *present tense* (0,86), *infinitives* (0,49), *modal verbs of possibility* (0,30); *topical adjectives* (0,34), *evaluative adjectives* (0,40) e *size adjectives* (0,21).

Vale lembrar que a coocorrência de itens linguísticos revela o(s) propósito(s) comunicativo(s) de um texto [Biber 1988]. Em relação ao CorAgrarian, os exemplos (1) e (2) ilustram o uso dos itens linguísticos que coocorrem com a função de comunicar resultados de pesquisas desenvolvidas, relatando suas contribuições para a compreensão do assunto específico abordado nos artigos (uso do aspecto perfeito, tempo presente, verbo modal) e as ações envolvidas no contexto (uso do infinitivo). Para tal, o texto relata as implicações da adoção de certos parâmetros, explicitando suas características (uso de adjetivos).

1) *Research has shown (aspecto perfeito) that FOP improves (tempo presente) consumer's understanding of nutrition information and their ability to identify (infinitivo) healthy (adjetivo atributivo) products. Currently, FOP nutritional schemes have been already implemented (aspecto perfeito) in over 40 countries. Brazil, which was one of the first countries to adopt (infinitivo) mandatory nutrition labeling, started in 2014 the 3rd regulatory intervention to implement (infinitivo) a FOP nutritional (adjetivo) scheme to processed (adjetivo) food products.* (CorAgrarian.Feg.cereal.15_introduction)

2) *Nocturnal temperature inversions may reverse (verbo modal - possibilidade) the usual (adjetivo) temperature gradient, which may impact (verbo modal - possibilidade) forest productivity in mountainous (adjetivo) areas. Low (adjetivo) temperatures on inversion days reduce (tempo presente) respiration in the valley (Novick et al., 2016), but restrict (tempo presente) photosynthesis if inversions*

persist (tempo presente) in the daytime. (CorAgrarian.For.silviculture.09_introduction)

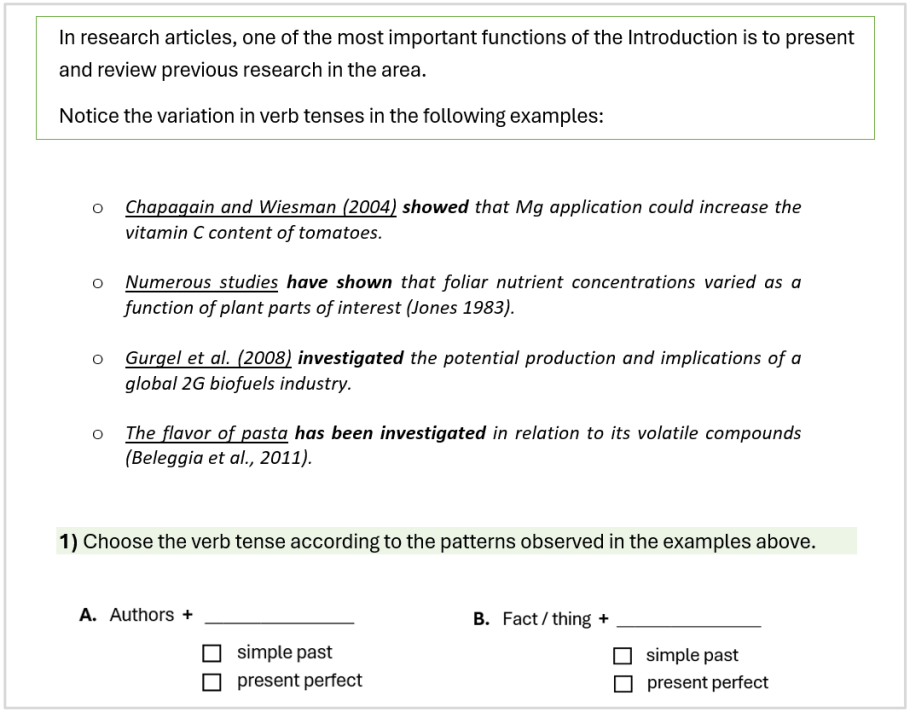
O resultado da AAC revelou que as funções comunicativas dos itens léxico-gramaticais [Biber e Conrad 2009] que coocorrem colaboram para a construção de um frequente movimento retórico presente na seção de Introdução, qual seja, a apresentação e revisão de estudos anteriores [Swales 1990; Feak e Swales 2009; Swales e Feak 2012; Cotos et al. 2015]. Além de variação de tempos verbais na Introdução (e.g. uso dos tempos presente simples, presente perfeito e passado na citações), nossos resultados apontam para a concomitante coocorrência de adjetivos atributivos, (exemplos 1 e 2), que contribuem para o caráter informacional das introduções, e verbos no infinitivo (exemplo 1), que descrevem as ações desenvolvidas em estudos prévios. A AAC detectou, também, como verbos modais (Exemplo 2), em trechos que trazem detalhamentos e possíveis explicações que podem carregar menos assertividade sobre as investigações feitas, colaboram para que as informações sejam logicamente organizadas, levando o leitor a seguir a interpretação dos dados dos estudos previamente publicados da maneira que os autor(es) da introdução escolheram por assim apresentar.

3.2. Proposta de atividade

Utilizando excertos de textos autênticos do CorAgrarian e as abordagens baseadas em corpus e em gêneros textuais, foram desenvolvidas atividades com vistas a auxiliar estudantes e pesquisadores da área de ciências agrárias na compreensão das especificidades linguísticas da escrita acadêmica em inglês. Neste trabalho, apresentamos uma amostra dessas atividades, tratando especificamente da seção de Introdução de artigos científicos. Com as atividades, espera-se que os alunos possam identificar e compreender que os itens léxico-gramaticais presentes na escrita acadêmica variam conforme a função retórica que desempenham no discurso.

No desenvolvimento das atividades, levou-se em consideração o fato de que o ensino de línguas para falantes não nativos precisa estar inserido em um contexto real de uso da língua, sobretudo no que tange o ensino/aprendizagem de questões gramaticais. Como observa Larsen-Freeman (2001), é importante considerar como um determinado gênero ou registro de discurso afeta o uso de uma estrutura sintática. A autora destaca que o aprendiz deve ser conduzido a refletir sobre questões como: "Quando ou por que um falante/escritor varia a forma de uma determinada estrutura linguística?" [Larsen-Freeman 2011, p. 252].

Nesse sentido, dado que o tempo/aspecto verbal *present perfect* destacou-se entre os itens léxico-gramaticais identificados na AAC, criamos uma atividade que aborda seu uso nas seções de Introdução. Mais especificamente, a atividade explora a variação entre os tempos *present perfect* e *simple past* sendo empregados em um movimento retórico comumente observado na seção de Introdução, qual seja: apresentar e revisar pesquisas anteriores, como também demonstrado por Swales e Feak (2012). Cabe esclarecer que as atividades propostas serão disponibilizadas em ambiente virtual, no formato de *self-study activities*, com correção e feedback automáticos. Para fins de ilustração, neste trabalho as atividades foram adaptadas para o formato analógico impresso, como mostrado na Figura 1.


In research articles, one of the most important functions of the Introduction is to present and review previous research in the area.

Notice the variation in verb tenses in the following examples:

- *Chapagain and Wiesman (2004)* ***showed*** *that Mg application could increase the vitamin C content of tomatoes.*
- *Numerous studies* ***have shown*** *that foliar nutrient concentrations varied as a function of plant parts of interest (Jones 1983).*
- *Gurgel et al. (2008)* ***investigated*** *the potential production and implications of a global 2G biofuels industry.*
- *The flavor of pasta* ***has been investigated*** *in relation to its volatile compounds (Beleggia et al., 2011).*

1) Choose the verb tense according to the patterns observed in the examples above.

A. Authors + _____
- ☐ simple past
- ☐ present perfect

B. Fact / thing + _____
- ☐ simple past
- ☐ present perfect

Figura 1. Amostra de atividade didática de conscientização linguística desenvolvida com base nos dados do CorAgrarian.

A atividade ilustrada na Figura 1 objetiva a conscientização (*consciousness-raising*) a respeito de como a escolha entre dois tempos verbais está diretamente ligada não só à estrutura da frase (o pesquisador ocupando a posição de sujeito da oração), mas também à função retórica que eles desempenham no discurso.

Trata-se, portanto, de um exercício com foco na aprendizagem baseada em dados, ou *data-driven learning* (DDL). Como definem Pérez-Paredes e Boulton (2025), a DDL refere-se a uma abordagem pedagógica no ensino de línguas que utiliza recursos e métodos da Linguística de Corpus para auxiliar os aprendizes a observar e compreender padrões de uso da língua em diferentes gêneros textuais. Os autores destacam, ainda, que a DDL permite que os professores criem ou adaptem materiais didáticos com base no uso autêntico da língua, tornando as aulas mais relevantes e envolventes para os alunos.

Na sequência, a variação entre os tempos verbais *simple past* e *present perfect* é apresentada em uma nova atividade, ilustrada na Figura 2. Nesse exercício, os alunos leem a seção de Introdução de um artigo científico que aborda o tema ‘gestão de enchentes’, extraído do CorAgrarian. No texto, as ocorrências dos tempos verbais em questão foram substituídas por lacunas. Com base na estrutura das orações e na função retórica dos tempos verbais, já apresentados na atividade anterior, deve-se escolher qual alternativa apresenta a forma verbal adequada para preencher cada uma das lacunas. Esse exercício foi pensado com o objetivo de praticar e consolidar o conhecimento adquirido sobre o tema.

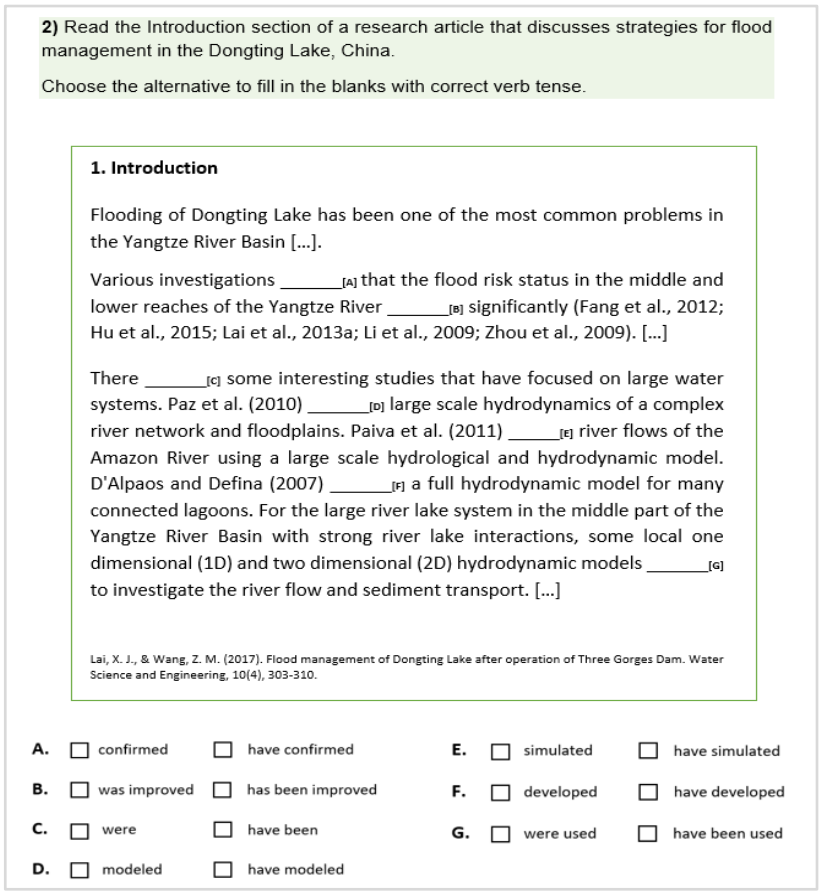
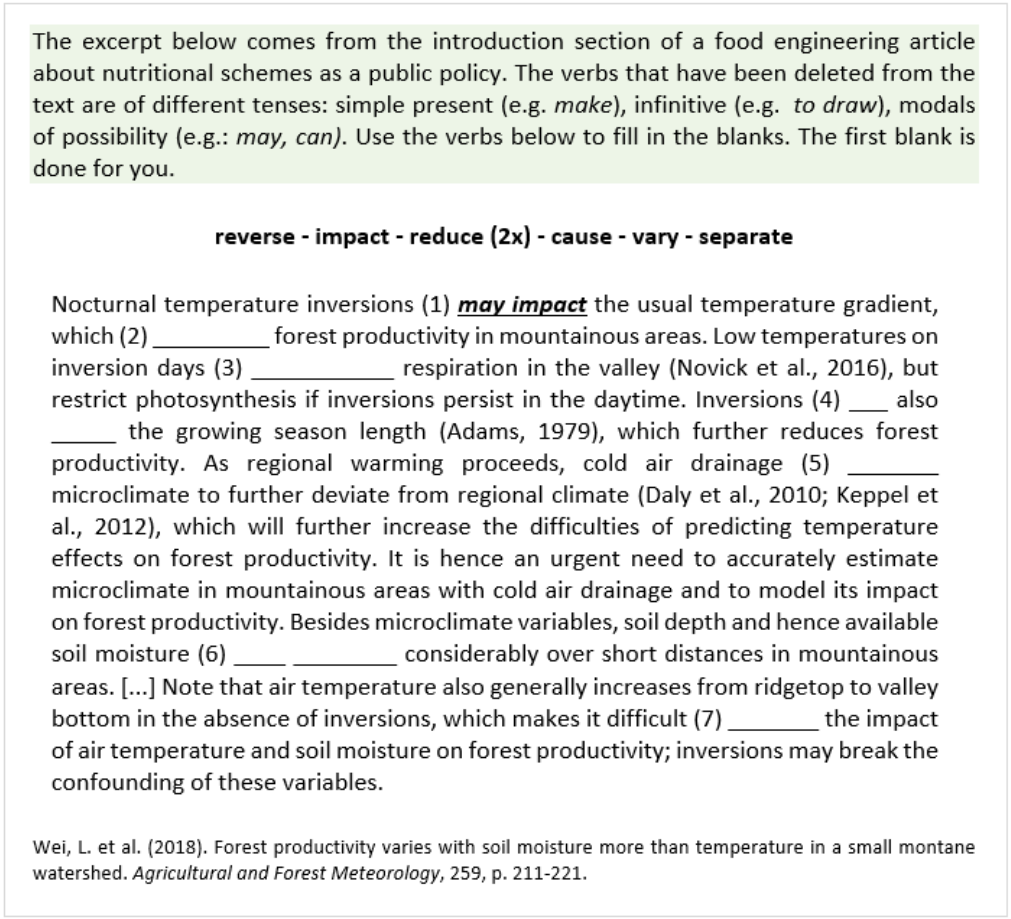


Figura 2. Amostra de atividade de fixação de conteúdo linguístico desenvolvida com base nos dados do CorAgrarian.

A atividade a seguir (Fig. 3) apresenta três formas verbais coocorrendo no mesmo trecho (presente simples, infinitivo e modais de possibilidade), demonstrando que na introdução de AC é comum que várias formas verbais sejam usadas. Esse foi um dos resultados marcantes de nossa AAC, pois demonstra que as formas coocorrem a fim de expressarem importantes funções comunicativas na seção de introdução. Ao passo que o uso de modais de possibilidade expressam fatos que podem ou não ocorrer, o presente simples apresenta algo estabelecido na área e o infinitivo carrega o sentido de uma ação.

Em síntese, essas atividades buscam, por meio da observação de padrões linguísticos e discursivos característicos da escrita acadêmica, fornecer ao estudante o conhecimento necessário para reconhecer e compreender as diferentes funções desempenhadas pelo itens léxico-gramaticais presentes na seção de Introdução de artigos científicos no campo das ciências agrárias. Espera-se, com isso, que esse conhecimento possa servir como apoio à produção escrita acadêmica.



The excerpt below comes from the introduction section of a food engineering article about nutritional schemes as a public policy. The verbs that have been deleted from the text are of different tenses: simple present (e.g. *make*), infinitive (e.g. *to draw*), modals of possibility (e.g.: *may, can)*. Use the verbs below to fill in the blanks. The first blank is done for you.

reverse - impact - reduce (2x) - cause - vary - separate

Nocturnal temperature inversions (1) ***may impact*** the usual temperature gradient, which (2) _____ forest productivity in mountainous areas. Low temperatures on inversion days (3) _____ respiration in the valley (Novick et al., 2016), but restrict photosynthesis if inversions persist in the daytime. Inversions (4) ___ also _____ the growing season length (Adams, 1979), which further reduces forest productivity. As regional warming proceeds, cold air drainage (5) _____ microclimate to further deviate from regional climate (Daly et al., 2010; Keppel et al., 2012), which will further increase the difficulties of predicting temperature effects on forest productivity. It is hence an urgent need to accurately estimate microclimate in mountainous areas with cold air drainage and to model its impact on forest productivity. Besides microclimate variables, soil depth and hence available soil moisture (6) ____ _____ considerably over short distances in mountainous areas. [...] Note that air temperature also generally increases from ridgetop to valley bottom in the absence of inversions, which makes it difficult (7) _____ the impact of air temperature and soil moisture on forest productivity; inversions may break the confounding of these variables.

Wei, L. et al. (2018). Forest productivity varies with soil moisture more than temperature in a small montane watershed. *Agricultural and Forest Meteorology*, 259, p. 211-221.

Figura 3. Amostra de atividade didática com coocorrência de itens linguísticos

4. Conclusão

Essas descobertas subsidiaram o desenvolvimento de atividades baseadas em corpus [Flowerdew 2012; Bocorny e Welp 2021] e na abordagem de gênero [Swales 1990] para aprimorar a escrita acadêmica em inglês nas Ciências Agrárias e apoiar a internacionalização da pesquisa brasileira, aumentando sua visibilidade e impacto.

A proposta busca facilitar e desenvolver conhecimentos linguísticos no contexto da escrita acadêmica por meio da observação de padrões sintáticos e retóricos, fundamentados nos princípios da linguística de corpus e da abordagem de gêneros acadêmicos. Ao analisar exemplos autênticos retirados de artigos científicos, os estudantes são incentivados a compreender como a linguagem é utilizada para cumprir objetivos comunicativos específicos na construção do conhecimento científico.

Agradecimentos

Os autores agradecem à FAPEMIG pelo financiamento do projeto de pesquisa APQ-01173-22, bem como à CAPES pelas bolsas de pesquisa n.º 88887.939578/2024-00 e 88881.126398/2025-01.

Referências

Berber Sardinha, T., Dutra, D. P., Delfino, C. N., Bocorny, A., Shimazumi, M., Kauffmann, C., Macedo, L. D., Victor, A. C. (2024). Key feature analysis of academic writing in the humanities. In: *International American Association for Corpus Linguistics Conference*, página 119. Anais eletrônicos.

Bocorny, A., Dutra D. P., Acorinti, W., Soares, K., Teixeira, G., Costa D. D., Marques, C. G. F., Bohórquez, C. (2026) "Lexical bundle families in Agrarian Sciences research articles", In: Revista Intercâmbio, v.LVIII, e76980. https://doi.org/10.23925/2237-759X2026V58e76980

Bocorny, A., Welp, A. (2021) "O desenho de tarefas pedagógicas para o ensino de Inglês para Fins Acadêmicos: conquistas e desafios da Linguística de Corpus", In: Revista de Estudos da Linguagem, Faculdade de Letras da UFMG, Belo Horizonte, p. 1529–1638.

CAPES. (2025) "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Plataforma Sucupira: Dados Cadastrais do Programa", https://sucupira-legado.capes.gov.br/sucupira/public/consultas/coleta/programa/listaPrograma.jsf, Setembro.

Cohen, J. (1988) Statistical power analysis for the behavioral sciences, Laurence Erlbaum Associates, 2ª edição.

Cortes, V. (2013), "The purpose of this study is to: Connecting lexical bundles and moves in research article introductions", In: Journal of English for Academic Purposes, Elsevier, p. 33–43.

Cotos, E., Huffman, S., Link, S. (2015) "Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy", In: Journal of English for Academic Purposes, Editado por Z. Lancaster, L. Aull e M. D. Perales-Escudero, Elsevier, p. 52–72

Dutra, D. P., Berber Sardinha, T. (2018). A linguistic typology of sections in research articles. A Multi-Dimensional perspective. *In:* In: *International American Association for Corpus Linguistics Conference.* Apresentação oral.

Dutra, D. P., Berber Sardinha, T. (2021). A multi-dimensional typology of English research articles sections. *In:* In: *International American Association for Corpus Linguistics Conference.* Online.

Dutra, D. P., Berber Sardinha, T., Delfino, C. N., Bocorny, A. E. P., Shimazumi, M., Kauffmann, C., Macedo, L. D. de, Taborda, A. C. (2024a). Key Feature Analysis of Academic Writing in the Humanities. In: *Semana de Eventos da Faculdade de Letras da UFMG.* Apresentação oral.

Dutra, D. P., Bocorny, A. E. P., Costa, D. D., Teixeira, G. L., Marques, C. G. F. (2024b) Desafios metodológicos na compilação do corpus de textos acadêmicos das Ciências Agrárias. In: *XVI Encontro de Linguística de Corpus e da XII Escola Brasileira de Linguística Computacional.* Anais eletrônicos.

Egbert, J., Biber, D. (2023) "Key feature analysis: a simple, yet powerful method for comparing text varieties", In: Corpora, Edinburgh University Press, p. 121–133.

Flowerdew, L. (2012), Corpora and language education, Palgrave Macmillan.

Gray, B. (2015), Linguistic Variation in Research articles: When Discipline Tells Only Part of the Story, John Benjamins.

Gray, B., Cotos, E., Smith, J. (2020) "Combining rhetorical move analysis with multi-dimensional analysis: Research writing across disciplines", Advances in Corpus-based Research on Academic Writing: Effects of discipline, register, and writer expertise, U. Römer, V. Cortes e E. Friginal, John Benjamins Publishing Company, p. 137–168.

Grondona, C. B., Oliveira, L. A. (2026) "Key-feature Analysis for Dummies: a comprehensive step-by-step guide for novice researchers", In: Revista Letras & Letras. P 1-22

Larsen-Freeman, D. (2001) "Teaching Grammar", Teaching English as a second or foreign language, M. Celce-Murcia,. Heinle & Heinle, p. 251–266.

Marques, C. G. de F. (2024). CorAgrarian: um corpus de artigos científicos de ciências agrárias. In: *Semana de Eventos da Faculdade de Letras da UFMG*. Apresentação oral.

Martínez, I. A., Beck, S. C., Panza, C. B. (2009) "Academic vocabulary in agriculture research articles: A corpus-based study", In: English for Specific Purposes, Elsevier, p. 183–198.

Melnikov, M. (2020) "Linguistic corpus as a means of adaptation of modern scientific agricultural approaches", In: BIO Web of Conferences.

Pérez-Paredes, P., Boulton, A. (2025) "Data-driven Learning in and out of the Language Classroom", In: Elements in Language Teaching, Cambridge University Press, Cambridge.

SAS INSTITUTE INC. (2026) "SAS OnDemand for Academics", Versão 9.4. Cary, NC: SAS Institute Inc. Disponível em: https://www.sas.com/pt_br/software/on-demand-for-academics.html.

Shi, H., Wannaruk, A. (2014) "Rhetorical Structure of Research Articles in Agricultural Science", In: English Language Teaching, Canadian Center of Science and Education, p. 1–13.

Swales, J. M. (1990), Genre Analysis: English in Academic and Research Settings, Cambridge University Press.

Teixeira, G. Bocorny, A. E. P., Dutra, D. P. (2022). Mapping the language needs of the Brazilian researchers in the Agrarian Sciences: a corpus-based study of abstracts.In: *International American Association for Corpus Linguistics Conference*. Anais eletrônicos.

Thompson, P. (2022) "Corpus analysis of disciplinary variation and the teaching of ESP/EAP", The Routledge handbook of corpora and English language teaching and learning, R. R. Jablonkai e E. Csoma, New York, Routledge, p. 177–192.

O corpus C-ORAL-BRASIL II

Bruno Rocha, Tommaso Raso

Universidade Federal de Minas Gerais

Abstract. *This article presents and describes the C-ORAL-BRASIL II corpus, a corpus of formal spoken Brazilian Portuguese in natural contexts, media, and telephone. It reports on the corpus architecture and its main measurements, outlines the compilation protocol, and presents the tables and other metrics provided with the corpus. This information is accompanied by considerations that assist the reader in understanding the resource's characteristics.*

Resumo. *Este artigo apresenta e descreve o corpus C-ORAL-BRASIL II, corpus de português brasileiro falado formal em contexto natural, mídia e telefone. São relatadas a arquitetura do corpus e as suas principais medidas, é exposto o protocolo de compilação, são apresentadas as tabelas e outras medidas fornecidas junto ao corpus. Essas informações são acompanhadas por considerações que ajudam o leitor na compreensão das características do recurso.*

1. Introdução

Neste artigo apresentamos a primeira descrição do corpus C-ORAL-BRASIL II [Raso, Mello, Rocha e Ferrari 2026] (OSF DOI: 10.17605/OSF.IO/W94JT), disponibilizado para download gratuito no endereço http://www.c-oral-brasil.org/c-oral-brasil-ii.php sob a licença CC BY-NC-SA 4.0. Neste endereço podem ser baixados os arquivos de áudio (.wav), as transcrições (.rtf e .txt), os alinhamentos nos softwares WinPitch [Martin 2004] e Elan [Witterburg et al. 2006] (.xml e .eaf), metadados (.txt) e as anotações léxico-morfossintáticas (.syn e .dep) feitas com o software Palavras [Bick 2000] adaptado para o C-ORAL-BRASIL [Bick 2012]. Além disso, é possível baixar tabelas com informações relativas aos falantes e a diversas medidas do corpus.

Trata-se de um corpus de fala espontânea do português brasileiro (PB), gravado entre 2012 e 2015, que inclui três subcorpora: um de fala formal em contexto natural, um de mídia (televisão e rádio) e um de fala telefônica. São, portanto, representadas quatro diamesias: a interação face-a-face e, aquelas mediadas por televisão, rádio e telefone. O C-ORAL-BRASIL II integra o C-ORAL-BRASIL I [Raso e Mello 2012; Raso 2012; Mello 2014], para a fala informal do PB. Os corpora são comparáveis por arquitetura e proporções aos corpora C-ORAL-ROM [Cresti e Moneglia 2005] para espanhol, francês, italiano e português europeu, com algumas pequenas mudanças indicadas abaixo e um aumento de tamanho em cerca de 50%.

O conjunto de corpora C-ORAL-BRASIL apresenta também uma ampla variação diafásica. O C-ORAL-BRASIL I é dividido igualmente em três tipologias interacionais: monólogos, diálogos e conversações. As duas últimas retratam uma grande variação situacional, porque os falantes quase sempre interagem enquanto estão realizando atividades variadas, inclusive em movimento (compras em supermercado, jogando futebol, tendo aula de direção, fazendo faxina, etc.); pouquíssimos são os bate-papos sem alguma atividade e as entrevistas. No monológico, os assuntos são muito variados

(histórias de vida, contos, relatos profissionais, receitas, etc.). No C-ORAL-BRASIL II, os subcorpora de contexto natural e mídia são divididos em gêneros, seguindo a estrutura do C-ORAL-ROM, por sua vez inspirado na subdivisão do Spoken Dutch Corpus [Boves e Oostijk 2003]. O subcorpus telefônico é dividido em interações públicas e privadas. Quando possível, as gravações pertencentes a cada subcorpus foram classificadas segundo as três tipologias interacionais, mas sem pretender um equilíbrio entre elas dentro de cada gênero e no corpus. Nem todos os gêneros permitem essa distinção (em especial, alguns do subcorpus de mídia).

2. As medidas gerais

O corpus C-ORAL-BRASIL II totaliza 290.565 palavras e 254 textos, assim distribuídos: 120.738 palavras e 74 textos no formal em contexto natural; 138.582 palavras e 101 textos em mídia; e 31.245 palavras e 79 textos no telefônico. Se somamos esses números àqueles do corpus informal (208.130 palavras e 139 textos), conseguimos um conjunto de corpora de PB com 498.695 palavras e 393 textos.

O subcorpus de contexto natural é dividido em 8 gêneros: 16.071 palavras e 18 textos em *political speech*; 16.995 palavras e 9 textos em *conference*; 16.134 palavras e 9 textos em *law*; 15.592 palavras e 9 textos em *political debate*; 12.777 palavras e 9 textos em *preaching*; 16.105 palavras e 8 textos em *professional explanation*; 16.071 palavras e 8 textos em *teaching*; 10.993 palavras e 4 textos em *business*.

O subcorpus de mídia é dividido em 7 gêneros e uma seção extra: 23.455 palavras e 29 textos de *reportage*; 43.689 palavras e 18 textos de *talk show*; 13.139 palavras e 12 textos de *scientific press*; 15.250 palavras e 9 textos de *interview*; 6.041 palavras e 9 textos de *news*; 12.162 palavras e 7 textos em *sports*; 232 palavras e 1 texto em *metheorology*; 24.614 palavras e 16 textos de extra, que recolhem textos que extrapolam as as proporções do C-ORAL-ROM. Apenas 50% do corpus de mídia foi publicado porque somente a Rede Globo deu a autorização. Até agora não conseguimos um contato com as outras emissoras.

O subcorpus de telefone contém 31.245 palavras e 79 textos, divididos em interações públicas (5.717 palavras e 29 textos) e privadas (25.528 palavras e 50 textos). No C-ORAL-ROM também há uma seção de interações homem-máquina, que não foi incluída no C-ORAL-BRASIL II, devido às mudanças tecnológicas intercorridas no tempo. Pela mesma razão, o subcorpus telefônico do C-ORAL-BRASIL apresenta muitas gravações em que um dos falantes usa um telefone celular.

As medidas dos textos variam muito: o maior é um texto de contexto natural de 24:03 minutos; o menor é uma gravação telefônica de 5 segundos. Os textos telefônicos são os mais curtos; o maior dura 5:24 minutos. Os textos de contextos natural e mídia variam entre pouco mais de 24 e 1:08 minutos.

3. Os critérios de compilação

Todas as gravações do corpus foram registradas em formato .wav com taxa de amostragem de 22.050 Hertz, com gravadores digitais de alta resolução (MARANTZ PMD660 ou TASCAM DR-100). As gravações do corpus telefônico e de mídia, por restrições dos aparelhos usados, possui um único canal de áudio (mono). No corpus de fala formal em contexto natural, a maior parte das gravações tem dois canais (estéreo).

As gravações de contexto natural foram feitas ligando microfones de lapela

omnidirecionais (Sennheiser ME 2) ao gravador por meio de *transmitters* e *receivers* sem fio que operam via rádio (Sennheiser Evolution EW100 G3). Nas gravações com mais de dois falantes, foi usada uma mesa de som (Behringer XEXYX 1222 FX). As gravações de mídia foram realizadas conectando a saída auxiliar de aparelhos de televisão e de rádio à entrada P2 dos gravadores. Para as gravações telefônicas, foi usado um aparelho de escuta que era conectado entre um telefone fixo e sua linha telefônica. O aparelho possuía uma saída de áudio em formato mono, ligada ao gravador, como no esquema a seguir.

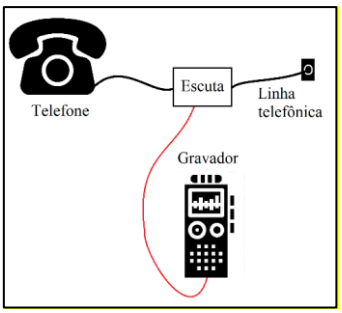


Figura 1. Equipamentos usados no subcorpus telefônico.

Um script de Praat foi desenvolvido para auxiliar na avaliação da qualidade acústica das gravações [Vieira, em preparação]. O script realiza a amostragem das gravações e fornece ao avaliador inputs visuais e auditivos para analisar a audibilidade e a adequação das medidas espectrais com relação a 4 parâmetros (ruído, sobreposição de fala, formantes e f0) classificados individualmente em cinco níveis (A, AB, B, BC, C), a partir dos quais é calculada uma nota geral (também de A a C). Intervalos de dez segundos são extraídos como amostras de trechos delimitados por pontos equidistantes no tempo. Por exemplo, para uma gravação de 1 minuto, seriam coletadas 3 amostras correspondentes aos primeiros 10 segundos dos intervalos 0-20, 20-40 e 40-60. O número de trechos aumenta proporcionalmente à duração das gravações. Para cada amostra, o avaliador estabelece inicialmente os intervalos de fala e de silêncio (caso existam), a partir dos quais a relação sinal-ruído é calculada, como uma diferença de intensidade (A: >= 20 dB; AB: >= 18 dB; B: >= 15 dB; BC: >= 10 dB; C: < 10 dB). No que concerne à f0 e aos formantes (F1 e F2), o avaliador atesta a adequação da estimação por meio da inspeção visual do rastreamento dos parâmetros e da similaridade entre o som original e uma versão sintetizada, baseada exclusivamente na estimação do Praat (para f0: erros de cálculo receberam peso 0.5 e a presença de candidatos corretos receberam peso 1; para formantes: visualização recebeu peso 3 e coerência do som recebeu peso 1). Por fim, com base nas transcrições, contabilizam-se os trechos de sobreposição de vozes ou ininteligíveis (A: <= 5%; AB: <= 12%; B: <= 20%; BC: <= 30%; C: > 30%). Ao término do processo, fundamentado em pesos previamente definidos para cada parâmetro e que dependem da finalidade, o script calcula uma média ponderada arredondada. Para o C-ORAL-BRASIL II, esses valores foram aplicados em uma fórmula de média ponderada (com divisor 7), cujos pesos foram distribuídos a depender da relevância de cada parâmetro. A frequência fundamental e formantes receberam pesos maiores (3 e 2, respectivamente), por serem primordiais em estudos fonéticos e fonológicos. Já a relação sinal-ruído e sobreposição receberam peso 1. 74 áudios foram classificados como qualidade A, 92 como AB, 63 como B, 10 como BC e 15 como C.

Todas as gravações são acompanhadas por um arquivo de metadados que fornece as informações relativas à gravação, à transcrição e aos falantes. A seguir colocamos como exemplo os metadados da gravação de uma conferência.

@Title: A cultural action forum
@File: bnatco05
@Participants: ID 290, VIV (female, B, 3, librarian, participant, Belo Horizonte/MG)
ID 289, GIL (female, C, 3, librarian, participant, Belo Horizonte/MG)
@Date: 17/10/2014
@Place: UFMG School of Information Science, Belo Horizonte/MG
@Situation: a forum on cultural initiatives
@Topic: organization and execution of the Mobile library project
@Source: C-ORAL-BRASIL
@Class: formal, natural context, conference, monologue
@Length: 4'13"
@Words: 507
@Acoustic_quality: general: B, silence_noise_ratio: AB, f0: AB, formant: BC, overlapping: A
@Transcriber: Angelo Coutinho
@Revisor: Bárbara Figueiredo, Bárbara Teixeira, Saulo Santos, Livia Winkler, Giulia Bossaglia, Tayane Soares
@Comments: Apheretic forms: tá (está).
Standardized forms: pr' (para), pra (para), ocê (você), cê (você).
Abbreviations: cedês (CDs: Compact Discs).
VIV stops speaking and asks GIL for help between 35" and 55", GIL laughs at 92".

Como todos os corpora da família C-ORAL, o C-ORAL-BRASIL II adota critérios específicos de transcrição e segmentação prosódica. Os critérios de transcrição são de base ortográfica e seguem o sistema CHILDES [MacWhinney 2000], mas algumas línguas, principalmente italiano e PB, introduzem critérios não ortográficos para capturar fenômenos da fala candidatos a processos de gramaticalização e lexicalização, como a perda de morfologia verbal ou da marca de plural, a redução das formas pronominais e preposicionais e as formas aferéticas. Para al ista dos fenômenos anotados e das suas motivações, vejam-se Mello e Raso (2009) e Mello et al. (2012).

A fala é segmentada prosodicamente em unidades terminadas e unidades prosódicas. A primeira, marcada com a dupla barra oblíqua (//), se define como o menor trecho de fala dotado de autonomia pragmática e prosódica. As unidades prosódicas internas são marcadas com uma barra oblíqua (/). Para uma discussão sobre esse tipo de segmentação prosódica, veja-se Izre'el et al. (2020). A transcrição CHILDES é portanto integrada com a anotação prosódica [Moneglia e Cresti 1997].

A transcrição anota também diversos fenômenos de disfluência, tanto prosódica quanto segmental. São marcadas as unidades interrompidas (unidades abandonadas por algum motivo interno ou externo ao falante) com o símbolo "+" e as retratações (repetições, correções, mudanças de planejamento) com uma barra simples seguida do número de palavras retratadas dentro de colchetes "[/n]". No plano segmental, são marcadas as palavras interrompidas transcrevendo a parte pronunciada precedida pelo símbolo *&*, enquanto as pausas preenchidas são anotadas como *&he*, independentemente do som com o qual são pronunciadas.

As transcrições e as segmentações prosódicas foram realizadas por alunos de graduação e pós-graduação que passaram por um processo de formação de pelo menos um semestre. Antes de começar as transcrições, os alunos foram submetidos a um teste

Kappa [Fleiss 1971] para verificar o grau de acordo com segmentadores experts. Os alunos começaram a transcrição somente depois de terem alcançado um acordo mínimo de 0.8. Os textos passaram por pelo menos 2 revisões (normalmente mais, e até 5 vezes) pelos segmentadores com Kappa mais alto (para maiores detalhes sobre o protocolo de anotação e avaliação, veja Mello et al. 2012). Texto e áudio foram alinhados em WinPitch e Elan [Martin 2004]. A unidade de alinhamento é a unidade terminada.

4. Características dos falantes e dos textos

Várias tabelas informam sobre o corpus e suas partes, e sobre os falantes.

4.1. Os falantes

Nos metadadossão indicadas, entre outras informações, as características dos falantes: sexo, idade, escolaridade, ocupação e origem. O sexo é indicado com as letras M, F ou X, nos poucos casos em que não foi possível tomar uma decisão; a idade é indicada com as letras A, entre de 18 e 25 anos, B, entre 26 e 40 anos, C, entre 41 e 60 anos, D, com mais de 60 anos, M, para menor de idade, e X, quando a idade é desconhecida (frequente no corpus de mídia); a escolaridade é indicada com os números 1, até o primário completo, 2, até o terceiro grau, se o título não é necessário para o trabalho exercido, 3, se o falante completou o terceiro grau e seu trabalho requer o título; a letra X indica a falta dessa informação (também mais frequente no corpus de mídia). O quadro dos falantes com base nessas informações é explorado brevemente a seguir.

Tabela 1. Número de falantes no C-ORAL-BRASIL II e seus subcorpora.

	C-ORAL-BRASIL II	Contexto natural	Mídia	Telefônico	Extra
Falantes diferentes	743	158	393	121	87
Total de falantes	916	185	473	169	90

A figura a seguir mostra o percentual de palavras produzido em cada subcorpus em função de sexo, idade e escolaridade. Na primeira linha de gráficos, estão os dados relativos ao corpus completo, em que é possível observar: (i). a prevalência de palavras de homens (59%) sobre a de mulheres (41%); (ii). a predominância de palavras de pessoas entre 41 e 60 anos (faixa C, com 33%) e 26 e 40 anos (faixa B, com 23%) sobre as demais; (iii). maior percentual de palavras de pessoas com terceiro grau completo e que exercem a função relativa à sua formação (faixa 3, com 64% das palavras) com relação à de pessoas com escolaridade mais baixa.

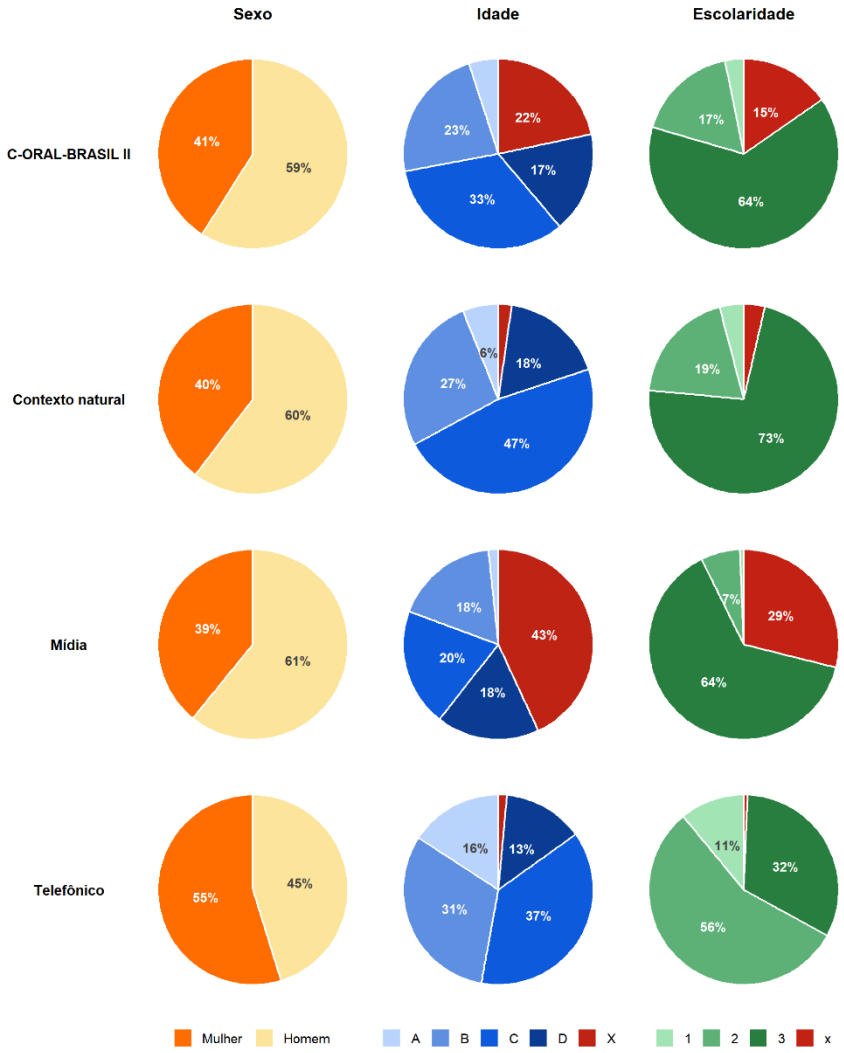


Figura 2. Percentual de palavras produzidas por falantes em função do sexo (esquerda), idade (centro) e escolaridade (direita) no C-ORAL-BRASIL II e em seus subcorpora.

Nota: Nos gráficos, foram omitidos valores inferiores a 5%.

A distribuição de palavras por sexo, idade e escolaridade é bastante heterogênea nos diferentes subcorpora. Por restrições de espaço, não será possível fazer uma análise pormenorizada, mas chamamos a atenção para algumas características. Primeiramente vale notar que nos subcorpora de contexto natural e mídia são especialmente frequentes as palavras produzidas por pessoas com a maior escolaridade (faixa 3, com 73% e 67%, respectivamente), o que se reflete na alta escolaridade do corpus como um todo. No subcorpus telefônico, por outro lado, que possui seja gravações em domínio público, seja gravações em domínio familiar/privado, predominam as palavras de pessoas na faixa intermediária de escolaridade (faixa 2, com 56% das palavras). Além disso, o subcorpus telefônico é o único que possui maior percentual de palavras produzidas por mulheres com relação a homens (55% e 45%, respectivamente). Por fim, no subcorpus de mídia, é notável o percentual de palavras produzidas por falantes cuja idade e escolaridade não são registradas, o que se deve ao método de coleta das gravações (sem contato direto com

os participantes). Isso faz com que também o C-ORAL-BRASIL II como um todo tenha muitas palavras produzidas por pessoas sem a informação de idade (22%) e escolaridade (15%) registradas. Para o contexto natural e o telefônico, esses percentuais são sempre inferiores a 4%.

O corpus também registra a origem dos falantes, que é prevalentemente mineira, em particular da área metropolitana de Belo Horizonte, para os subcorpora de contexto natural e telefônico. Para mídia, a origem é mais variada, mas muito centrada na região sudeste.

Tabela 2. Percentual de palavras produzidas em função da origem do falante.

Região	C-ORAL-BRASIL II	Contexto natural	Mídia	Extra	Telefônico
Norte	3%	5%	2%	0%	0%
Nordeste	6%	6%	9%	8%	4%
Centro-Oeste	5%	6%	4%	5%	1%
Sudeste	81%	85%	76%	74%	87%
Sul	4%	0%	9%	15%	1%
Estrangeiro	1%	0%	2%	0%	1%
Base de cálculo	235.096	115.662	73.503	14.689	31.242

Nota: A base de cálculo inclui somente o número de palavras produzido por falantes com origem registrada. Por esse motivo, os totais do corpus e dos subcorpora são sempre inferiores aos registrados na tabela 1.

4.2. O corpus e os textos

Uma tabela geral disponibilizada no site do projeto fornece, com muito mais detalhes, as informações já expostas nas seções anteriores. A partir dessa tabela é possível saber, para cada subcorpus e cada domínio interno a eles, o número de textos, de palavras, diversas medidas de tamanho (palavras por turno, palavras por unidade terminada, palavras por unidade prosódica, unidades prosódicas por unidade terminada, unidades terminadas por turno), de duração (tempo total, palavras por segundo, unidades prosódicas por segundo, unidades terminadas por minuto) e de disfluências (unidades interrompidas por unidades terminadas, unidades prosódicas retratadas, palavras retratadas, palavras interrompidas, unidades prosódicas com pausas preenchidas). Todas essas medidas vêm com média, mediana, desvio padrão, mínimo e máximo. Uma outra tabela fornece essas informações para cada texto. Todas essas informações nos permitem ter uma ideia das características de cada tipologia textual e de cada texto.

5. Os minicorpora etiquetados

Um corpus tratado como o C-ORAL-BRASIL permite estudos de caráter lexical e morfossintático, facilitados pelos critérios de transcrição que capturam um grande

número de formas não padrão. Mas a característica mais original dos corpora da família C-ORAL é permitir pesquisas pragmáticas. Isso é possível pela arquitetura geral do corpus, que valoriza muito a variação diafásica, mas principalmente pela segmentação da fala, que permite estudos em interface com a prosódia.

De fato, a segmentação da fala em unidades terminadas e unidades prosódicas permite a identificação e quantificação imediata das menores unidades pragmática e prosodicamente autônomas e da sua estruturação interna. As unidades terminadas apresentam pelo menos um padrão prosódico constituído por um núcleo ilocucionário e eventuais outras unidades prosódicas com função não ilocucionária. Os padrões podem ser separados por fronteiras terminais ou não terminais. A segmentação em unidades prosódicas é importante por pelo menos duas razões: ela constitui o domínio das relações sintáticas estreitas; de fato, as fronteiras prosódicas se comportam como um recurso que interrompe a composicionalidade sintática, menos nos casos em que elas são devidas a disfluência.

Muitos estudos realizados a partir dos corpora da família C-ORAL dizem respeito às ilocuções e à estrutura informacional (citamos apenas alguns entre os mais recentes: Cresti e Moneglia 2023; Raso et al. 2025; Cavalcante et al. 2023). A maioria desses estudos se inspira na *Language into act theory* [L-AcT; Cresti 2000; Moneglia e Raso 2014; Cresti e Moneglia 2018].

Para facilitar ainda mais esses estudos, os grupos que realizam os corpora da família C-ORAL compilam também subcorpora etiquetados informacionalmente de maneira manual e disponibilizados gratuitamente e integralmente no site do projeto. A partir do C-ORAL-BRASIL, até agora foram realizados os minicorpora de fala informal em contexto natural [Raso et al. 2018] e de fala telefônica [Raso et al. 2019], além de um minicorpus de inglês americano [Cavalcante et al. 2018] extraído do corpus de Santa Barbara [Du Bois et al. 2000-2005]. Os minicorpora extraídos dos outros dois subcorpora estão em curso de compilação. Todos os minicorpora são etiquetados com base na L-AcT.

Questões éticas

O corpus foi realizado depois da aprovação do comité de ética da UFMG (COEP: CAAE 0209.0.203.000-07). Todos os participantes das sessões de fala formal em contexto natural e de fala telefônica assinaram um TCLE em conformidade com a legislação atual. As informações que poderiam identificar os participantes foram bipadas e, nas transcrições e metadados, os participantes são identificados por acrônimos. A idade e a escolaridade são registradas em faixas, também com o objetivo de impedir sua identificação. No subcorpus de mídia, foram disponibilizadas apenas as gravações de programas da emissora Rede Globo, a qual concedeu autorização para sua disponibilização.

Referências

Bick, E. (2000). The parsing system “PALAVRAS”: automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus: Aarhus University Press,

Bick, E. (2012). A anotação gramatical do C-ORAL-BRASIL. In: Raso, T.; Mello, H. (org.). *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora da UFMG: 223-254.

Boves, L. e Oostdijk, N. (2003). Spontaneous speech in the spoken Dutch Corpus. Proceeding ISCA and IEEE Worjshop on Spontaneous Speech Processing and Recognition (SSPR). Tokio.

Cavalcante, F., Raso, T. e Ramos, A. (2018). Minicorpus de inglês americano etiquetado informacionalmente. https://www.c-oral-brasil.org/minicorpus_ingles_americano.php

Cavalcante, F., Raso, T., Barbosa, P. (2023). The Topic information unit: modeling prosodic forms in a crosslinguistic perspective. Em: Castagneto M. E Ravetto, M. (Orgs.). *La comunicazione parlata*. Roma: Aracne: 427-460.

Cresti, E. e Moneglia, M. (2005). *C-ORAL-ROM: integrated reference corpora for spoken romance languages*. Amsterdam/Philadelphia: John Benjamins.

Cresti, E. E Moneglia, M. (2018). The illocutionary basis of information structure. Em: Adamou, E., Haude, K. E Vanhove, M. *Information Structure in Lesser-described Languages*. Amsterdam: John Benjamins: 359-402.

Cresti, E. e Moneglia, M. (2023). The role of prosody in the identification of illocutionary types. Polar questions vs. Request of confirmation in the L-AcT taxonomy. Em: Studi Italiani di Linguistica Teorica e Applicata, LII, 1: 24-40.

Du Bois, J., Chafe, W., Meyer, Ch., Thompson, S., Englebretson, R. e Martey, N. (2000-2005). *Santa Barbara corpus of spoken American English*, 1-4. Philadelphia: Linguistic Data consortium.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, v. 76, n. 5, p. 378-382, 1971.

Izre'el, Sh.; Mello, H.; Panunzi, A.; Raso, T. (Orgs.). *In search of basic units of spoken language: A corpus-driven approach*. Amsterdam: John Benjamins, 2020.

MacWhinney, B. J. *The CHILDES Project: tools for analyzing talk*. 3rd ed. Mahwah: Lawrence Erlbaum Associates, 2000.

Martin, Ph. (2004). WinPitch Pro. A Tool for Speech Alignment and Prosodic Analysis. *Proceedings of SpeechProsody*: 545-548. https://www.isca-archive.org/speechprosody_2004/martin04b_speechprosody.html#

Mello, H. R. Methodological issues for spontaneous speech corpora compilation: The case of C-ORAL-BRASIL. In: RASO, T.; MELLO, H. R. (org.). *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins Publishing Company, 2014. p. 27-68.

Mello, H., Raso, T. Para a transcrição da fala espontânea: o caso do C-ORAL-BRASIL. Revista Portuguesa de Humanidades, v. 13, p. 301-326, 2009.

Mello, H.; Raso, T.; Mittmann, M.; Vale, H.; Côrtes, P. Transcricao e segmentação prosodica do corpus C-ORAL-BRASIL: criterios de implementacao e validacao. In: Raso, T.; Mello, H. (org.). *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora da UFMG, 2012. p. 125-145.

Moneglia, M.; Cresti, E. L'intonazione e i criteri di transcrizione del parlato adulto e infantile. In: Bortolini, U.; Pizzuto, E. (org.). *Il Progetto CHILDES Italia*. Pisa: Del Cerro, 1997. p. 57-90.

Moneglia, M.; Raso, T. Notes on Language into Act Theory (L-AcT). In: Raso, T.; Mello, H. (org.). *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins, 2014. p. 468-494.

Raso, T. O corpus C-ORAL-BRASIL. In: RASO, T.; MELLO, H. (org.). *C-ORALBRASIL I: Corpus de referência do português brasileiro falado informal*. 1. ed. Belo Horizonte: Editora UFMG, 2012a, p. 55-90.

Raso, T.; Cavalcante, F.; Bossaglia, G. (2018). Minicorpus de portugês brasileiro falado informal etiquetado informacionalmente. https://www.c-oral-brasil.org/minicorpus_portugues_brasileiro_2018.php

Raso, T.; Soares, E.; Miranda, I. (2019). Minicorpus de portuês brasileiro telefônico etiquetado informacionalmente. https://www.c-oral-brasil.org/minicorpus_telefonico.php

Raso, T.; Mello, H. (2012). C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.

Raso, T.; Mello, H.; Rocha, B.; Ferrari, L. (2026). C-ORAL-BRASIL II. Corpus de fala espontânea do português brasileiro formal em contexto natural, mídia e telefone. http://www.c-oral-brasil.org/c-oral-brasil-ii.php

Raso, T. Santos, S., Rilliard, A., Moraes, J. A. (2025) Defining and Idenfying Discourse Markers in Spontaneous Speech. In: Oliveira, M. Jr. (Org.). *Prosodic Interfaces: Interdisciplinary Perspectives on Sound Patterns and Human Interaction*. Berlin: DeGruyter: 65-102.

Vieira, M. (Em preparação). Automatic Audio-Quality Evaluation in Speech Corpora.

Witterburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjies, H. (2006) ELAN: A Professional Framework for Multimodality Research. In: *Proceedings of the LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC)*, Paris, European Language Resources Association (ELRA): 1556-1559.

Mudanças linguísticas em abstracts acadêmicos após a difusão do ChatGPT: uma análise baseada em corpus

Renato de Vargas Matos, Simone Sarmento

Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre - RS - Brasil

renatovargas1983@gmail.com; simone.sarmento@ufrgs.br

Abstract. *This study investigates linguistic changes in English-language abstracts published before and after the diffusion of ChatGPT. The corpus comprises 1,600 abstracts from Brazilian and international Linguistics journals. Sketch Engine was used to examine normalized frequencies and retrieve linguistic patterns through CQL queries. Log-likelihood tests were then applied to identify statistically significant differences in frequency between the two periods. The results reveal increases in verbs associated with argumentative emphasis and in discourse-condensing multilexical patterns, alongside reductions in structures such as modal + verb and that-clause. The findings indicate localized changes in textual organization and argumentative positioning in academic abstracts.*

Keywords: Corpus Linguistics; academic writing in English; generative artificial intelligence.

Resumo*. Este estudo investiga mudanças linguísticas em abstracts escritos em inglês e publicados antes e após a difusão do ChatGPT. O corpus é composto por 1.600 abstracts provenientes de periódicos brasileiros e internacionais da área de Linguística. O Sketch Engine foi utilizado para examinar frequências normalizadas e recuperar padrões linguísticos por meio de consultas em CQL. Em seguida, testes de log-likelihood permitiram identificar diferenças de frequência estatisticamente significativas entre os dois períodos. Os resultados revelam aumento de verbos associados à ênfase argumentativa e de padrões multilexicais relacionados à condensação discursiva, além da redução de estruturas como modal + verbo e that-clause. Os achados apontam para mudanças localizadas na organização textual e no posicionamento argumentativo em abstracts acadêmicos.*

Palavras-chave: Linguística de Corpus; escrita acadêmica em inglês; inteligência artificial generativa.

1. Introdução

A difusão de sistemas de IA generativa a partir de 2022, após o lançamento do ChatGPT, expandiu as discussões sobre os impactos dessas ferramentas na escrita acadêmica em inglês. Embora a literatura recente tenha dedicado atenção a aspectos éticos, pedagógicos e regulatórios relacionados ao seu uso, ainda são relativamente escassos os estudos que investigam se a popularização dos *large language models* (LLMs) está associada a mudanças linguísticas em textos acadêmicos publicados.

Essa questão gera interesse no contexto da comunicação científica internacional, marcado pela centralidade do inglês na circulação do conhecimento. Estudos em *English*

for Research Publication Purposes (ERPP) mostram que a publicação acadêmica é regulada por convenções relacionadas à organização textual, às escolhas lexicais e às formas de posicionamento discursivo (Swales, 1990; Hyland, 2015). Os *abstracts* são, assim, considerados objetos produtivos para investigação por concentrarem informações importantes do estudo, seguirem padrões discursivos compartilhados e circularem de forma independente em bases de dados e sistemas de indexação.

Partindo desse contexto, este estudo investiga possíveis diferenças linguísticas em *abstracts* acadêmicos escritos em inglês e publicados antes e depois da difusão do ChatGPT. A pesquisa busca responder à seguinte questão: em que medida a popularização dos sistemas de IA generativa está associada a alterações em padrões linguísticos recorrentes de *abstracts* da área da Linguística? O objetivo é identificar mudanças na frequência e na distribuição de recursos linguísticos apontados pela literatura recente como potencialmente sensíveis ao uso dessas tecnologias.

O *corpus* analisado é composto por 1.600 *abstracts* publicados em periódicos Qualis A1 da área da Linguística, incluindo revistas brasileiras e internacionais. Organizado em dois conjuntos temporais, o *corpus* reúne 800 *abstracts* publicados entre 2020 e 2022 (Pré-ChatGPT) e 800 publicados entre 2024 e 2025 (Pós-ChatGPT). Foram incluídos apenas periódicos que mantiveram continuidade editorial ao longo dos dois períodos analisados e que publicaram artigos em língua inglesa.

As denominações Pré-ChatGPT e Pós-ChatGPT funcionam exclusivamente como marcadores temporais e não implicam que os *abstracts* do segundo período tenham sido produzidos ou revisados com o ChatGPT ou outras ferramentas de IA generativa. A comparação permite, portanto, identificar diferenças linguísticas entre os dois períodos, sem atribuí-las diretamente ao uso dessas ferramentas pelos autores.

A análise no *Sketch Engine* envolveu o exame de frequências normalizadas e a recuperação de padrões linguísticos por meio de consultas em *Corpus Query Language* (CQL). Em seguida, testes de *log-likelihood* foram aplicados para identificar diferenças de frequência estatisticamente significativas entre os períodos Pré-ChatGPT e Pós-ChatGPT. Ao todo, foram investigados 197 itens linguísticos, selecionados com base em estudos recentes sobre escrita acadêmica e IA generativa e em análises exploratórias do *corpus*. Esses itens abrangem recursos relacionados à apresentação de resultados, ao posicionamento argumentativo, à avaliação e à condensação discursiva.

2. Fundamentação teórica

Estudos em ERPP mostram que a escrita acadêmica em inglês costuma seguir convenções discursivas consolidadas na comunicação científica, especialmente na forma como os textos organizam informações e apresentam seus resultados (Swales, 1990; Hyland, 2015). No caso dos *abstracts*, estudos anteriores já observaram recorrência de estruturas mais compactas ligadas à organização do gênero e à condensação das informações presentes no texto (Graetz, 1982; Lorés, 2004; Swales e Feak, 2012).

Essas características tornam o *abstract* um objeto pertinente para investigar possíveis mudanças nos padrões linguísticos da escrita acadêmica no contexto de difusão da IA generativa. Como os LLMs geram textos com base em padrões linguísticos recorrentes identificados em grandes volumes de dados (Bender *et al*., 2021), é pertinente

investigar se o período de popularização dessas ferramentas coincide com mudanças observáveis nos *abstracts* acadêmicos publicados após 2022.

Estudos baseados em Linguística de Corpus têm identificado diferenças entre textos de autoria humana e textos gerados por IA. Por meio da Análise Multidimensional (AMD), Berber Sardinha (2024) identificou perfis linguísticos distintos nesses dois tipos de produção. Em uma investigação mais próxima ao objeto deste estudo, Ali e Ali (2024) encontraram diferenças relacionadas à densidade informacional e à argumentação ao compararem *abstracts* de Linguística escritos por pesquisadores e gerados pelo ChatGPT. Demir e Egbert (2026) também observaram diferenças entre textos acadêmicos humanos e gerados pelo ChatGPT em artigos científicos e livros didáticos. Diferentemente dessas pesquisas, o presente estudo não utiliza a AMD nem compara diretamente textos humanos e textos gerados por IA. Seu foco recai sobre mudanças na frequência de recursos linguísticos específicos em *abstracts* publicados antes e após a difusão do ChatGPT.

Estudos recentes têm investigado características linguísticas de textos gerados por IA em comparação com produções de autoria humana (Ali e Ali, 2024; Goulart *et al*., 2024; Wu, 2025). Outros trabalhos examinaram mudanças em textos produzidos antes e após o lançamento do ChatGPT (Aslan e Iqbal, 2025; Bao *et al*., 2025). Permanecem escassas, contudo, as análises baseadas em *corpora* comparáveis de textos acadêmicos efetivamente publicados em periódicos ao longo desse período. É nessa lacuna que se insere a presente pesquisa.

3. Metodologia

Este estudo comparativo adota uma abordagem quantitativo-interpretativa baseada em Linguística de Corpus e orientada teoricamente pelo ERPP para investigar possíveis diferenças linguísticas em *abstracts* acadêmicos publicados antes e após a difusão do ChatGPT.

O *corpus* utilizado, denominado *Corpus of Abstracts* (CABS), é composto por 1.600 *abstracts,* selecionados em 10 periódicos nacionais e 14 periódicos internacionais do estrato Qualis A1 e dividido em dois grupos - Pré-ChatGPT e Pós-ChatGPT. O *subcorpus* Pré-ChatGPT reúne *abstracts* de periódicos nacionais e internacionais publicados entre 2020 e 2022, totalizando 800 *abstracts* (158.780 *tokens*). De modo equivalente, o *subcorpus* Pós-ChatGPT também conta com 800 *abstracts*, esses publicados entre 2024 e 2025 (169.561 *tokens*).

Essa etapa busca verificar se há aumento ou diminuição na ocorrência de itens lexicais que pesquisas anteriores apontam como sensíveis ao uso de IA. Para isso, foi constituída uma lista consolidada de 197 itens lexicais (divididos entre verbos, substantivos, adjetivos, advérbios, pronomes e itens multilexicais), organizada a partir dos levantamentos realizados por Geng e Trotta (2024); Kobak *et al.* (2024); Aslan e Iqbal (2025); Bao *et al.* (2025); Juzek e Ward (2025); He e Bu (2025); Terčon e Dobrovoljc (2025).

O *Sketch Engine* foi utilizado para organizar o *corpus*, extrair frequências normalizadas e realizar consultas em *Corpus Query Language* (CQL). As frequências dos itens linguísticos nos períodos Pré-ChatGPT e Pós-ChatGPT foram comparadas por meio do teste de *log-likelihood*. Foram considerados estatisticamente significativos os contrastes com valor absoluto de LL igual ou superior a 3,84.

4. Resultados e discussão

A maior parte dos recursos linguísticos investigados permaneceu estável entre os períodos Pré-ChatGPT e Pós-ChatGPT. Apenas uma parcela dos itens apresentou diferenças estatisticamente significativas, concentradas em escolhas lexicais e construções específicas. Esse predomínio da estabilidade sugere a continuidade das convenções que caracterizam o gênero *abstract*, acompanhada de mudanças linguísticas localizadas.

Entre os verbos que apresentaram diferenças significativas, predominaram os aumentos no período Pós-ChatGPT. O maior contraste foi observado em *highlight*, cuja frequência passou de 62 para 137 ocorrências (LL = -24,26). Também aumentaram *emphasize* (de 11 para 32; LL = -9,38), *address* (de 63 para 108; LL = -9,21) e *reveal* (de 110 para 166; LL = −8,06), além de *foster*, *align with*, *employ*, *underscore*, *delve into*, *underpin*, *encompass*, *explore*, *synthesize*, *deepen* e *enrich*. *Show* foi o único verbo que apresentou redução significativa, passando de 340 para 308 ocorrências (LL = +4,38). Embora a maioria dos verbos tenha permanecido estável, as mudanças observadas apontam para maior recorrência de formas associadas à ênfase discursiva e à explicitação de resultados e contribuições da pesquisa.

No conjunto dos substantivos, *identity* apresentou o maior aumento no período Pós-ChatGPT, passando de 61 para 140 ocorrências (LL = -26,93), seguido de perspective, que passou de 104 para 196 (LL = -22,95). Também aumentaram *study* (de 665 para 831; LL = -9,17), *findings* (de 135 para 198; LL = -8,21), *voice* (de 53 para 91; LL = -7,81), *impact* (de 45 para 77; LL = -6,52) e *research* (de 366 para 458; LL = -5,14), além de *nuance*, *engagement*, *journey*, *integration*, *challenge*, *element* e *trajectory*. Os aumentos concentram-se em substantivos relacionados à apresentação da pesquisa, ao posicionamento epistemológico e à construção do objeto investigado. Como a maioria dos itens permaneceu estável, esses contrastes não indicam uma alteração ampla no repertório nominal do gênero.

As diferenças observadas nos adjetivos foram mais restritas, destacando-se os aumentos de *significant* (48 para 100 ocorrências; LL = -15,41), *critical* (111 para 180; LL = -12,30), *analytical* (15 para 41; LL = -10,90), *notable* (2 para 10; LL = -5,31) e *potential* (48 para 75; LL = -4,33). Também cresceram *systematic* e *methodological*. Em sentido contrário, *important* diminuiu de 54 para 38 ocorrências (LL = +3,95) e *empirical* passou de 46 para 31 ocorrências (LL = +4,01).

Entre os advérbios, *particularly* cresceu de 28 para 61 ocorrências (LL = -10,46) e *effectively* passou de 3 para 16 ocorrências (LL = -8,93). Em contraste, *similarly* reduziu de 7 para 1 ocorrência (LL = +5,46). Esses resultados sugerem intensificação pontual de recursos avaliativos e de realce discursivo.

As mudanças mais evidentes ocorreram nas expressões multilexicais. Construções com *vírgula + verbo em -ing* apresentaram o maior aumento de toda a análise, passando de 475 para 802 ocorrências (LL = -64,57). Esse foi o contraste mais acentuado entre os dois períodos. Também aumentaram *deeply rooted* (de 0 para 4 ocorrências; LL = -5,29) e *in summary* (de 0 para 3; LL = -3,97). Em contrapartida, houve redução de *modal + verb* (de 596 para 507; LL = +14,23), *that-clause* (de 880 para 789; LL = +12,74), *relative clauses* (de 1.182 para 1.117; LL = +8,59), *tríade de substantivos* (de 81 para 58; LL = +6,38), *as well as* (de 116 para 88; LL = +5,91), *shed light on* (de 20 para 9; LL = +5,03), *verb + adverb + preposition* (de 160 para 134; LL = +4,33) e *on the other hand* (de 21

para 11; LL = +3,87). Esses resultados sugerem maior frequência de estruturas compactas e condensadas, acompanhada pela redução de alguns padrões mais expansivos.

Considerados em conjunto, os resultados apontam para uma variação seletiva na frequência dos recursos linguísticos investigados. As diferenças concentram-se em escolhas lexicais e construções específicas, enquanto a maioria dos itens permanece estável entre os períodos. Como a análise se restringiu a um conjunto previamente delimitado de recursos, os dados não permitem avaliar o possível surgimento de outros padrões discursivos.

5. Considerações finais

Este estudo investigou diferenças linguísticas em 1.600 *abstracts* da área de Linguística publicados antes e após a difusão do ChatGPT. Embora alguns recursos tenham apresentado diferenças estatisticamente significativas entre os períodos, a maior parte dos itens examinados permaneceu estável. Os contrastes observados são, portanto, localizados e não indicam uma transformação generalizada no repertório linguístico analisado.

Entre as principais mudanças, destacam-se o aumento de verbos associados à ênfase e à apresentação de contribuições científicas, o crescimento de substantivos relacionados ao posicionamento da pesquisa e a maior frequência de orações participiais em *-ing*, identificadas pelo padrão *vírgula + verbo em -ing*. Também foi observada a redução de algumas estruturas subordinadas e modalizadas. Esses resultados apontam para mudanças pontuais na maneira como resultados, procedimentos e contribuições são apresentados nos *abstracts*.

Considerados em conjunto, os achados sugerem uma reorganização das preferências por determinados recursos lexicais e sintáticos no período Pós-ChatGPT. Como a análise se restringiu a um conjunto previamente delimitado de itens, os resultados não permitem avaliar o possível surgimento de outros padrões discursivos nem sustentar que tenha ocorrido uma ruptura nas práticas de escrita acadêmica.

A comparação estabelece uma associação entre o período de publicação e as diferenças de frequência identificadas, mas não permite atribuir uma relação causal entre essas diferenças e a difusão do ChatGPT. O *corpus* não informa se os autores utilizaram ferramentas de IA generativa nem quais sistemas teriam sido empregados. Também não permite excluir a influência de outros fatores históricos, editoriais ou discursivos sobre as mudanças observadas. Pesquisas futuras poderão ampliar o recorte temporal e disciplinar, examinar outros gêneros acadêmicos e comparar textos cuja produção ou revisão com diferentes ferramentas de IA generativa esteja documentada.

Questões éticas

A pesquisa utilizou exclusivamente dados de acesso público, provenientes de artigos já publicados, não sendo necessária sua submissão a um Comitê de Ética em Pesquisa. Ferramentas de inteligência artificial foram empregadas como apoio à revisão linguística e à formatação do trabalho. Renato de Vargas Matos organizou e processou o *corpus*, realizou as consultas no *Sketch Engine* e elaborou a primeira versão do artigo. Simone

Sarmento orientou todas as etapas da pesquisa, colaborou na interpretação dos resultados e realizou a revisão integral do texto.

Referências

ALI, M.; ALI, S. Can artificial intelligence preferences be an alternative to human linguistic choices? A multidimensional analysis of research abstracts of English linguistics. *Journal of Nusantara Studies* (JONUS), v. 9, n. 2, p. 514-536, 2024.

ASLAN, M. H.; IQBAL, C. A comparative analysis of syntactic complexity in academic writing before and after the advent of large language models. Rochester, NY: SSRN, 2025. Preprint. *Southwestern University of Finance and Economics.* Disponível em: https://ssrn.com/abstract=5903446. Acesso em: 5 dez. 2025.

BAO, T. *et al.* Examining linguistic shifts in academic writing before and after the launch of ChatGPT: a study on preprint papers. *Scientometrics*, Budapest: Akadémiai Kiadó Zrt., v. 130, p. 3597-3627, 2025. DOI: https://doi.org/10.1007/s11192-025-05341.

BENDER, E. M. *et al*. On the dangers of stochastic parrots: Can language models be too big? In: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT '21), 2021, Virtual Event, Canada. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* New York: ACM, 2021. p. 610-623.

BERBER SARDINHA, T. AI-generated vs human-authored texts: a multidimensional comparison. *Applied Corpus Linguistics*, v. 4, n. 1, art. 100083, 2024.

DEMIR, N. Y.; EGBERT, J. Register alignment of ChatGPT-generated academic texts. *Applied Corpus Linguistics*, v. 6, n. 1, art. 100174, 2026.

GENG, M.; TROTTA, R. Is ChatGPT transforming academics' writing style? In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 41., 2024, Vienna, Austria. *Proceedings of the 41st International Conference on Machine Learning – Workshop The Next Generation of AI Safety.* [S.l.]: PMLR, 2024. p. 1-14.

GOULART, L. *et al.* AI or student writing? Analyzing the situational and linguistic characteristics of undergraduate student writing and AI-generated assignments. *Journal of Second Language Writing,* Amsterdam: Elsevier, v. 66, p. 101160, 2024.

GRAETZ, N. Teaching EFL students to extract structural information from abstracts. In: *International Symposium on Language for Special Purposes (LSP),* 1982, Eindhoven, The Netherlands. Proceedings… Eindhoven, 1982. 22 p.

HE, Y.; BU, Y. Academic journals' AI policies fail to curb the surge in AI-assisted academic writing. *Proceedings of the National Academy of Sciences,* Washington, DC: National Academy of Sciences, v. 123, n. 9, p. e2526734123, 2026.

HYLAND, K. *Genre, discipline and identity. Journal of English for Academic Purposes*, Amsterdam, v. 19, p. 32–43, 2015.

JUZEK, T. S.; WARD, Z. B. Why does ChatGPT "Delve" so much? Exploring the sources of lexical overrepresentation in Large Language Models. In: *International Conference on Computational Linguistics,* 31., 2025, Abu Dhabi. Proceedings […]. Abu Dhabi: International Committee on Computational Linguistics, 2025. p. 6397-6411.

KOBAK, D. *et al*. Delving into ChatGPT usage in academic writing through excess vocabulary. *Patterns*, New York: Cell Press, v. 6, n. 2, 2025, e100987. DOI: https://doi.org/10.1016/j.patter.2024.100987

LORÉS, R. On RA abstracts: from rhetorical structure to thematic organisation. *English for specific purposes.* Amsterdam: Elsevier, v. 23, n. 3, p. 280–302, 2004. DOI: 10.1016/j.esp.2003.06.001.

SWALES, J. M.; FEAK, C. B. *Academic writing for graduate students:* essential tasks and skills. 3. Ed. Ann Arbor: University of Michigan Press, 2012. P. 279-302.

SWALES, J. M. Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press, selected 45-47, 52-60. In: *The discourse studies reader: main currents in theory and analysis.* John Benjamins Publishing Company, 1990. P. 306-316.

TERČON, L.; DOBROVOLJC, K. Linguistic Characteristics of AI-Generated Text: A Survey. *Linguistic Features,* Ljubljana: University of Ljubljana, preprint, v. 84, p. 1-84, 2025.

WU, J. A Corpus-Based Multidimensional Analysis of Linguistic Features between Human-Authored and ChatGPT-Generated Compositions. *International Journal of Linguistics, Literature and Translation,* Dubai: Al-Kindi Center for Research and Development, v. 8, n. 5, p. 102-110, 2025.

Automatização da Identificação de Funções Retóricas de Pacotes Lexicais em Artigos Científicos das Ciências Agrárias: uma análise exploratória do uso de Large Language Models (LLM)

Karol Traslatti Soares, Ana Eliza Pereira Bocorny

Universidade Federal do Rio Grande do Sul

karoltsletras@gmail.com; abocorny@gmail.com

Resumo: *A identificação de funções retóricas em artigos científicos constitui uma etapa importante para pesquisas em análise de gêneros acadêmicos e Linguística de Corpus. Através dessa análise, podemos identificar padrões recorrentes na escrita, mapear as particularidades do gênero e área de estudo, e fundamentar hipóteses sobre o uso da linguagem acadêmica. Entretanto, a categorização manual desses pacotes lexicais pode ser um processo intensivo. A análise do uso de Largue Language Models (LLMs), em uma tentativa de automatizar o processo de identificação desses dados, surgiu através do meu Trabalho de Conclusão de Curso (TCC), onde fiz a análise manual de 304 pacotes lexicais identificados em quatro corpora de seções de artigos (Introdução, Metodologia, Resultados/Discussão e Conclusão), totalizando cerca de quatro milhões de palavras. Estudos recentes demonstram que LLMs apresentam potencial para automatizar tarefas de anotação discursiva em diferentes gêneros acadêmicos (Kim & Lu, 2024). O presente trabalho pretende testar essa hipótese através da análise da capacidade que o ChatGPT e Gemini têm de agrupar e categorizar pacotes lexicais extraídos da seção Introdução de artigos de pesquisa da área das Ciências Agrárias, conforme sua função retórica.*

Palavras-chave: linguística de corpus; estrutura retórica; grandes modelos de linguagem.

Abstract*: Identifying rhetorical functions in scientific articles is an important step in research on academic genre analysis and Corpus Linguistics. Through this analysis, we can identify recurring patterns in academic writing, map the particularities of the genre and area of study, and authenticate hypotheses about the use of academic language. However, the manual categorization of these Lexical Bundles (LB) can be an intensive process. The analysis of the use of Large Language Models (LLMs), in an attempt to automate the process of identifying this data emerged from my Undergraduate Thesis, where I manually analyze of 304 lexical packages identified in four corpora of article sections (Introduction, Methodology, Results/Discussion, and Conclusion), totaling approximately four million words. Recent studies show that LLMs have the potential to automate discursive annotation tasks in different academic genres (Kim & Lu, 2024). This study aims to test these hypotheses by analyzing the ability of ChatGPT and Gemini to group and categorize lexical packages extracted from the Introduction section of research articles in the field of Agricultural Sciences, according to their rhetorical function.*

Keywords: corpus linguistics, rhetorical structure; large language models.

1. Introdução

A escrita acadêmica é caracterizada pela recorrência de padrões linguísticos que auxiliam autores a organizar o discurso científico. Dentre esses padrões, os pacotes lexicais (lexical bundles), definidos como sequências contínuas e recorrentes de palavras que aparecem com frequência em determinados registros discursivos (Biber et al., 1999), se destacam. Em artigos de pesquisa, essas unidades desempenham funções importantes relacionadas à organização textual.

Nos últimos anos, diferentes estudos têm procurado aproximar a análise de padrões linguísticos recorrentes nos gêneros acadêmicos, tentando compreender a relação entre forma e função (Moreno & Swales, 2018; Gray et al., 2020). Dentro dessa perspectiva, Soares, 2023) investigou a frequência e as funções discursivas de pacotes lexicais em artigos científicos da área das Ciências Agrárias.

Tradicionalmente, a identificação dessas funções retóricas é realizada por meio de análises manuais baseadas em modelos de gênero, como o modelo CARS de Swales (1990). Mas a análise e categorização manual desses pacotes lexicais constitui uma das etapas mais trabalhosas da análise de gêneros, muitas vezes limitando o tamanho dos *corpora* e o crescimento das pesquisas (Kim e Lu, 2024).

Recentemente, *Large Language Models* (LLMs), como o ChatGPT, tem demonstrado capacidade de executar tarefas de classificação textual e análise discursiva quando estes modelos são ajustados por *fine-tuning* (Kim e Lu, 2024). Nesse contexto, o presente estudo investigou o potencial de duas LLMs, ChatGPT e Gemini, para identificar automaticamente funções retóricas associadas a pacotes lexicais encontrados na seção Introdução de artigos científicos da área das Ciências Agrárias. O estudo, ainda em fase inicial, busca comparar os achados prévios encontrados através da classificação manual realizada por Soares (2023) e a categorização produzida pela IA.

2. Fundamentação Teórica

2.1 Pacotes Lexicais e Análise das Funções Retóricas

Pacotes Lexicais (Lexical Bundles) são sequências contínuas e recorrentes de palavras que aparecem com frequência em determinados registros discursivos (Biber et al., 1999). Estudos demonstram que esses pacotes lexicais estão associados a movimentos retóricos específicos e desempenham funções discursivas relevantes na organização do texto acadêmico.

Segundo Swales (1990), gêneros acadêmicos podem ser descritos a partir de movimentos retóricos que são responsáveis por diferentes propósitos comunicativos. Cada movimento é composto por etapas específicas relacionadas aos diferentes objetivos do estudo, como formulação de hipóteses, revisão da literatura, identificação de lacunas de pesquisa e etc.. A partir desses achados, Swales desenvolveu o modelo *Create a Research Space* (CARS) (1981), que auxilia autores a organizarem seus trabalhos de forma eficaz dentro da literatura existente.

Na seção Introdução de artigos científicos, por exemplo, pacotes lexicais como "*the aim of this study was*", "*the objective of this paper is to*", são frequentemente utilizados para realizar a função retórica de apresentação dos objetivos de pesquisa.

2.2 Large Language Models e Automatização da Análise Retórica

Kim e Lu (2024) argumentam que os LLMs apresentam vantagens importantes dentro do contexto da automatização da análise das funções retóricas por possuírem maior capacidade de compreender relações contextuais e discursivas complexas que muitas vezes acabam deixando os pesquisadores em dúvida durante a anotação manual. Segundo os autores, a combinação de refinamento de prompts (few-shot learning[3] e fine-tuning[4]), permite que modelos como o *ChatGPT* realizem tarefas de anotação retórica com níveis de desempenho similares às anotações manuais.

3. *Corpus* e procedimentos metodológicos

Os dados analisados neste estudo derivam do corpus previamente analisado de forma manual por Soares (2023), onde foram coletados quatro *corpora* de aproximadamente um milhão de palavras cada, compostos por textos das seções (Introdução, Metodologia, Resultados e Discussão e Conclusão) de artigos de pesquisa das Ciências Agrárias publicados em inglês em periódicos internacionais de acesso aberto com revisão por pares. A ferramenta AntCorGen (Anthony, 2019) foi utilizada para compilar o *corpus,* extraindo conteúdo da plataforma PLoS ONE com base nos critérios escolhidos pelo pesquisador.

Tabela 1. *Corpora* em números

Discipline	*Corpus*	Total number of words (*tokens*)	Total number of words without repetition (*types*)	Texts	Average number of words per section
Agrarian Sciences	Introduction	952,414	41,421	1,188	801
	Materials & Methods	953,541	60,189	620	1,537
	Results & Discussion	996,553	59,909	363	2,745
	Conclusions	870,294	40,766	4,108	211

Para a seção Introdução, com o auxílio da ferramenta *Sketch Engine,* foram extraídos 54 pacotes lexicais, que foram agrupados e classificadas de acordo com o framework de Shi e Wannaruk (2014), inspirado no modelo CARS de Swales, mas voltado especificamente para a área das Ciências Agrárias.

Posteriormente, os mesmos dados foram submetidos ao ChatGPT (GPT-4o) e ao Gemini (3.5 Flash), utilizando *prompts* desenvolvidos para duas tarefas específicas: (i) o agrupamento dos pacotes lexicais em famílias fraseológicas; (ii) a identificação das suas funções retóricas correspondentes aos aos movimentos e etapas propostos por Shi e

[3] *Few-shot learning* ou **aprendizado com poucos exemplos**, consiste em fornecer instruções acompanhadas de alguns exemplos práticos (geralmente 2 a 5) diretamente no prompt.

[4] *Fine-tuning* ou **ajuste fino**, consiste em trabalhar com um modelo de IA já existente e "continuar o treinamento" dessa IA alimentando-a com uma base de dados própria e específica.

Wannruk (2014).

Para garantir a comparabilidade dos resultados, os mesmos conjuntos de pacotes lexicais e as mesmas instruções foram submetidos ao ChatGPT e ao Gemini, permitindo observar diferenças e semelhanças nas classificações produzidas pelos dois modelos. Foi utilizado um único prompt, elaborado a partir das categorias propostas por Shi e Wannaruk (2014), aplicado de forma idêntica aos dois modelos. As classificações foram realizadas em contexto *zero-shot*[5], sem fornecimento prévio de exemplos classificados.

4. Resultados e Futuros Estudos

A comparação entre a análise manual realizada por Soares (2023) e as análises assistidas pelos LLMs (*ChatGPT* e *Gemini*) revelou semelhanças e convergências importantes na identificação das funções retóricas. Os dois métodos identificaram pacotes lexicais transparentes[6] e formulaicos com maior precisão.

Durante o agrupamento das famílias lexicais, o *prompt* pedia para que as ferramentas organizassem os pacotes lexicais por palavras-chave. Enquanto o agrupamento realizado pelo *ChatGPT* estava mais de acordo com o *prompt* e a categorização foi feita pela identificação de palavras-chave. A plataforma *Gemini*, durante o agrupamento, realizou uma pré-análise e organizou os dados já de acordo com suas funções retóricas. Essa organização foi realizada a partir de informações que a plataforma buscou da web, utilizando como referência o modelo CARS.

A análise automática apresentou concordância com a classificação manual em aproximadamente 47 dos 54 pacotes lexicais analisados (87%), enquanto 7 pacotes (13%) apresentaram divergências de classificação. Esse resultado é coerente com a natureza dos erros observados: a IA teve melhor desempenho em pacotes lexicais transparentes e altamente formulaicos (objetivos, revisão da literatura e lacunas de pesquisa), mas encontrou mais dificuldades em unidades opacas[7], especialmente aquelas relacionadas à descrição de procedimentos e à apresentação de hipóteses.

Os achados fraseológicos mais produtivos identificados automaticamente foram aqueles relacionados à apresentação dos objetivos de pesquisa. Expressões como "*the aim of this study*", "*the objective of this study was*" e "*the purpose of this study is*", foram agrupados e associados ao Move 3, Step 1, categoria utilizada no framework de Shi e Wannaruk (2014) para identificar a apresentação dos objetivos do estudo. Essa classificação reproduz diretamente os resultados encontrados por Soares (2023).

Uma das divergências observadas está relacionada à categorização de unidades associadas a funções retóricas menos frequentes. Enquanto a análise manual identificou categorias específicas relacionadas a apresentação de hipóteses como, por exemplo, "*in many parts of the world*" como *Move 3: Introducing the present study. Step 3: Describing*

[5] ***Zero-Shot*** é uma técnica de engenharia de prompt onde é solicitado que a Inteligência Artificial execute uma tarefa sem fornecer nenhum exemplo prévio. A IA baseia-se unicamente no seu conhecimento prévio (através de sua base de daos) e compreensão de linguagem para gerar o resultado esperado.

[6] Segundo Lameira (2022), **pacotes lexicais transparentes** são aqueles pacotes lexicais mais fáceis de serem identificados e categorizados.

[7] Segundo Lameira (2022), **pacotes lexicais opacos** são aqueles pacotes lexicais de mais difícil identificação e categorização.

procedures, o modelo geracional frequentemente agrupou essas unidades em categorias mais amplas. Nesse caso, algumas unidades foram associadas ao *Move 1: Stating why the topic is important. S2: Making topic generalizations*, isto é, ao movimento de justificar a importância do tema por meio de apresentações de generalizações sobre o tópico. Essa divergência, é consistente com a observação de Kim e Lu (2024) que apontam que "etapas retóricas mais específicas tendem a apresentar menor precisão de classificação quando comparadas aos movimentos retóricos mais amplos".

De acordo com os resultados apresentados por Kim e Lu (2024), os achados desse estudo reforçam o potencial dos LLMs como ferramentas de apoio para a automatização da função retórica em pesquisas de Linguística de Corpus. Como sequência do estudo, será feita a implementação de um *fine-tuning* utilizando dados mais amplos da anotação manual para alimentar estes LLMs, visando o aperfeiçoamento dos *prompts*. Essa etapa poderá aumentar significativamente a precisão da classificação, especialmente em categorias menos frequentes e mais complexas (Kim e Lu, 2024).

Além disso, futuras investigações serão expandidas para a análise das demais seções do artigo científico da área das Ciências Agrárias, a fim de verificar se os padrões identificados na análise da seção Introdução se mantêm ao longo de toda a estrutura do gênero artigo científico. A ampliação dos dados permitirá uma avaliação mais abrangente do potencial desses LLMs para a automatização da identificação de funções retóricas em artigos científicos.

Referências

BIBER, D. , & CONRAD, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora.- Studies in honor of Stig Johansson* (pp. 182—190). Amsterdam: Rodopi.

CORTES, V.; FRIGINAL, E. *Advances in Corpus-based Research on Academic Writing: Effects of discipline, register, and writer expertise.* Amsterdam: John Benjamins, p. 137-168, 2020.

GRAY, B.; COTOS, E.; SMITH, J. Combining rhetorical move analysis with multi-dimensional analysis: research writing across disciplines. In: RÖMER, Ute; CORTES, Viviana; FRIGINAL, Eric (ed.). *Advances in corpus-based research on academic writing: effects of discipline, register, and writer expertise*. Amsterdam: John Benjamins Publishing Company, 2020. p. 137–168.

KIM, Minjin; LU, Xiaofei. *Exploring the potential of using ChatGPT for rhetorical move-step analysis: the impact of prompt refinement, few-shot learning, and fine-tuning". Journal of English for Academic Purposes*, [S. l.], v. 71, 2024, 101422. DOI: 10.1016/j.jeap.2024.101422.

LAMEIRA, V. M. *A análise funcional de estruturas lexicais (lexical frames): dados linguísticos extraídos de corpora para subsidiar o ensino de IFA*. Dissertação (Mestranda em Letras). Instituto de Letras, Universidade Federal do Rio Grande do Sul. 2022

MORENO, A. I., & SWALES, J. M. (2018). Strengthening Move Analysis Methodology. Towards Bridging the Function-Form Gap. *English for Specific Purposes*, 50, 40-63.https://doi.org/10.1016/j.esp.2017.11.006

SHI, H., & WANNARUK, A. (2014). Rhetorical Structure of Research Articles in Agricultural Science. *English Language Teaching*, 7(8), 41. https://doi.org/10.5539/elt.v7n8p1

SOARES, K. T. (2023). *Bridging the gap between form and function: a corpus-driven study of lexical bundles and lexical frames across sections of agrarian sciences research articles*. Instituto de Letras, Universidade Federal do Rio Grande do Sul. 2023.

Aplicações iniciais de modelos de língua na anotação *Rhetorical Structure Theory* em textos jurídicos do português

Winycius Alberto Nascimento do Brasil[1], Jackson Wilke da Cruz Souza[2]

[1]Instituto de Ciência, Tecnologia e Inovação (ICTI) – Universidade Federal da Bahia (UFBA) – Camaçari/BA
[2]Programa de Pós-Graduação em Língua e Cultura (PPGLinc) – Universidade Federal da Bahia (UFBA) – Salvador/BA

winyciusbrasil@ufba.br, jacksoncruz@ufba.br

Abstract. *This study demonstrates how the Document-Level Multilingual RST (DMRST) model can be used to identify rhetorical relations in the Brazilian Consumer Protection Code (Código de Defesa do Consumidor – CDC). To this end, data preprocessing and model fine-tuning were carried out, considering two analysis approaches: intersentential and intrasentential. The study was motivated by the scarcity of discourse-annotated resources for the legal domain in Portuguese and by the need to investigate the behavior of discourse parsing models in normative texts. The results revealed limitations in the intrasentential analysis due to the excessive fragmentation of discourse units, whereas the intersentential approach achieved better performance in identifying rhetorical relations. A high occurrence of the Joint relation was also observed, suggesting that the sequential structure of CDC articles influences the discourse interpretation performed by the model. The findings indicate the potential of DMRST for legal NLP applications and highlight challenges related to discourse segmentation and model adaptation to this specific domain.*

Resumo. *Neste trabalho demonstramos como o modelo Document-Level Multilingual RST (DMRST) pode ser utilizado para identificar relações retóricas no Código de Defesa do Consumidor. Para tanto, foram realizados pré-processamento dos dados e fine-tuning do modelo, visando duas abordagens de análise (inter e intrassentencial). O estudo foi motivado pela escassez de recursos discursivamente anotados para o domínio jurídico em português e pela necessidade de investigar o comportamento de modelos de análise discursiva em textos normativos. Os resultados apresentaram limitações na análise intrassentencial, devido à fragmentação excessiva das unidades discursivas, enquanto a abordagem intersentencial apresentou melhor desempenho na identificação de relações retóricas. Observou-se ainda uma elevada ocorrência da relação Joint, sugerindo que a estrutura sequencial dos artigos do CDC influencia a interpretação discursiva realizada pelo modelo. Os resultados indicam o potencial do DMRST para aplicações em PLN jurídico e apontam desafios relacionados à segmentação e adaptação de modelos para esse domínio específico.*

1. Introdução

A análise discursiva investiga como diferentes segmentos de um texto se organizam e se relacionam para construir uma unidade de sentido [Cardoso, Souza e Rodrigues, 2023]. Nesse contexto, não se busca identificar apenas os conteúdos apresentados, mas também as relações de sentido que podem ser estabelecidas entre as partes do texto. Uma das teorias utilizadas para representar essas relações e que tem bastante aderência ao Processamento de Linguagem Natural (PLN) é a *Rhetorical Structure Theory* (RST) [Mann; Thompson, 1988], que descreve a organização retórica de um texto a partir de unidades discursivas relacionadas entre si.

Na RST, essas relações podem ser estabelecidas entre segmentos que desempenham diferentes funções discursivas, sendo classificadas, em geral, como núcleo (N), que apresenta a informação mais central, e satélite (S), que fornece informações complementares ao núcleo. As relações entre N e S se dão de maneira hierárquica, sendo possível identificar certa organização de importância entre elas; a essas relações a RST classifica como mononucleares. Já a organização não hierárquica, a teoria classifica como multinuclear. Ao final, tem-se uma representação arbórea de unidades discursivas, relações e sinalizadores que indicam a existência dessas relações. Essa representação permite modelar explicitamente como diferentes segmentos contribuem para a organização e interpretação do discurso e tem sido aplicada em tarefas de PLN, incluindo análise discursiva, sumarização automática e *discourse parsing* [Souza, Cardoso e Rodrigues; 2024].

Na Figura 1, apresenta-se uma representação simplificada de uma árvore RST com a relação *Elaboration*. Nesse tipo de relação, a unidade nuclear (N) é complementada por outra unidade satélite (S), responsável por fornecer detalhes ou explicações adicionais. No exemplo, a *Elementary Discourse Unit* (EDU) 1 apresenta a informação principal sobre o direito do consumidor à informação adequada, enquanto a EDU 2 especifica quais informações devem ser fornecidas. Essa organização evidencia como a RST representa explicitamente as relações de dependência entre segmentos textuais, contribuindo para a compreensão da estrutura discursiva do texto [Mann e Thompson, 1988; Cardoso *et al.*, 2011].

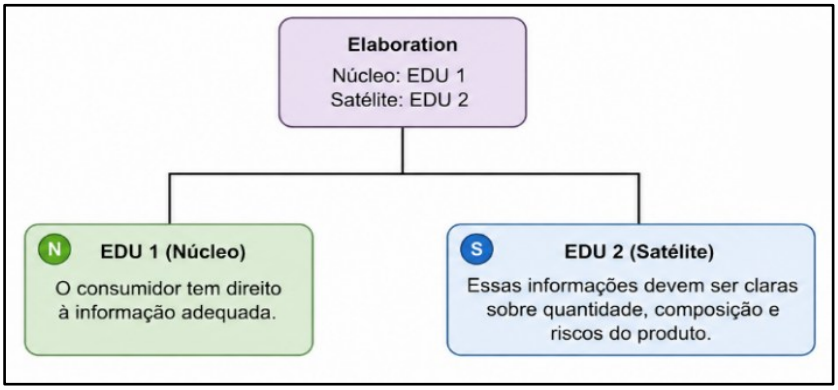


Figura 1. Exemplo de árvore RST.

Os trabalhos de análise discursiva a partir da RST têm se baseado em *corpora* de gêneros textuais cuja norma linguística é a formal, como notícias jornalísticas [Cardoso *et al.*, 2011] e textos acadêmicos [Pardo e Seno, 2005], por exemplo, e, mais recentemente, em

textos informais, como conteúdo de redes sociais [Pereira e Souza, 2025]. No entanto, ainda há uma escassez de *corpora* anotados segundo a RST para outros gêneros textuais, e essa ausência dificulta o desenvolvimento e a adaptação de modelos de língua pré-treinados para determinados domínios, com o jurídico, cujos textos apresentam organização, vocabulário e estratégias discursivas específicas. Além disso, a escassez de dados anotados em padrão-ouro para gêneros específicos pode limitar o desempenho de modelos quando aplicados a gêneros distintos daqueles utilizados em seu treinamento, pois a diferença entre esses domínios pode envolver padrões discursivo e construções linguísticas particulares [Gururangan *et al.*, 2020].

A construção de recursos discursivamente anotados, entretanto, envolve um processo especializado e de alto custo, uma vez que exige segmentação dos textos em EDUs e das relações retóricas que se estabelecem entre as unidades. Diante desse cenário, modelos de língua pré-treinados e *Large Language Models* (LLMs) têm sido investigado como alternativas para auxiliar etapas de anotação e reduzir o esforço necessário à construção de *corpora* linguísticos [Zhang *et al.*, 2023]. Entretanto, essa alternativa não descarta a necessidade de anotação e validação humanas, especialmente em domínios especializados, mas permite, por outro lado, explorar inicialmente o comportamento dos modelos automáticos diante de dados ainda não anotados.

Especificamente no domínio jurídico, essa questão é relevante devido às especificidades linguístico-discursivas dos textos. Xue *et al.*, (2024), por exemplo, demonstram, com base na construção do *corpus* LegalDiscourse, que a identificação automática de segmentos e relações discursivas em textos legais ainda apresenta desafios quando comparada à anotação realizada por especialistas humanos. Os achados neste estudo reforçam tanto a importância da construção de recursos discursivamente anotados para o domínio em questão, quanto a necessidade de investigar em qual medida modelos pré-treinados podem contribuir para esse processo.

Partindo dessa perspectiva, este estudo constitui uma etapa exploratória anterior à construção de um corpus jurídico manualmente anotado segundo a RST. Ao invés de assumir as estruturas produzidas automaticamente como anotações de referência, buscou-se investigar inicialmente do modelo *Document-Level Multilingual RST* (DMRST) [Liu, Shi e Chen, 2021] quando aplicado a textos jurídico-normativos em português. Os resultados dessa etapa poderão subsidiar a posterior elaboração de diretrizes de anotação e a construção de uma amostra manualmente anotada, necessária para a avaliação objetiva do desempenho do modelo, além de seu *fine tuning* para escalabilidade da tarefa.

Para tanto, escolheu-se a área do Direito Civil, mais especificamente o Direito do Consumidor, por esta pesquisa estar vinculada ao projeto DIGGER, que visa desenvolver recursos de PLN para essa área, como sistemas de perguntas e respostas ou recuperação da informação [Athaydes *et al.*, 2024].

Este artigo está estruturado em três seções, além desta Introdução. Na Seção 2, apresenta-se a metodologia adotada na pesquisa, incluindo a descrição do corpus utilizado, o processo de pré-processamento dos dados e as estratégias empregadas para o treinamento e avaliação do modelo DMRST. Na Seção 3, são apresentados e discutidos os resultados obtidos nos experimentos realizados com o corpus jurídico. Por fim, na Seção 4, são expostas as considerações finais, destacando as principais contribuições do estudo, suas limitações e perspectivas para trabalhos futuros.

2. Metodologia

Metodologicamente, utilizamos o Código de Defesa do Consumidor (CDC) [Brasil, 1990] como *corpus* de análise, por ser um documento representativo do gênero jurídico-normativo, permitindo investigar o comportamento do DMRST em um domínio distinto daqueles que foram utilizados em seu treinamento. Nesse caso, foi necessário considerar especificidades linguísticas desse gênero textual, como a linguagem técnico-jurídica e a estrutura normativa e formal (como a presença de artigos, incisos e parágrafos). Para este estudo piloto, foram utilizados os três primeiro capítulos do CDC, totalizando 42 sentenças. A escolha dessa amostra decorre do caráter piloto do estudo, em que buscou-se avaliar inicialmente a aplicabilidade do modelo ao domínio jurídico antes da ampliação do experimento para a totalidade do documento.

Neste estudo, foram realizadas análises em dois níveis de construção sentencial: *intra* e inter*ssentencial*. Na abordagem intrassentencial, cada sentença foi submetida ao modelo para segmentação em EDUs e identificação das relações estabelecidas entre essas unidades. Na abordagem intersentencial, foram considerados pares de sentenças consecutivas, em que buscou-se identificar a relação discursiva estabelecida entre elas.

A avaliação realizada neste estudo possui caráter exploratório e qualitativo, tendo como objetivo observar o comportamento do DMRST diante de diferentes estratégias de construção das unidades de análise no domínio jurídico. Não foi utilizada, nesta etapa, uma amostra manualmente anotada segundo a RST que pudesse atuar como *gold standard* para comparação das relações discursivas identificadas pelo modelo. Dessa forma, não foram calculadas métricas quantitativas de desempenho (como concordância ou acurácia). A anotação automática das EDUs, portanto, devem ser consideradas, nesta etapa, como resultados exploratórios do modelo, e não anotações discursivas de referência.

A análise apresentada na Seção 3 concentra-se, portanto, na análise de (i) estruturas discursivas produzidas pelo modelo, (ii) distribuição das relações identificadas e (iii) diferenças observadas entre as abordagens intra e intersentencial. Especificamente sobre (ii), foram contabilizadas as relações atribuídas pelo DMRST às unidades analisadas, agrupando-se as ocorrências por capítulo do CDC e por tipo de relação retórica.

Com relação à anotação automática, foi realizado o *fine tuning* do modelo DMRST [Liu, Shi e Chen, 2021]. Tal modelo realiza a anotação RST com base na segmentação do texto em EDUs e na identificação de relações retóricas entre essas unidades. O modelo identifica a quantidade e a relação entre EDUs (nucleares ou satélites) em cada sentença analisada, resultando em uma estrutura da relação discursiva inferida. O *fine-tuning* realizado buscou aplicar adaptações que refletissem as especificidades ortográficas e morfossintáticas do português e do gênero jurídico.

Assim, o texto foi inicialmente segmentado em sentenças por meio do Stanza [Qi *et al.*, 2020], desconsiderando pontuação, já que o fato de haver ponto-final pode ser compreendido como a abreviação de um artigo e/ou capítulo e não a finalização de uma sentença, por exemplo. Esse pré-processamento foi necessário para que esses elementos não fossem fundidos à sentença subsequente antes de serem processados pelo modelo DMRST.

Como técnica de vetorização textual, o DMRST utiliza o XLM-RoBERTa (2019), um modelo multilíngue de arquitetura *transformer*. A vetorização é necessária porque modelos de aprendizado de máquina não processam texto diretamente, exigindo que as informações linguísticas sejam convertidas em representações numéricas. Dessa forma, o XLM-RoBERTa converte o material textual em *embeddings* contextualizados que capturam informações lexicais, sintáticas e semânticas das sentenças, permitindo ao DMRST identificar os limites das EDUs e as relações discursivas associadas a elas.

Todo o desenvolvimento do sistema e *fine-tuning* do modelo foi realizado em ambiente Python 3.8 (venv38), utilizando o editor *Visual Studio Code* [Uzayr, 2022] como ambiente de desenvolvimento integrado. Foram empregadas as bibliotecas necessárias para execução do DMRST, incluindo PyTorch [Paszke *et al.*, 2019] e *Transformers* [2019], conforme especificado na documentação oficial do modelo.

3. Resultados e discussão

A análise dos resultados foi organizada a partir de três aspectos centrais neste estudo, a saber: (i) a segmentação das unidades discursivas produzidas pelo DMRST; (ii) as diferenças observadas entre as estratégias intra e intersentencial; e (iii) a distribuição das relações retóricas identificadas pelo modelo. Por conta do caráter exploratório deste estudo, os resultados são discutidos como evidências do comportamento do modelo quando aplicado ao domínio jurídico-normativo, e não como medidas de desempenho da anotação automática.

No experimento da análise intrassentencial, observou-se uma segmentação excessivamente fragmentada nas sentenças. Em (1a), primeiro artigo do CDC, por exemplo, o modelo identificou como unidades independentes segmentos como "*inciso XXXII"*, "*170"* e "*inciso V"*, por exemplo. Essas ocorrências sugerem dificuldades do modelo no tratamento de elementos característicos da organização jurídico-normativa, sobretudo referências legais, numerações e remissões a outros dispositivos da lei.

(1)

a) Art. 1° O presente código estabelece normas de proteção e defesa do consumidor, de ordem pública e interesse social, nos termos dos arts. 5°, inciso XXXII, 170, inciso V, da Constituição Federal e art. 48 de suas Disposições Transitórias.
b) Art. 2° Consumidor é toda pessoa física ou jurídica que adquire ou utiliza produto ou serviço como destinatário final.
c) Parágrafo único. Equipara-se a consumidor a coletividade de pessoas, ainda que indetermináveis, que haja intervindo nas relações de consumo.

É possível que esse comportamento esteja relacionado à diferença entre características do domínio jurídico analisado e os dados utilizados no treinamento e desenvolvimento do modelo original do DMRST. Embora o modelo utilize *embeddings* contextualizadas produzidas pelo XLM-RoBERTa e realize conjuntamente a segmentação das EDUs e a análise da estrutura discursiva [Liu, Shi e Chen, 2021], os resultados observados não permitem determinar separadamente em que medida a tokenização, os *embeddings* ou as

particularidades estruturais do CDC contribuíram para a fragmentação identificada. Liu, Shi e Chen (2021) destacam, inclusive, que *parsers* treinados em *treebanks* de um único domínio podem apresentar problemas de generalização *out-of-domain*.

Diante desses resultados, realizou-se uma segunda análise, buscando inferir relações em nível intersentencial. Nessa estratégia, o modelo analisou pares de sentenças consecutivas e, nesse caso, cada sentença seria uma EDU diferente. Essa decisão visou investigar se unidades textuais mais extensas e contextualmente mais informativas reduziriam os efeitos da fragmentação observada na análise intrassentencial. Entre (1a) e (1b), por exemplo, o modelo inferiu a relação *Joint,* que é multinuclear, indicando que ambas as unidades possuem importância equivalente, conforme ilustrado na Figura 2. No caso, (1a), que atua como a EDU 1, apresenta o objetivo e os fundamentos legais do CDC, enquanto (1b) º, atuando como a EDU 2, introduz a definição de consumidor, estabelecendo uma continuidade temática sem que haja uma relação explícita de dependência hierárquica entre os segmentos.

A atribuição da relação RST *Joint* pelo modelo é compatível com a ausência de uma relação explícita de dependência hierárquica entre (1a) e (1b) e com a continuidade temática estabelecida entre elas. Essa interpretação, no entanto, deve ser considerada uma hipótese de análise, já que não foi confrontada com uma anotação padrão-ouro.

Nesse sentido, na análise intersentencial, a utilização de sentenças completas como unidades de análise produziu estruturas discursivas menos fragmentadas, em que é possível propor interpretações a partir da estrutura retórica delas. Difernte da abordagem intrassentencial, em que foram identificadas EDUs constituídas por segmentos muito curtos, a estratégia intersentencial preservou maior quantidade de informação linguística em cada unidade. Isso favoreceu a análise das relações atribuídas pelo modelo. Esse comportamento sugere que unidades textuais maiores podem fornecer ao modelo maior contexto linguístico para a geração dos *embeddings* pelo XLM-RoBERTa, reduzindo os efeitos da fragmentação observada na análise intrassentencial.

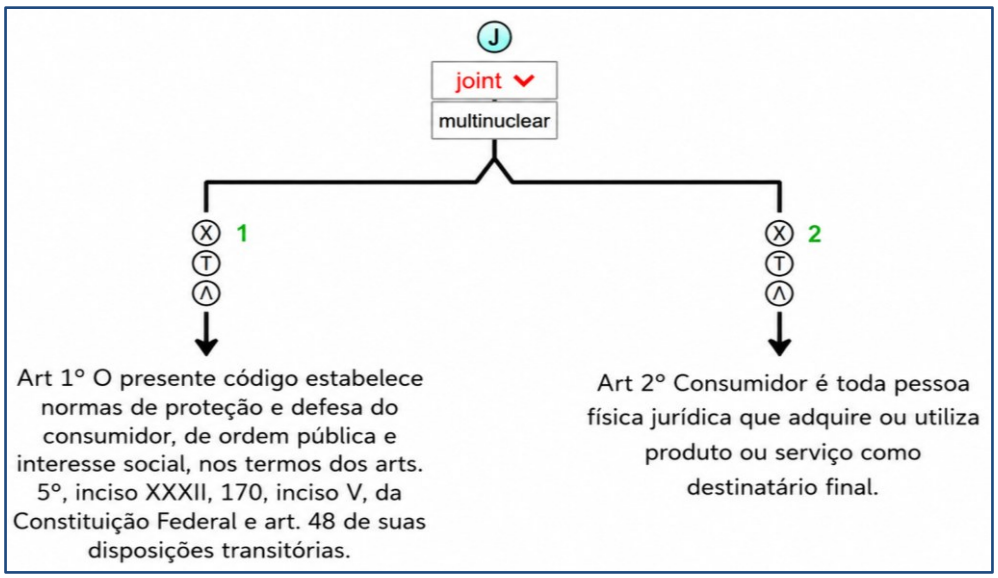


Figura 2. Exemplo de análise intersentencial entre sentenças consecutivas do CDC.

Os resultados da Tabela 1 sintetizam a distribuição das relações eRST por capítulo do CDC na análise intersentencial. Observa-se que a relação Joint foi a mais frequente em todos os capítulos, especialmente no Capítulo 2, o que pode ser explicado pela própria organização do texto normativo, cujos artigos são apresentados de forma sequencial e frequentemente abordam temas relacionados sem estabelecer uma dependência retórica explícita.

Nesse contexto, ao analisar pares de sentenças consecutivas, o modelo tende a interpretar os segmentos como unidades de mesma relevância discursiva, favorecendo a atribuição da relação *Joint***.** Além disso, foram identificadas relações como *Elaboration*, *Background*, *Condition* e *Contrast*, indicando que o modelo também produziu diferentes tipos de relações discursivas ao processar o texto jurídico. Entretanto, a análise qualitativa dos resultados revelou situações em que relações classificadas como Elaboration seriam mais adequadamente interpretadas como Joint, indicando dificuldades do modelo em distinguir casos de detalhamento informacional daqueles em que existe apenas continuidade temática entre os segmentos, característica recorrente em textos normativos.

Tabela 1. Distribuição das relações eRST por Capítulo na análise interssentencial.

RELAÇÃO RST	CAPÍTULO 1	CAPÍTULO 2	CAPÍTULO 3
Condition	0	0	2
Background	1	1	1
Joint	2	11	7
Elaboration	2	3	4
Contrast	1	0	2
Text organization	0	1	0
Sem identificação	0	5	0

A distribuição apresentada na Tabela 1 evidencia uma concentração da relação *Joint*, com 20 ocorrências entre os casos analisados. Esse resultado pode estar relacionado à organização sequencial característica do texto normativo, em que os artigos e dispositivos consecutivos frequentemente acrescentam e/ou detalham informações juridicamente relacionadas. Nesse caso, não estabelecem, necessariamente, uma relação de dependência entre si. Entretanto, a predominância de *Joint* também pode refletir a própria decisão metodológica adotada na abordagem intersentencial, em que sentenças consecutivas foram tomadas como EDUs. Desse modo, não foi possível determinar se a frequência observada decorre predominantemente de uma propriedade discursiva do corpus analisado ou do comportamento do modelo diante da estratégia de segmentação utilizada.

4. Considerações finais

Neste trabalho, investigou-se, em caráter exploratório, a aplicação do modelo DMRST na identificação automática de relações retóricas em textos do CDC, considerando estratégias de análise intra e intersentencial. Os experimentos evidenciaram que a adaptação de modelos de análise discursiva para o domínio jurídico apresenta desafios específicos, principalmente em razão da estrutura normativa dos textos. Como forma de minimizar esses desafios, foi utilizada uma etapa de pré-processamento com o Stanza para segmentação das sentenças e tratamento das particularidades do texto jurídico.

Os resultados evidenciaram comportamentos distintos entre as duas estratégias de análise. Na abordagem intrassentencial, foram observados casos de segmentação bastante fragmentada, numerações e remissões a outros dispositivos. Na abordagem inersentencial, por sua vez, a utilização de sentenças completas como EDUs reduziu esse tipo de fragmentação e produziu estruturas discursivas passíveis de interpretação qualitativa. Esses resultados, entretanto, não permitiram afirmar que a abordagem intersentencial apresenta desempenho superior, já que as estruturas produzidas pelo modelo não foram comparadas a uma anotação de referência.

A distribuição das relações identificadas pelo DMRST ainda demonstrou uma predominância da relação *Joint*, com 20 ocorrências na amostra analisada. Ainda que esse resultado possa estar relacionado à organização sequencial do gênero textual em questão, a frequência dessa relação também pode decorrer da própria estratégia intersentencial adotada no experimento. Assim, os dados ainda não permitem determinar em que medida esse padrão constitui uma característica discursiva do CDC ou um comportamento do modelo diante das EDUs analisadas.

Ainda, os experimentos indicam possibilidades de poder aplicar o DMRST à análise discursiva de textos jurídicos, mas também evidenciaram a necessidade de adaptações para esse domínio. As análises construídas automaticamente pelo DMRST permitiram identificar fenômenos que deverão ser considerados nas etapas posteriores da pesquisa, sobretudo dificuldades de segmentação associadas à organização retórica desse gênero textual, além de possíveis efeitos da estratégia de construção das EDUs sobre as relações inferidas pelo modelo.

Assim, como continuidade deste trabalho, pretende-se construir uma amostra do CDC manualmente anotada segundo a RST, estabelecendo um padrão-ouro que permita avaliar quantitativamente a segmentação e a inferência das relações retóricas produzidas pelo DMRST. Essa etapa vai permitir verificar tanto a adequação das relações identificadas automaticamente quanto as hipóteses levantadas neste estudo, incluindo a predominância da relação Joint, além das diferenças observadas entre as estratégias intra e intersentencial.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Além disso, o trabalho contou com suporte do pedido 5065 no processo PPP0031/2024 da Chamada FAPESB/CNPq 004/2023 –Programa Primeiros Projetos.

Referências

Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D., Souza, M. e Pita, R. (2024). Brazilian Consumer Protection Code: a methodology for a dataset to Question-Answer (QA) Models. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, p. 493–500. Porto Alegre: SBC. DOI: 10.5753/stil.2024.31168.

Brasil (1990). Lei nº 8.078, de 11 de setembro de 1990. Código de Defesa do Consumidor. *Diário Oficial da União*, Brasília, DF, 12 set. 1990.

Cardoso, P. C. F. et al. (2011). CSTNews: A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, p. 88–105. Cuiabá, Brasil.

Cardoso, P. C. F., Souza, J. W. C. e Rodrigues, R. (2024). Modelos discursivos. In Caseli, H. M. e Nunes, M. G. V. (eds.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3. ed., cap. 11. BPLN.

Conneau, A. et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451. Online: Association for Computational Linguistics.

Gururangan, S. et al. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360. Online: Association for Computational Linguistics.

Liu, Z., Shi, Y. e Chen, N. F. (2021). DMRST: A joint framework for document-level multilingual rhetorical structure parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, p. 154–164. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics.

Mann, W. C. e Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3), p. 243–281.

Pardo, T. A. S. e Seno, E. R. M. (2005). Rhetalho: Um corpus de referência anotado retoricamente. In *Anais do V Encontro de Corpora*, p. 24–25.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, v. 32, p. 8024–8035.

Pereira, M. A. e Souza, J. W. C. (2025). Anotação Enhanced Rhetorical Structure Theory em textos de User-Generated Content. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, p. 713–717. Porto Alegre: SBC. DOI: 10.5753/stil.2025.37877.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. e Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, p. 101–108. Online: Association for Computational Linguistics.

Souza, J. W. C., Cardoso, P. C. F. e Rodrigues, R. (2024). Revisão sistemática de estudos sobre Rhetorical Structure Theory (RST). *Revista de Estudos da Linguagem*, 31(3), p. 1643–1675. DOI: 10.17851/2237-2083.31.3.1643-1675.

Zhang, R., Li, Y., Ma, Y., Zhou, M. e Zou, L. (2023). LLMaAA: Making Large Language Models as Active Annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 13088–13103. Singapore: Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-emnlp.872.

Uma Metodologia de Extração de Dados Dialetais de Documentos Científicos do Atlas Linguístico do Brasil

Alfredo Victor do N. S. Sena[1], Amanda Sammer do N. Brandão[1], Felipe B. S. dos Reis[1], Laila P. M. Santos[1], Daniela B. Claro[1]

[1]FORMAS Research Center on Natural Language Institute of Computing – Federal University of Bahia (UFBA) Salvador – BA – Brazil

{alfredosena,amandabrandao,felipe.sanches,laila.pereira,
dclaro}@ufba.b

Abstract. *Dialectal data on Brazilian Portuguese documented by the ALiB (Atlas Linguístico do Brasil) Project are scattered across articles, master's dissertations, and doctoral theses, making their retrieval challenging. This study proposes a methodology based on Large Language Models (LLMs) and Prompt Engineering to extract and organize lexical variants, Semantic-Lexical Questionnaire (QSL) items, localities, and bibliographic references into a structured format. The extracted data are subsequently validated by human experts. Preliminary results indicate the potential of this approach to support geolinguistic surveys and the development of dialectal knowledge bases for Brazilian Portuguese.*

Resumo. *Os dados dialetais do português brasileiro documentados pelo Projeto Atlas Linguístico do Brasil (ALiB) estão dispersos em artigos, dissertações e teses, dificultando sua recuperação. Este trabalho propõe uma metodologia baseada em Modelos de Linguagem de Grande Escala (LLMs) e Engenharia de Prompt para extrair e organizar variantes lexicais do Questionário Semântico-Lexical (QSL), localidades e referências bibliográficas em formato estruturado. Os dados extraídos passam por validação humana. Resultados preliminares indicam potencial para apoiar levantamentos geolinguísticos e a construção de base de conhecimento dialetal.*

1. Introdução

A Geolinguística tem desempenhado um papel central na documentação, descrição e análise da diversidade linguística do português brasileiro [Cardoso 2012]. Segundo os autores [Silva and Romano 2026], o foco tem sido a investigação da variação linguística em sua dimensão espacial, buscando representar a distribuição diatópica de variantes lexicais, fonético-fonológicas e morfossintáticas. No contexto brasileiro, esse campo de investigação assume particular relevância em razão da extensão territorial e da diversidade linguística e sociocultural do país, contribuindo para a descrição das diferentes variedades do português falado no território nacional. Nesse âmbito, o Projeto Atlas Linguístico do Brasil (ALiB) estabeleceu uma metodologia de alcance nacional voltada à coleta, organização e representação de dados dialetais, contribuindo de forma significativa para os estudos da variação lexical, fonética e morfossintática do português falado no país [da Silva and Romano 2024]. Com base nos dados reunidos pelo projeto, diversos trabalhos investigam campos semânticos específicos, registrando variantes

lexicais distribuídos por diferentes regiões brasileiras e ampliando o conhecimento acerca da diversidade linguística nacional.

Embora o ALiB seja uma das principais referências da Geolinguística brasileira, parte do conhecimento produzido a partir de seus dados permanece dispersa em artigos, dissertações e teses, dificultando a recuperação sistemática das variantes lexicais e a comparação entre estudos [de Oliveira et al. 2018]. Nesse âmbito, os Modelos de Linguagem em Grande Escala (LLMs) oferecem novas possibilidades para recuperação, organização e estruturação de informações especializadas [Jia 2023, Ghali et al. 2025]. No campo da Linguística, pesquisas também têm investigado o uso desses modelos em tarefas relacionadas à identificação de variedades linguísticas, indicando possibilidades de aplicação em estudos dialetais [Freitag and Gois 2024].

Diante deste cenário, o presente trabalho propõe uma metodologia fundamentada em Engenharia de *Prompt* e LLMs para apoiar a recuperação e a organização de dados dialetais presentes em artigos científicos, dissertações e teses baseadas nos dados do Projeto ALiB. A metodologia orienta o modelo a identificar variantes lexicais, seus respectivos itens do Questionário Semântico-Lexical (QSL), as localidades associadas e as referências bibliográficas correspondentes, organizando essas informações em uma estrutura previamente definida. O objetivo não é substituir a análise realizada por especialistas, mas investigar o potencial dos LLMs como ferramentas auxiliares na sistematização de informações geolinguísticas dispersas e no apoio as pesquisas sobre a variação lexical do português brasileiro. Assim, embora desenvolvida no contexto do Projeto ALiB, a metodologia proposta apresenta uma estrutura passível de adaptação a outros conjuntos documentais e iniciativas de pesquisa linguística com características semelhantes.

2. Geolinguística e o Projeto ALiB

A Geolinguística é um dos principais campos da Dialetologia que investiga a distribuição espacial dos fenômenos linguísticos por meio de procedimentos sistemáticos de coleta, análise e representação cartográfica dos dados. Tradicionalmente, seus estudos resultam na elaboração de atlas linguísticos, que registram e representam a distribuição territorial das variantes linguísticas a partir de entrevistas realizadas em diferentes localidades segundo critérios metodológicos padronizados [Paim and Ribeiro 2018].

O Projeto ALiB, principal vetor da Geolinguística no Brasil, reúne pesquisadores de diversas instituições e constitui a principal referência para estudos da variação lexical, fonética e morfossintática do português brasileiro. Além de documentar a diversidade linguística nacional, o projeto consolidou procedimentos metodológicos que se tornaram referência para pesquisas geolinguísticas regionais [da Silva and Romano 2024].

A coleta de dados do ALiB é realizada por meio de entrevistas presenciais e instrumentos específicos, entre os quais se destaca o Questionário Semântico-Lexical (QSL), composto por perguntas organizadas em diferentes campos semânticos. Os dados obtidos têm subsidiado numerosas pesquisas sobre a variação lexical do português brasileiro, abrangendo temas como acidentes geográficos, alimentação, corpo humano, fauna e fenômenos atmosféricos [Sanches et al. 2018, Pereira Romano 2022, do Socorro Silva Espíndola 2021, Melo and Pereira 2024].

Após a coleta, os dados são organizados e divulgados em um atlas [Cardoso et al. 2014], livros, dissertações, teses e artigos científicos. Entretanto, essa produção encontra-se dispersa em diversas publicações, o que dificulta a recuperação sistemática das variantes linguísticas e reforça a necessidade de metodologias computacionais capazes de auxiliar na localização, organização e consolidação dessas informações.

3. Trabalhos Relacionados

As tarefas de PLN tem ampliado significativamente as possibilidades de recuperação, organização e extração de conhecimento a partir de documentos textuais com o uso dos LLMs. Em diferentes domínios científicos, essas tecnologias vêm sendo utilizadas para localizar informações especializadas, sintetizar grandes volumes de dados e estruturar conteúdos anteriormente dispersos em coleções documentais.

Na área de recuperação de conhecimento, [Ghali et al. 2025] propõem uma arquitetura que integra busca semântica, bancos vetoriais e aprendizagem em contexto (*In-Context Learning*), demonstrando que a combinação entre mecanismos de recuperação e LLMs pode produzir respostas mais precisas sem a necessidade de retreinamento dos modelos. Os resultados demonstram o potencial dessas tecnologias para consolidar informações provenientes de múltiplas fontes documentais, característica particularmente relevante para domínios científicos cuja literatura encontra-se distribuída em diferentes publicações.

Em uma perspectiva semelhante, a revisão sistemática realizada por [Abu-Salih et al. 2025] demonstra que os LLMs têm sido empregados para promover interoperabilidade semântica entre bases heterogêneas de conhecimento, permitindo alinhar diferentes estruturas de informação e facilitar o compartilhamento de dados entre sistemas distintos. Os autores destacam que, apesar dos avanços, aplicações especializadas ainda dependem de estratégias de validação humana para garantir consistência, confiabilidade e rastreabilidade das informações recuperadas.

No âmbito da extração de informação, [Chun et al. 2025] investigam o uso de LLMs na Extração Automática de Termos (*Automatic Term Extraction* - ATE). Os autores demonstram que estratégias de recuperação baseadas em similaridade sintática produzem resultados superiores às abordagens tradicionais fundamentadas apenas em similaridade semântica, o que reforça a importância da Engenharia de *Prompt* e da seleção criteriosa de exemplos para aumentar a precisão da recuperação textual em domínios especializados.

Também, na área biomédica, [Xu et al. 2025] apresentam um amplo levantamento sobre aplicações de PLN na Medicina Baseada em Evidências, demonstrando que técnicas de recuperação, sumarização, extração de evidências e organização automática da literatura científica vêm desempenhando papel crescente no apoio à análise documental. Embora direcionado ao domínio médico, o estudo evidencia como metodologias baseadas em PLN podem reduzir o esforço necessário para explorar grandes coleções de documentos científicos.

No contexto específico da Linguística, [Jia 2023] discute as transformações promovidas pela Inteligência Artificial na pesquisa linguística contemporânea, destacando que ferramentas baseadas em LLMs tendem a modificar processos tradicionais de análise, documentação e organização de recursos linguísticos. Entre as

possibilidades apontadas pelo autor estão a automatização da recuperação de conhecimento, a estruturação de bases lexicais e o apoio à pesquisa em Linguística Computacional.

Mais especificamente para o português brasileiro, [Freitag and Gois 2024] investigam o desempenho de LLMs em tarefas de identificação de variedades dialetais. Os resultados demonstram que esses modelos já apresentam capacidade para reconhecer características linguísticas regionais, embora ainda existam limitações relacionadas à precisão em variedades menos representadas. O estudo representa um dos primeiros esforços em aproximar modelos de linguagem dos estudos dialetais brasileiros.

Apesar desses avanços, observa-se que a literatura concentra-se principalmente em tarefas como recuperação de conhecimento, busca semântica, extração terminológica, sumarização ou classificação textual. Não foram identificados trabalhos que proponham uma metodologia voltada especificamente para a recuperação automatizada de variantes lexicais documentadas em publicações derivadas do ALiB, preservando simultaneamente a associação entre conceito, variante lexical, distribuição geográfica e referência bibliográfica.

Dessa forma, este trabalho diferencia-se por investigar o uso de Engenharia de *Prompt* aplicada a LLMs como estratégia metodológica para localizar, extrair e estruturar informações dialetais dispersas em artigos, dissertações e teses baseadas nos dados do ALiB. Em vez de produzir novos dados linguísticos ou treinar modelos específicos, a proposta concentra-se na sistematização automática do conhecimento já publicado, mantendo a validação humana como etapa indispensável para assegurar a fidelidade das informações recuperadas.

4. Metodologia

Esta pesquisa caracteriza-se como um estudo de natureza metodológica, cujo objetivo é investigar o uso de LLMs como ferramenta de apoio à recuperação e organização de dados dialetais documentados em publicações científicas relacionadas ao Atlas. A metodologia integra levantamento bibliográfico, Engenharia de *Prompt*, recuperação e extração automatizada de informações e validação humana, buscando avaliar o potencial dessas tecnologias na sistematização da literatura geolinguística.

Os documentos não constituíram previamente um corpus fechado nem foram fornecidos ao modelo por meio do envio de arquivos PDF. A recuperação documental foi realizada pelo próprio LLM a partir das instruções definidas nos *prompts*, que orientavam a busca no portal oficial do ALiB e em publicações relacionadas ao projeto. Foram consideradas, como potenciais fontes, artigos científicos, dissertações, teses, livros, cartas linguísticas e outros documentos que apresentassem informações lexicais relacionadas aos itens do QSL.

4.1. Caracterização do Conjunto Documental

A recuperação documental foi realizada de forma iterativa, a partir de dois *prompts*. Na primeira etapa, o LLM foi orientado a localizar publicações relacionadas ao ALiB e aos itens do QSL, resultando em 14 referências. Na segunda etapa, com a ampliação da recuperação para diferentes blocos do QSL, o modelo indicou referências adicionais que

não constavam no primeiro conjunto. Assim, o conjunto documental foi sendo ampliado ao longo da execução dos *prompts*, de acordo com as fontes localizadas pelo modelo.

As referências recuperadas contemplam diferentes tipos de produção acadêmica, incluindo artigos, dissertações e teses. Entretanto, nem todas as referências adicionais apresentadas pelo modelo possuem informações bibliográficas suficientes para determinar com segurança sua classificação documental. Por essa razão, não é apresentada uma distribuição quantitativa dos trabalhos por tipo de publicação. Os documentos recuperados abordam diferentes campos semânticos do QSL e estudos relacionados à variação linguística, à variação lexical, aos estudos semântico-lexicais e à geolinguística.

Como critérios de inclusão, foram consideradas publicações relacionadas ao ALiB ou dele derivadas que apresentassem dados lexicais associados aos itens do QSL. Os LLMs foram empregados para localizar e estruturar informações previamente publicadas, incluindo variantes lexicais, localidades e referências bibliográficas, não sendo utilizados para gerar novos dados linguísticos.

4.2. Engenharia de *Prompt*

A Engenharia de *Prompt* compreende o conjunto de estratégias empregadas na formulação e estruturação de instruções destinadas a orientar LLMs na execução de tarefas específicas. Na literatura recente, diferentes técnicas de *prompting* têm sido sistematizadas e classificadas segundo a forma de construção das instruções, o contexto fornecido ao modelo e as características esperadas para a saída [Schulhoff et al. 2024]. Além disso, padrões de construção de *prompts* têm sido propostos para orientar a definição de instruções, restrições e formatos de saída de acordo com as características da tarefa [White et al. 2023].

Após a seleção dos documentos, foi desenvolvida uma estratégia de Engenharia de *Prompt* para orientar o LLM no processo de recuperação e organização das informações. A formulação do *prompt* foi refinada progressivamente a partir dos resultados observados nos testes exploratórios e das discussões metodológicas realizadas com pesquisadores vinculados ao Projeto ALiB. Esses refinamentos buscaram aprimorar a capacidade do modelo de identificar as variantes lexicais, associá-las aos respectivos itens do Questionário Semântico-Lexical (QSL) e preservar sua relação com as informações geográficas e bibliográficas apresentadas nas publicações.

A versão final do *prompt* foi definida qualitativamente com base nos testes exploratórios, considerando sua aderência à estrutura de saída estabelecida e sua capacidade de preservar a relação entre variante lexical, localização geográfica e referência bibliográfica. Não foi realizada, nesta etapa, uma comparação quantitativa sistemática entre as diferentes versões do *prompt*.

Os refinamentos foram realizados ao longo dos próprios testes exploratórios, não tendo sido estabelecido, nesta etapa do estudo, um conjunto documental independente destinado exclusivamente à avaliação da metodologia. Essa ausência de um conjunto de avaliação independente constitui uma limitação do experimento e será considerada em avaliações futuras, nas quais versões alternativas do *prompt* poderão ser comparadas sob condições controladas.

O *prompt* foi construído com instruções explícitas para restringir a atuação do modelo à identificação e extração de informações documentadas nas publicações analisadas. Entre as principais instruções fornecidas ao modelo destacam-se: localizar a variante lexical registrada; identificar o item correspondente do QSL; recuperar a localidade geográfica associada à variante, incluindo estado e região; registrar a referência bibliográfica completa; e informar a fonte consultada.

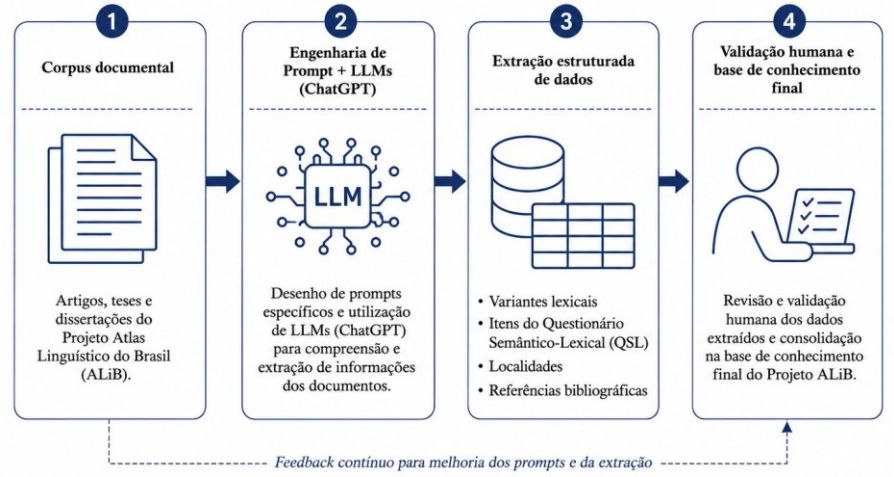


Figura 1. Metodologia baseada em LLMs para extração de dados

Trecho representativo do *prompt* utilizado[8]: Você é um assistente especializado em linguística dialetal e análise de dados sociolinguísticos. Sua tarefa é realizar um levantamento exaustivo e sistemático das variações lexicais registradas no Projeto ALiB. O objetivo é mapear as palavras registradas para cada item do Questionário Semântico-Lexical, identificando: (i) a variante lexical exata; (ii) a região geográfica correspondente; (iii) o(s) estado(s) de ocorrência; e (iv) as referências bibliográficas utilizadas. A saída deve ser apresentada em formato tabular contendo os campos: Item do QSL, Palavra/Variação, Estado(s), Região e Referência.

Também foram incorporadas restrições destinadas a reduzir respostas imprecisas e inferências não documentadas. O modelo foi instruído a preservar a grafia original das variantes lexicais, evitar normalizações linguísticas, registrar múltiplas variantes quando presentes em uma mesma publicação e indicar explicitamente as referências utilizadas para cada ocorrência recuperada. Essas restrições foram empregadas para manter o vínculo entre as unidades lexicais identificadas e as evidências documentais utilizadas pelo modelo, favorecendo a posterior conferência dos resultados. Tais estratégias estão alinhadas aos princípios de estruturação de instruções, definição de restrições e especificação das características de saída discutidos na literatura sobre Engenharia de *Prompt* [Schulhoff et al. 2024, White et al. 2023]. A abordagem também é compatível com trabalhos que empregam Engenharia de *Prompt* em tarefas de recuperação e extração de informações em domínios especializados [Chun et al. 2025].

[8] Prompt completo disponível em: https://github.com/avsena/ELC-2026

4.3. Extração e Organização dos Dados

A etapa de extração foi realizada utilizando assistentes baseados em LLMs, empregados como ferramentas auxiliares de recuperação, identificação e estruturação de informações documentais. Para cada documento analisado, o modelo recebeu o *prompt* elaborado na etapa anterior e foi instruído as variantes lexicais documentadas na publicação.

Como formato de saída, definiu-se uma estrutura tabular composta pelos seguintes campos: questionário semântico-lexical (QSL), palavra ou variante lexical, Estado(s) de ocorrência, região geográfica e referência da publicação.

A padronização dessas informações teve como objetivo preservar o vínculo entre a variante lexical, sua distribuição geográfica e a fonte bibliográfica correspondente, facilitando posteriores consultas, comparações e processos de validação. A estrutura tabular também favorece a reutilização dos dados em estudos futuros e possibilita a consolidação de informações originalmente distribuídas em diferentes documentos científicos.

4.4. Execução do Experimento

Os experimentos foram realizados por meio da plataforma ChatGPT (OpenIA), utilizando o modelo GPT-5 [Singh et al. 2026] na plataforma em novembro de 2025. Foi utilizado um *prompt* abrangente destinado à recuperação de informações referentes a todos os blocos do QSL do Projeto ALiB. Utilizamos a interface do ChatGPT, sem uso de API e sem fine-tuning, opção compatível com o caráter exploratório do estudo e com a abordagem baseada em *prompt*.

O processamento da consulta demandou aproximadamente 1 minuto e 23 segundos. Como saída, o modelo produziu uma tabela estruturada contendo informações dialetais organizadas por blocos temáticos do QSL, incluindo variantes lexicais, regiões geográficas e referências bibliográficas quando disponíveis.

Apesar de o *prompt* orientar o modelo a realizar uma recuperação ampla e sistemática das variantes documentadas, o primeiro experimento teve caráter exploratório e buscou avaliar a capacidade inicial do modelo de localizar e organizar automaticamente essas informações antes dos refinamentos sucessivos da estratégia de Engenharia de *Prompt*.

4.5. Validação dos Resultados

Embora os LLMs tenham sido utilizados na recuperação inicial das informações, todos os registros selecionados foram submetidos à validação manual. Foram verificadas a correspondência entre conceito e variante lexical, a localização geográfica e a exatidão das referências bibliográficas, a fim de identificar possíveis omissões, inconsistências ou alucinações. Dessa forma, os LLMs foram empregados como ferramenta de apoio à recuperação e organização do conhecimento linguístico, mantendo-se a supervisão humana como elemento essencial para garantir a confiabilidade dos dados [Ghali et al. 2025, Abu-Salih et al. 2025, Freitag and Gois 2024].

A validação dos resultados contou com a participação de dois professores vinculados ao Projeto ALiB, com formação e experiência na área de Linguística. Durante

essa etapa, foram discutidos os seguintes dados recuperados pelo LLM e suas variantes: “sanga”, “arroio”, “igarapé” e “córrego”. Observou-se a relação com as localidades e as referências indicadas. A análise contribuiu para a identificação de inconsistências e para o aprimoramento da estratégia de extração.

5. Resultados Preliminares

A comparação entre as duas versões do *prompt* evidenciou mudanças no comportamento do LLM durante a recuperação e organização dos dados dialetais. A primeira execução, realizada com a versão inicial do *prompt*, apresentou caráter predominantemente exploratório e concentrou-se inicialmente no bloco de Acidentes Geográficos. No QSL 1, cuja pergunta investiga “um rio pequeno, de uns dois metros de largura?”, foram recuperados registros de variantes como igarapé, córrego, lago/lagoa, riacho, rio/riozinho, grota/grotão, ribeirão e paraná/braço. Entretanto, a variante sanga não foi recuperada nessa primeira execução, sendo posteriormente identificada pela equipe envolvida na pesquisa.

A partir dessa observação, o *prompt* foi refinado com instruções mais específicas para a identificação dos itens do QSL, recuperação das variantes, localização das fontes e associação dos registros às respectivas referências bibliográficas. Na segunda execução, o modelo passou a consultar e relacionar diferentes trabalhos que utilizam dados do Projeto ALiB, ampliando a diversidade das fontes recuperadas. Para o QSL 1, por exemplo, além das variantes apresentadas por [Sanches et al. 2018], foram recuperadas as formas sanga e arroio, associadas ao trabalho de [Romano 2017].

Para exemplificar os resultados da segunda execução, foram selecionados itens do QSL pertencentes ao bloco de Acidentes Geográficos e a outros itens relacionados a fenômenos atmosféricos. O QSL 1 corresponde à pergunta um rio pequeno, de uns dois metros de largura?”; o QSL 2 investiga o tronco, pedaço de pau ou tábua que serve para passar por cima de um...”; o QSL 3 refere-se ao lugar onde o rio termina ou encontra com outro rio”; e o QSL 4 aborda o fenômeno em que, muitas vezes, num rio, a água começa a girar, formando um buraco, na água, que puxa para baixo”. Para os demais exemplos apresentados no quadro, o QSL 12 solicita outros nomes para temporal”; o QSL 13 investiga uma chuva de pouca duração, muito forte e pesada”; o QSL 15 trata da chuva em que “podem cair bolinhas de gelo”; e o QSL 17 refere-se à faixa colorida que aparece no céu depois da chuva, isto é, o arco-íris.

Como resultado, foi produzida uma tabela contendo os campos Item do QSL, Palavra/Variação, Estado(s), Região e Referência (artigo, livro, tese, dissertação, página ou carta linguística). Essa organização permitiu reunir informações originalmente dispersas em diferentes tipos de documentos científicos, facilitando sua consulta e posterior conferência.

Os resultados apresentados nesta etapa não foram utilizados para calcular medidas de precisão, revocação ou acurácia, uma vez que não foi realizada uma classificação exaustiva de todos os registros em categorias de acerto e erro. A conferência humana ocorreu de forma direcionada, especialmente para a identificação de omissões e inconsistências observadas durante as execuções e para o refinamento dos *prompts*. Dessa forma, os resultados quantitativos apresentados devem ser compreendidos como registros recuperados pelo modelo, e não como medidas de desempenho ou de correção dos dados.

O Quadro 1 apresenta variantes lexicais recuperadas na segunda execução do *prompt*, abrangendo diferentes questões do QSL e fontes do ALiB. Os exemplos evidenciam a diversidade de documentos, regiões e itens recuperados.

Observa-se que um mesmo conceito pode apresentar diferentes denominações conforme a região investigada, evidenciando a diversidade lexical documentada nos trabalhos relacionados ao ALiB. Destaca-se, ainda, que as variantes recuperadas não se concentraram em uma única fonte documental. No caso do QSL 1, por exemplo, igarapé e córrego foram associados a [Sanches et al. 2018], enquanto sanga e arroio foram associados a [Romano 2017], evidenciando a recuperação de informações provenientes de trabalhos distintos.

Quadro 1. Variantes lexicais recuperadas pelo segundo *prompt*.

Item do QSL	Palavra/Variação	Estado(s)	Região	Referência
QSL 1	igarapé	AP, AM, PA, AC, RO	Norte	Sanches et al. (2018)
QSL 1	córrego	AP, AM, PA, TO	Norte	Sanches et al. (2018)
QSL 1	sanga / arroio	PR, SC, RS	Sul/Centro-Sul	Romano (2018)
QSL 2	ponte	Nordeste	Nordeste	Estudo da variação do item lexical
QSL 3	foz	PR	Sul	Semântico-lexical de Foz do Iguaçu/PR
QSL 4	redemoinho	Diversos estados	Diversas regiões	Documentos do Projeto ALiB; Carta 230
QSL 12	tempestade / vendaval	Diversos estados	Centro-Oeste/Norte	Carvalho (2015) e atlas regionais
QSL 12	temporal	Diversos estados	Centro-Oeste/Norte	Carvalho (2015) e atlas regionais
QSL 13	panca	MT e outras localidades	Centro-Oeste/Norte	Azevedo/UNEMAT
QSL 13	pancada de chuva	MT e outras localidades	Centro-Oeste/Norte	Azevedo/UNEMAT
QSL 13	manga de chuva	MT e outras localidades	Centro-Oeste/Norte	Azevedo/UNEMAT
QSL 15	granizo / granito	Diversos estados	Diversas regiões	FS Lima e documentos do ALiB
QSL 17	arco-da-velha / arco-da-aliança	Capitais brasileiras	Diversas regiões	Maria do Socorro Aragão; dados do Projeto ALiB

A organização dos resultados preservou o vínculo entre os registros recuperados e as referências indicadas pelo modelo, permitindo identificar as fontes associadas às informações e facilitando sua conferência posterior. Os registros apresentados correspondem aos resultados da recuperação automatizada realizada pelo LLM. A própria resposta do segundo *prompt* indicou que a identificação das variantes por município e informante demandaria a consulta às cartas, volumes e tabelas primárias do ALiB.

6. Discussão

Os resultados preliminares indicam que a metodologia proposta foi capaz de recuperar e estruturar informações dialetais provenientes de diferentes publicações baseadas no Projeto ALiB. A estratégia de Engenharia de *Prompt* permitiu identificar variantes lexicais, localidades associadas e referências bibliográficas, produzindo uma saída padronizada adequada para consulta e organização dos dados.

O desempenho foi mais consistente em documentos que apresentavam tabelas, cartas linguísticas ou listas explícitas de variantes lexicais. Em contrapartida, publicações com resultados predominantemente descritivos exigiram maior conferência manual para verificar a associação entre conceitos, variantes e localidades.

A estruturação automática das informações mostrou potencial para reduzir o esforço necessário ao levantamento bibliográfico e à organização de dados geolingüísticas dispersas. Entretanto, foram observadas inconsistências e omissões que reforçam a necessidade de validação humana. Assim, os LLMs devem ser compreendidos como ferramentas de apoio metodológico, capazes de auxiliar a recuperação e organização das informações, mas não de substituir a análise especializada. Embora aplicada ao contexto do ALiB, a metodologia pode ser adaptada a outros atlas linguísticos e conjuntos documentais, mediante ajustes no *prompt* e na estrutura de saída.

7. Limitações

Durante a aplicação da metodologia foram observadas limitações inerentes ao emprego de Modelos de Linguagem de Grande Escala na recuperação de dados dialetais. Em diferentes momentos, o modelo deixou de recuperar todas as variantes lexicais presentes nas publicações consultadas, exigindo refinamentos sucessivos da estratégia de Engenharia de *Prompt*.

Os resultados também devem ser interpretados considerando algumas ameaças à validade do estudo. A metodologia depende da recuperação automática de documentos acessíveis pelos mecanismos de busca utilizados pelos LLMs, de modo que publicações indisponíveis em formato digital ou com acesso restrito podem não ter sido consideradas. Além disso, o comportamento não determinístico desses modelos implica que alterações no *prompt*, na versão do sistema ou no ambiente de execução podem produzir respostas distintas para uma mesma consulta.

Outra ameaça à validade decorre do fato de a metodologia ter sido avaliada exclusivamente em publicações derivadas do ALiB. Embora o procedimento seja potencialmente aplicável a outros atlas linguísticos ou coleções documentais, sua generalização ainda requer investigações adicionais. Por fim, a conferência manual das informações, embora aumente a confiabilidade dos resultados, permanece sujeita à interpretação dos pesquisadores durante o processo de validação. Essas limitações podem afetar a cobertura, a estabilidade e a confiabilidade dos resultados recuperados, reforçando a necessidade de validação humana e de avaliações adicionais da metodologia.

Como continuidade desta pesquisa, pretende-se estabelecer conjuntos documentais independentes para o desenvolvimento e a avaliação da metodologia, permitindo uma análise mais controlada de seu desempenho. Trabalhos futuros também deverão comparar a abordagem baseada em LLMs com estratégias alternativas, como busca manual assistida, métodos baseados em regras e outros modelos de linguagem, além de realizar execuções repetidas para avaliar a estabilidade dos resultados. Pretende-se ainda investigar o efeito de diferentes configurações do *prompt* e calcular métricas formais de avaliação, como precisão, cobertura (recall) e taxa de erro, que não foram mensuradas nesta etapa do estudo.

Adicionalmente, trabalhos futuros poderão investigar estratégias alternativas para documentos predominantemente discursivos e avaliar a metodologia em outros atlas linguísticos e conjuntos documentais.

8. Conclusão

Este trabalho apresentou uma metodologia baseada em Engenharia de *Prompt* e Modelos de Linguagem de Grande Escala para apoiar a recuperação de dados dialetais documentados em publicações derivadas do ALiB.

Os resultados indicam que os LLMs podem auxiliar na localização, organização e estruturação de variantes lexicais associadas a conceitos, localidades e referências bibliográficas, contribuindo para a sistematização de informações geolinguísticas dispersas. A principal contribuição da pesquisa consiste na proposição de um procedimento estruturado que integra recuperação automática, organização dos dados e validação humana.

As informações recuperadas pelo modelo foram submetidas à verificação humana e confrontadas com as fontes documentais originais, preservando-se a rastreabilidade entre os registros extraídos e as respectivas referências bibliográficas. Considera-se ainda que o processamento foi realizado por meio de uma plataforma externa de Inteligência Artificial, aspecto relevante para a transparência metodológica do estudo.

Agradecimentos

Os autores agradecem a FAPESB CCE 0022/2023, Projeto ALiB, INCT-TILDIAR, FAPESB BOL0129/2025.

Ética e Uso da Inteligência Artificial

Os LLMs foram utilizados neste estudo como ferramentas auxiliares na recuperação, organização e revisão de informações. Não foram empregados para produzir novos dados dialetais, nem foram submetidos ao modelo registros individuais de informantes. Os dados analisados correspondem a publicações disponibilizadas no site do ALiB, de acordo com a solicitação do *prompt*. Os modelos também foram utilizados para revisão gramatical, aprimoramento textual e organização de ideias.

Referências

Abu-Salih, B., Alotaibi, S., Alanazi, A. L., Abu Khurma, R., Al-Shboul, B., Khouri, A., and Aljaafari, M. (2025). Using large language models for semantic interoperability: A systematic literature review. ICT Express, 11(4):819–837.

Cardoso, S. (2012). ALiB – um atlas nacional com dimensões continentais: problemas, soluções, resultados. In Lobo, T., Carneiro, Z., Soledade, J., Almeida, A., and Ribeiro, S., editors, Rosae: *linguística histórica, história das línguas e outras histórias*, pages 499–508. EDUFBA, Salvador.

Cardoso, S. A. M., Mota, J. A., Ferreira, C., Aragão, M. d. A., and Souza, E. R. d. S. d. (2014). *Atlas Linguístico do Brasil*, volume 1,2. EDUEL, Londrina.

Chun, Y., Kim, M., Kim, D., Park, C., and Lim, H. (2025). Enhancing automatic term extraction with large language models via syntactic retrieval. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics*: ACL 2025, pages 9916–9926, Vienna, Austria. Association for Computational Linguistics.

da Silva, G. A. and Romano, V. P. (2024). O atlas linguístico do brasil (alib) e os atlas de pequeno domínio: comparações e contrastes. *Fórum Linguístico*, 21(1):1–20.

de Oliveira, M. B., Costa, C. S., and da Silva Paz, F. H. (2018). Reinterpretando vazios dialetológicos no norte do brasil. Signum: *Estudos da Linguagem*, 21(1):12–31.

do Socorro Silva Espíndola, M. (2021). Denominações para lombo no português falado no norte e nordeste do brasil: contribuições do alib. TEXTURA - Revista de Educação e Letras, 23(56):7–28.

Freitag, R. and Gois, T. (2024). Performance in a dialectal profiling task of llms for varieties of brazilian portuguese. In Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pages 317–326, Porto Alegre, RS, Brasil. SBC.

Ghali, M.-K., Farrag, A., Won, D., and Jin, Y. (2025). Enhancing knowledge retrieval with in-context learning and semantic search through generative ai. *Knowledge-Based System*s, 311:113047.

Jia, Y. (2023). Research on the development trends of linguistics in the age of artificial intelligence. *Lecture Notes on Language and Literature*, 6(12):1–8.

Melo, L. L. and Pereira, L. C. P. (2024). Um estudo lexical sobre o campo semântico "fauna" nos dados do atlas linguístico do amapá. Trabalho de Conclusão de Curso (Graduação em Letras Português/Inglês) – Departamento de Letras e Artes, Universidade Federal do Amapá, Macapá. Orientador: Romário Duarte Sanches.

Paim, M. M. T. and Ribeiro, S. S. C. (2018). Contribuições do projeto atlas linguístico do brasil para a metodologia de pesquisa geolingüística: o papel do inquiridor. S*ignum: Estudos da Linguagem,* 21(1):106–125.

Pereira Romano, V. (2022). "mandioca brava" in southeast brazil from data of the atlas linguistic project of brazil – alib. Dialectologia, (29):221–243.

Romano, V. (2017). Áreas lexicais no centro-sul do brasil sob uma perspectiva geolinguística. *Revista de Estudos da Linguagem*, 26:103–145.

Sanches, R. D., Moreira, S. T., and Razky, A. (2018). Designações para riacho/córrego na região norte do brasil. EntreLetras, 9(2):466–479.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., and Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques.

Silva, G. A. d. and Romano, V. P. (2026). Tipologia dos atlas linguísticos: proposta de (re)categorização para o contexto brasileiro. ALFA: *Revista de Linguística*, 69.

Singh, A., Fry, A., Perelman, A., Tart, A., and Wang, Z. (2026). Openai gpt-5 system card.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt.

Xu, Z., Ma, H., Ding, Y., Zhang, G., Weng, C., and Peng, Y. (2025). Natural language processing in support of evidence-based medicine: A scoping review. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics*: ACL 2025, pages 21421–21443, Vienna, Austria. Association for Computational Linguistics.

Ressignificação terminológica e conceptualização infantil da experiência oncológica: um estudo com crianças em tratamento para leucemia linfoblástica aguda

Josefa Mônica Almeida Silva Alves[1], Soraya Carvalho Souza Biller Teixeira[2]

Pesquisadora independente[1], Programa de Pós-Graduação em Letras - Universidade Federal de Sergipe[2]

josefamonicaju@gmail.com; sorayabiller@gmail.com[2]

Abstract. *This study investigates how children undergoing treatment for Acute Lymphoblastic Leukemia (ALL) re-signify biomedical terminology related to the disease and chemotherapy. Grounded in Corpus Linguistics (Berber Sardinha, 2004) and Sociocognitive Terminology (Temmerman, 2000), the study analyzes interviews with five children aged 8 to 12. The results reveal processes of meaning reconstruction anchored in lived experience, expressed through generalizations, functional descriptions, metaphors, and bodily references. The study highlights children's forms of conceptualizing biomedical knowledge and contributes to discussions on health literacy and textual and terminological accessibility.*

Resumo. *Este estudo investiga como crianças em tratamento para Leucemia Linfoblástica Aguda (LLA) ressignificam a terminologia biomédica relacionada à doença e à quimioterapia. Fundamentada na Linguística de Corpus (Berber Sardinha, 2004) e na Terminologia Sociocognitiva (Temmerman, 2000), a pesquisa analisa entrevistas com cinco crianças entre 8 e 12 anos. Os resultados evidenciam processos de reconstrução de sentidos ancorados na experiência, expressos por generalizações, descrições funcionais, metáforas e referências corporais. O estudo evidencia formas infantis de conceptualização do conhecimento biomédico e contribui para as discussões sobre letramento em saúde e acessibilidade textual e terminológica.*

1. Introdução

A comunicação em saúde constitui um elemento central nos processos de adoecimento, tratamento e cuidado. Entretanto, em contextos de elevada complexidade terminológica, como a oncologia pediátrica, a circulação do conhecimento biomédico nem sempre resulta em efetiva apropriação das informações pelos pacientes. Conceitos relacionados à doença, aos medicamentos e aos procedimentos terapêuticos são frequentemente apresentados por meio de terminologias especializadas, produzidas em comunidades científicas e profissionais cujos referenciais nem sempre coincidem com aqueles dos sujeitos que vivenciam a experiência do tratamento.

Essa questão torna-se particularmente relevante no contexto da LLA, neoplasia caracterizada pela produção acelerada de linfoblastos imaturos, que comprometem o funcionamento da medula óssea e substituem células saudáveis (Lima et al., 2023). Seu

tratamento é longo, intensivo e altamente protocolarizado, envolvendo diferentes fases, medicamentos, exames e procedimentos, o que coloca a criança em contato recorrente com uma terminologia médica densa e especializada. Nesse contexto, torna-se relevante investigar os modos pelos quais pacientes pediátricos nomeiam, associam e ressignificam conceitos relacionados à doença e ao tratamento.

No campo da Terminologia, a Terminologia Sociocognitiva (Temmerman, 2000) oferece contribuições importantes para a compreensão desse fenômeno ao conceber os conceitos especializados como construções dinâmicas, produzidas na interface entre linguagem, experiência e cognição. Nessa perspectiva, os sujeitos não apenas reproduzem conhecimentos especializados, mas os reinterpretam e reorganizam a partir de suas próprias experiências. Tal compreensão dialoga com abordagens contemporâneas de letramento em saúde, que entendem a construção do conhecimento como um processo social de produção de sentidos, e não apenas como assimilação de informações técnicas.

Partindo desse enquadramento teórico, este estudo investiga como crianças em tratamento para LLA ressignificam a terminologia biomédica relacionada à doença e à quimioterapia. A pesquisa busca responder à seguinte questão: como crianças em tratamento oncológico reinterpretam conceitos especializados presentes no discurso biomédico e o que essas reformulações revelam sobre os processos de conceptualização e letramento em saúde?

Para responder a essa questão, o estudo articula procedimentos da Linguística de Corpus (LC) e da Terminologia Sociocognitiva (TSC). O corpus é constituído pela terminologia especializada extraída do protocolo terapêutico ALL-IC BFM 2009 e por entrevistas realizadas com crianças em tratamento para LLA. A análise examina as reformulações produzidas pelos participantes, identificando processos de generalização, descrições funcionais, metáforas, modelos explicativos e outras estratégias de construção de sentido.

Ao focalizar as formas pelas quais crianças atribuem significado à terminologia biomédica, o estudo busca contribuir para as discussões sobre LC, TSC, letramento em saúde e comunicação centrada no paciente pediátrico.

2. Fundamentação Teórica

2.1 Terminologia Sociocognitiva, acessibilidade terminológica e letramento em saúde

A produção e a apropriação de conhecimentos especializados em contextos de saúde exigem abordagens capazes de considerar não apenas a precisão terminológica, mas também os processos pelos quais os sujeitos atribuem sentido às informações que recebem. Em contraste com a Teoria Geral da Terminologia, orientada pelos princípios da univocidade e da estabilidade conceitual (Wüster, 1998), a Terminologia Sociocognitiva (Temmerman, 2000) propõe compreender os conceitos especializados como construções dinâmicas produzidas na interação entre linguagem, experiência e cognição. Nessa perspectiva, categorias conceituais apresentam fronteiras flexíveis e são continuamente reconstruídas pelos sujeitos em situações concretas de uso.

Ao discutir os limites dos modelos terminológicos tradicionais, Temmerman (2000) argumenta que muitos domínios especializados não são organizados por conceitos

rigidamente delimitados, mas por categorias prototípicas e dinâmicas. Assim, fenômenos como polissemia, sinonímia e metáfora deixam de ser compreendidos como desvios e passam a ser considerados elementos constitutivos dos processos de conceptualização e produção de conhecimento.

Essa compreensão mostra-se especialmente relevante no campo da saúde, onde pacientes frequentemente precisam interpretar conceitos complexos relacionados a diagnósticos, tratamentos e procedimentos. Ao discutir a Acessibilidade Textual e Terminológica (ATT), Finatto e Paraguassu (2022) destacam que informar não é necessariamente comunicar, sobretudo quando a linguagem utilizada não corresponde às condições de compreensão do destinatário. Nesses casos, a informação existe, mas não cumpre plenamente sua função comunicativa, exigindo estratégias de mediação capazes de aproximar o conhecimento especializado das necessidades dos diferentes públicos.

Essa discussão aproxima-se das perspectivas dos Novos Estudos do Letramento. Para Street (2014), o letramento constitui uma prática social situada, atravessada por relações culturais e ideológicas. Em direção semelhante, Kleiman (1995) ressalta que a construção de conhecimentos depende dos contextos sociais em que os sujeitos participam e produzem sentidos. Assim, a produção de sentidos em torno das informações de saúde não se reduz à reprodução de terminologias especializadas, mas envolve interpretar, negociar e integrar conhecimentos à própria experiência.

Tal posicionamento também dialoga com a Linguística Aplicada contemporânea. Conforme argumenta Moita Lopes (2008), compreender fenômenos sociais implica reconhecer a legitimidade de diferentes formas de conhecimento e de vozes historicamente pouco valorizadas. Essa perspectiva converge com a proposta da TSC de compreender os conceitos especializados como construções dinâmicas produzidas na interação entre linguagem, experiência e cognição (Temmerman, 2000). Nesse sentido, as reformulações produzidas pelas crianças não são interpretadas como desvios ou déficits em relação ao discurso biomédico, mas como evidências de processos de conceptualização e produção de sentidos por meio dos quais os participantes atribuem significado à experiência do adoecimento e do tratamento.

2.2 Infância, agência e produção de sentidos

As contribuições da Sociologia da Infância permitiram superar concepções que compreendiam a criança como sujeito incompleto ou mero receptor da cultura adulta. Segundo Sarmento (2005), as crianças devem ser reconhecidas como atores sociais competentes e participantes da construção das relações sociais. Em perspectiva semelhante, Corsaro (2011) argumenta que as crianças não apenas reproduzem conhecimentos produzidos pelos adultos, mas os reinterpretam e transformam por meio de processos de reprodução interpretativa.

Essas contribuições mostram-se particularmente relevantes em contextos de adoecimento crônico. Ao conviverem cotidianamente com procedimentos, medicamentos e terminologias especializadas, as crianças constroem interpretações próprias sobre a doença e o tratamento, produzindo formas singulares de significação da experiência terapêutica. Nessa direção, Alderson (2005) defende que crianças devem ser reconhecidas como participantes legítimas das pesquisas que investigam suas experiências, uma vez que possuem perspectivas próprias e conhecimentos relevantes sobre os contextos em que

vivem.

Com base nessas contribuições, este estudo compreende as narrativas infantis como produções legítimas de sentido. As reformulações terminológicas analisadas não são interpretadas como versões incompletas do discurso biomédico, mas como evidências dos processos pelos quais as crianças atribuem significado à experiência do adoecimento e do tratamento.

3. Procedimentos Metodológicos

Este estudo deriva de uma pesquisa de doutorado desenvolvida no Programa de Pós-Graduação em Letras da Universidade Federal de Sergipe (Alves, 2025), voltada à investigação da terminologia da quimioterapia para LLA sob a perspectiva da TSC. A pesquisa foi aprovada por Comitê de Ética em Pesquisa (Parecer nº 6.505.772), os responsáveis assinaram o Termo de Consentimento Livre e Esclarecido (TCLE), e as crianças participaram mediante assentimento livre e esclarecido obtido por meio de um material multimodal elaborado para a pesquisa (HQ-TALE).

Participaram do estudo cinco crianças do sexo masculino, com idades entre 8 e 12 anos, diagnosticadas com LLA e em tratamento oncológico no Hospital de Urgências de Sergipe (HUSE), acompanhadas também pela AVOSOS, onde foram realizadas as entrevistas. O grupo correspondeu ao quantitativo de crianças encontradas no campo que atendiam aos critérios de elegibilidade estabelecidos: faixa etária entre 8 e 12 anos e até dois anos de cadastro na instituição. Os participantes foram identificados pelos códigos A12, K12, M10, P10 e W8, de modo a preservar suas identidades.

O corpus analisado é constituído por dois conjuntos de dados: (i) a terminologia especializada extraída do protocolo terapêutico ALL-IC BFM 2009 e (ii) entrevistas semiestruturadas realizadas com os participantes. As entrevistas ocorreram em dois momentos complementares. Na primeira etapa, buscou-se explorar livremente as experiências das crianças com a doença e o tratamento. Na segunda, foram utilizados recursos multimodais, incluindo cartas ilustradas do corpo humano, imagens de procedimentos hospitalares e atividades de desenho, com o objetivo de favorecer a evocação de experiências e conceitos relacionados ao tratamento.

A análise foi orientada pelos pressupostos da Terminologia Sociocognitiva (Temmerman, 2000). Inicialmente, realizou-se o levantamento dos termos especializados presentes no protocolo terapêutico. Em seguida, as narrativas infantis foram examinadas qualitativamente, buscando identificar processos de ressignificação terminológica por meio de reformulações, generalizações, descrições funcionais, metáforas e modelos explicativos construídos pelos participantes. A comparação entre a terminologia biomédica e as expressões produzidas pelas crianças permitiu examinar como conceitos especializados são reinterpretados e integrados às suas experiências de adoecimento e tratamento.

Para a constituição do *corpus* terminológico, o protocolo ALL-IC BFM 2009 foi convertido para formato textual e submetido a procedimentos de Linguística de Corpus. No AntConc 4.3.1, foram utilizadas listas de frequência e concordâncias (KWIC) para o levantamento de candidatos a termos e sua observação em contexto. Complementarmente, o TermoStat Web 3.0 foi empregado para identificar candidatos com base em sua especificidade estatística em relação a um *corpus* de referência. Os

resultados das duas ferramentas foram confrontados com a leitura manual do protocolo e submetidos a critérios qualitativos de seleção.

A constituição desse *corpus* permitiu identificar os principais conceitos biomédicos presentes no protocolo e estabelecer parâmetros de comparação com as formas de conceptualização produzidas pelas crianças. Após as etapas de filtragem e validação por especialista, foram selecionadas 90 unidades terminológicas, sendo 34 termos simples e 56 compostos. À luz da Terminologia Sociocognitiva, essas unidades foram categorizadas em entidades, atividades e categorias guarda-chuva, constituindo o repertório especializado de referência para a análise das produções infantis.

4. Resultados e discussão

4.1 Generalização, avaliação experiencial e diferentes formas de apropriação terminológica

Os dados do corpus evidenciam que a apropriação da terminologia biomédica pelas crianças não ocorre por meio da simples reprodução dos termos especializados presentes no discurso médico. Em vez disso, observa-se a mobilização de categorias mais amplas, avaliações subjetivas e explicações construídas a partir da experiência concreta do tratamento. Nesse processo, os participantes reinterpretam conceitos biomédicos segundo critérios perceptivos, afetivos e experienciais, produzindo formas particulares de significação da doença e da terapêutica.

Um exemplo representativo desse movimento pode ser observado na substituição do termo quimioterapia pela expressão remédio forte. Nesse caso, o conceito especializado é incorporado a uma categoria mais ampla (remédio), acompanhada de um qualificativo que destaca a intensidade de seus efeitos. De modo semelhante, a designação da leucemia como doença ruim evidencia a incorporação de avaliações afetivas que sintetizam sofrimento, medo e restrições associadas à experiência do adoecimento.

As entrevistas revelam diferentes formas de apropriação dos conceitos biomédicos. K12 destacou-se por apresentar maior familiaridade com determinados termos e procedimentos, mobilizando explicações relativamente elaboradas sobre a ação da quimioterapia e sobre exames realizados durante o tratamento. Em contraste, A12 demonstra reconhecer a leucemia como parte de sua experiência terapêutica, mas apresenta dificuldades para explicar seu significado. P10, por sua vez, produz explicações próprias para a doença, como a narrativa segundo a qual dela teria "nascido um caroço" que posteriormente "foi crescendo mais outros", revelando um esforço de construção de sentido baseado em elementos concretos e perceptíveis.

O caso de M10 demanda interpretação particular. Durante a entrevista gravada, o participante apresentou respostas breves e mobilizou poucas referências terminológicas. Entretanto, em situações informais de interação e brincadeira, passou a mencionar a leucemia e a compartilhar interpretações próprias sobre o tratamento. Esse contraste sugere que a baixa ocorrência de terminologia especializada não deve ser interpretada automaticamente como ausência de conhecimento, mas como resultado das condições interacionais em que os dados foram produzidos.

Os resultados indicam que a apropriação da terminologia biomédica não pode ser compreendida exclusivamente pela frequência de uso dos termos especializados. Sob a

perspectiva da Terminologia Sociocognitiva (Temmerman, 2000), generalizações, avaliações afetivas, explicações alternativas e silêncios constituem formas legítimas de conceptualização e revelam modos distintos pelos quais as crianças atribuem sentido à experiência do tratamento oncológico.

4.2 Procedimentos, memória corporal e experiência terapêutica

Os procedimentos diagnósticos e terapêuticos constituem um dos domínios em que a distância entre a terminologia biomédica e a conceptualização infantil se mostra mais evidente. Diferentemente de conceitos mais gerais, como doença ou remédio, procedimentos como o mielograma e a administração intratecal raramente são recuperados pelas crianças por meio de suas denominações técnicas. Em seu lugar, emergem referências associadas à dor, à sedação, às posições corporais e às experiências vividas durante o tratamento.

Entre os participantes, apenas K12 mobilizou espontaneamente uma aproximação ao termo especializado ao mencionar o "mielo", fato que parece relacionado à realização recente do exame. Nos demais casos, os procedimentos foram recuperados principalmente a partir das imagens utilizadas durante as entrevistas, que funcionaram como gatilhos de memória. Contudo, as narrativas raramente preservam as distinções técnicas estabelecidas pelo discurso biomédico.

As referências produzidas concentram-se em elementos como "injeção nas costas", "tirar líquido", "agulha", "dormir" e "dor". Em uma das narrativas, P10 descreve a intratecal como uma injeção realizada nas costas, associada à sedação e ao desconforto. Já A12 identifica como intratecal uma imagem originalmente selecionada para representar o mielograma e explica que o procedimento serviria para "tirar líquido das costas". Esses exemplos sugerem que as fronteiras entre diferentes procedimentos invasivos nem sempre aparecem claramente delimitadas na conceptualização infantil.

Tal resultado não deve ser interpretado como simples desconhecimento terminológico. Os dados sugerem que esses eventos são organizados, nas narrativas infantis, a partir de características concretas compartilhadas, como a região do corpo envolvida, o uso de agulhas e as sensações experimentadas durante sua realização. Assim, mielograma e intratecal tendem a integrar uma mesma categoria experiencial construída pela memória corporal do tratamento. Sob a perspectiva da TSC, esse fenômeno evidencia que os conceitos especializados são reorganizados segundo critérios relevantes para a experiência do sujeito, revelando formas próprias de construção de sentido sobre os procedimentos terapêuticos.

4.3 Reinterpretações funcionais e experiência corporal

Além da memória corporal associada aos procedimentos, as entrevistas revelam um segundo movimento recorrente de ressignificação terminológica: a interpretação dos conceitos biomédicos a partir de suas funções percebidas no contexto do tratamento. Nesses casos, os participantes tendem a atribuir significado a medicamentos, substâncias e componentes terapêuticos não por suas definições técnicas, mas pelos efeitos e finalidades que lhes são associados em sua experiência cotidiana.

Um exemplo desse fenômeno refere-se à interpretação do antibiótico como um

"remédio para não pegar doença", deslocando o foco da classificação biomédica para sua função protetiva. Processo semelhante pode ser observado em referências a componentes sanguíneos e alterações hematológicas, reinterpretados por meio de descrições amplas relacionadas ao funcionamento do organismo, como na expressão "sangue fraco".

Esse padrão sugere que a conceptualização infantil privilegia relações de finalidade e consequência. Mais do que identificar nomes ou classificações especializadas, os participantes atribuem sentidos aos medicamentos a partir das funções que lhes associam e dos efeitos que reconhecem em seus corpos. Sob a perspectiva da TSC, essas reformulações constituem estratégias legítimas de construção de sentido, permitindo integrar conhecimentos especializados a esquemas cognitivos já disponíveis no universo experiencial das crianças.

Os dados indicam, portanto, que medicamentos, componentes sanguíneos e procedimentos terapêuticos são frequentemente significados menos por aquilo que são, segundo o discurso biomédico, e mais por aquilo que fazem ou permitem fazer. Esse deslocamento constitui um dos mecanismos mais recorrentes de ressignificação terminológica identificados no corpus.

4.4 Metáforas e modelos explicativos da experiência terapêutica

Entre os fenômenos observados no corpus, as metáforas e os modelos explicativos produzidos pelas crianças constituem evidências significativas dos processos de conceptualização mobilizados para atribuir sentido à doença e ao tratamento. Diferentemente dos casos analisados anteriormente, as narrativas aqui discutidas revelam tentativas mais elaboradas de construir sentido para experiências biomédicas complexas e frequentemente inacessíveis à observação direta.

Um dos exemplos mais expressivos foi identificado na narrativa de K12 ao explicar a ação da quimioterapia. Ao afirmar que o tratamento serve para eliminar as "células ruins" e permitir que permaneçam as "células boas", o participante constrói um modelo explicativo baseado na oposição entre forças antagônicas. Embora simplificada em relação ao discurso biomédico, essa formulação revela uma interpretação funcional da lógica terapêutica e funciona como mecanismo de inteligibilidade para processos celulares invisíveis à experiência cotidiana.

Outro exemplo significativo aparece na descrição dos efeitos da quimioterapia por meio da expressão "o remédio machuca por dentro". A formulação traduz a experiência paradoxal do tratamento oncológico, na qual o mesmo medicamento que provoca dor e desconforto também representa a possibilidade de recuperação. A metáfora permite integrar sofrimento e cuidado em uma mesma estrutura de significado.

As narrativas revelam ainda que a construção de sentido sobre a doença frequentemente ultrapassa os limites da terminologia especializada. Ao afirmar que da doença "nasceu um caroço" que posteriormente "foi crescendo mais outros", P10 elabora uma explicação própria para a origem e o desenvolvimento da leucemia. Ainda que distante das explicações biomédicas, a narrativa evidencia uma tentativa de construir sentido para a experiência do adoecimento por meio de relações causais acessíveis ao universo infantil.

Esses resultados sugerem que as crianças não apenas simplificam a terminologia

biomédica, mas produzem formas próprias de conceptualização do tratamento oncológico. As metáforas e os modelos explicativos identificados no corpus funcionam como recursos de construção de sentido que aproximam o conhecimento especializado da experiência vivida, permitindo que conceitos complexos sejam integrados a estruturas de significado mais acessíveis. Sob essa perspectiva, a apropriação do conhecimento em saúde depende não apenas da aprendizagem de termos técnicos, mas da possibilidade de construir narrativas capazes de tornar inteligíveis experiências frequentemente marcadas pela dor, pela incerteza e pela transformação do cotidiano.

4.5 Ressignificação terminológica, produção de sentidos e letramento em saúde

Os resultados apresentados ao longo desta seção evidenciam que a relação das crianças com a terminologia biomédica ultrapassa a simples aprendizagem de termos especializados. As narrativas analisadas revelam processos contínuos de interpretação e reconstrução de significados, nos quais conceitos relacionados à doença e ao tratamento são incorporados às experiências vividas pelos participantes. Generalizações, descrições funcionais, metáforas e modelos explicativos constituem, nesse contexto, formas legítimas de produção de sentidos sobre a experiência terapêutica.

Essa constatação permite discutir o letramento em saúde a partir dos processos de produção de sentidos observados no corpus. Os dados indicam que a relação das crianças com o conhecimento biomédico não se reduz à reprodução de terminologias especializadas ou à assimilação integral das explicações oferecidas pelos profissionais de saúde. As crianças constroem interpretações próprias sobre a leucemia, a quimioterapia e os procedimentos terapêuticos, articulando informações recebidas, experiências corporais, emoções e interações sociais. Sob essa perspectiva, o letramento em saúde pode ser concebido como um processo social de construção de sentidos, produzido na interface entre conhecimento especializado e experiência vivida.

Tais resultados possuem implicações importantes para a comunicação em saúde e para a acessibilidade terminológica. Em vez de interpretar as reformulações produzidas pelas crianças como evidências de déficit ou incompreensão, torna-se produtivo reconhecê-las como pontos de partida para o diálogo entre saberes especializados e experiências cotidianas. As formas de conceptualização identificadas neste estudo revelam modos pelos quais o conhecimento biomédico se torna significativo para os sujeitos e reforçam a necessidade de práticas comunicativas que reconheçam os pacientes pediátricos como participantes ativos na construção de sentidos sobre sua própria experiência terapêutica.

5. Considerações finais

Este estudo investigou como crianças em tratamento para LLA ressignificam a terminologia biomédica relacionada à doença e à quimioterapia, examinando o que essas reformulações revelam sobre os processos de conceptualização e letramento em saúde.

A análise das entrevistas evidenciou que a apropriação dos conceitos especializados não ocorre, predominantemente, por meio da reprodução da terminologia médica, mas por processos de reconstrução de sentidos ancorados na experiência vivida. Generalizações, descrições funcionais, metáforas, modelos explicativos próprios e

referências à experiência corporal constituíram estratégias recorrentes de produção de sentidos sobre a doença e o tratamento. Os resultados reforçam os pressupostos da TSC ao evidenciarem que os conceitos especializados são continuamente reinterpretados pelos sujeitos em contextos concretos de uso. Em vez de serem interpretadas como evidências de desconhecimento ou déficit informacional, as reformulações observadas nas narrativas infantis revelam processos legítimos de conceptualização e produção de sentidos sobre experiências marcadas pela dor, pela incerteza e pela complexidade do tratamento oncológico. As análises também evidenciam que diferentes participantes apresentam distintas formas de apropriação terminológica, sugerindo que fatores experienciais, emocionais e interacionais influenciam a construção de significados sobre a doença e a terapêutica.

Do ponto de vista do letramento em saúde, os achados indicam que a relação dos participantes com o conhecimento especializado não se reduz à reprodução da terminologia biomédica, mas envolve processos de construção de sentidos que integram esse conhecimento à experiência cotidiana dos sujeitos. Nesse contexto, a acessibilidade textual e terminológica ultrapassa a simples simplificação vocabular, exigindo o reconhecimento das formas pelas quais os pacientes elaboram explicações próprias sobre sua condição de saúde.

Considerando o caráter qualitativo e situado da investigação e o número reduzido de participantes, os resultados não pretendem generalização estatística, mas oferecem indícios sobre processos de ressignificação terminológica que podem ser explorados em estudos com grupos mais amplos. Ao dar visibilidade às narrativas de crianças em tratamento oncológico, este estudo contribui para as discussões sobre Linguística de Corpus, Terminologia Sociocognitiva, letramento em saúde e comunicação centrada no paciente, apontando para a necessidade de práticas comunicativas mais sensíveis aos modos infantis de significar a experiência terapêutica.

Questões éticas

A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Universidade Federal de Sergipe (Parecer nº 6.505.772), com assinatura do Termo de Consentimento Livre e Esclarecido pelos responsáveis e assentimento das crianças participantes. Ferramentas de Inteligência Artificial generativa foram utilizadas exclusivamente como apoio à revisão linguística e editorial do manuscrito, incluindo sugestões de reformulação, concisão e adequação textual, sob avaliação e decisão das autoras. A IA não foi utilizada na produção ou coleta dos dados, nem substituiu a análise e a interpretação dos resultados pelas pesquisadoras. Josefa Mônica Almeida Silva Alves foi responsável pela concepção da pesquisa, coleta e análise dos dados, interpretação dos resultados e redação do manuscrito. Soraya Biller Teixeira contribuiu para os procedimentos metodológicos relacionados à Linguística de Corpus, especialmente na preparação do corpus e na utilização das ferramentas AntConc e TermoStat.

Referências

Alderson, P. (2005) As crianças como pesquisadoras: os efeitos dos direitos de participação sobre a metodologia de pesquisa, *Educação & Sociedade*, 26(91), pp. 419-442.

Alves, J. M. A. S. (2025) *Terminologia da quimioterapia para Leucemia Linfoblástica Aguda: uma análise sociocognitiva com foco na experiência de pacientes pediátricos*, Tese (Doutorado em Letras) – Universidade Federal de Sergipe, São Cristóvão.

Berber Sardinha, T. (2004) *Linguística de Corpus*. Manole, Barueri.

Corsaro, W. A. (2011) *Sociologia da infância*, 2. ed., Artmed, Porto Alegre.

Finatto, M. J. B. and Paraguassu, L. B. (2022) *Acessibilidade Textual e Terminológica*, EDUFU, Uberlândia.

Kleiman, A. B. (1995) *Os significados do letramento: uma nova perspectiva sobre a prática social da escrita*, Mercado de Letras, Campinas.

Lima, M. S. et al. (2023) Correlação entre fadiga e qualidade de vida em pacientes oncológicos pediátricos, *Research, Society and Development*, 12(1), p. e17312126892. Available at: https://rsdjournal.org/index.php/rsd/article/view/26892 [Accessed: 7 Jan. 2025].

Moita Lopes, L. P. da (org.) (2008) *Por uma linguística aplicada indisciplinar*, 2. ed., Parábola Editorial, São Paulo.

Sarmento, M. J. (2005) Gerações e alteridade: interrogações a partir da sociologia da infância, *Educação & Sociedade*, 26(91), pp. 361-378. DOI: 10.1590/S0101 73302005000200003.

Street, B. V. (2014) *Letramentos sociais: abordagens críticas do letramento no desenvolvimento, na etnografia e na educação*, Traduzido por M. Bagno, Parábola Editorial, São Paulo.

Temmerman, R. (2000) *Towards New Ways of Terminology Description*: the Sociocognitive Approach", John Benjamins, Amsterdam/Philadelphia.

Wüster, E. (1998) *Introducción a la teoría general de la terminología y a la lexicografía terminológica*, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

Acessibilidade textual e terminológica em ambiente digital: análise de *corpus* e impacto social na educação em saúde

Bruna Rodrigues da Silva

Universidade Federal do Rio Grande do Sul - UFRGS

thuduh@gmail.com

Resumo: *No momento digital atual, a disseminação de informações técnico-científicas nesses formatos (e-books, portais, aplicativos) tornou-se quase que obrigatória, quando se fala em educação e saúde. Entretanto, o simples fato do material ser/estar digitalizado não garante a democratização do conhecimento se não houver Acessibilidade Textual e Terminológica. Esta pesquisa investiga o impacto social da compreensão (ou não) de materiais digitais de saúde por jovens de escola pública, utilizando a Linguística de Corpus como método para diagnosticar e propor soluções para barreiras linguísticas encontradas. O corpus de estudo é o livro digital "Aprendendo sobre vírus e vacinas", testado com estudantes da rede municipal de Porto Alegre-RS.*

Abstract: *In the digital era, disseminating technical and scientific information through formats such as e-books, portals, and apps has become virtually mandatory in education and healthcare. However, digitalization alone does not guarantee the democratization of knowledge without Textual and Terminological Accessibility. This study investigates the social impact of whether public school students comprehend digital health materials, using Corpus Linguistics to diagnose and address linguistic barriers. The study corpus is the digital book "Aprendendo sobre vírus e vacinas" (Learning about viruses and vaccines), tested with municipal school students in Porto Alegre, Brazil.*

1. Introdução

Este estudo integra tese de doutorado inserida nos estudos de Acessibilidade Textual e Terminológica, que investiga como materiais educativos da área da saúde, adaptados para jovens, são efetivamente compreendidos pelo público. O *corpus* de estudo é o livro digital "Aprendendo sobre vírus e vacinas", publicado em 2020 pela Editora da UFCSPA[9], como um material simplificado para o público de 12 anos ou mais. O objetivo principal consiste em descrever e analisar como esse material, gratuitamente disponível, poderia ser compreendido por jovens do Ensino Fundamental II, matriculados em uma escola pública de Porto Alegre-RS. Para estimar prováveis pontos de complexidade e de acessibilidade, seu léxico foi caracterizado a partir do contraste com outros *corpora*: a) um *corpus* de redações escolares de escolas públicas do Ensino Fundamental do RS; b) o Corpop[10] – *corpus* do português popular escrito culto (Pasqualini, 2018) e, c) Somos Heróis – os

[9] UFCSPA: Universidade Federal de Ciências da Saúde de Porto Alegre.
[10] Para acessar o CorPop: < https://www.ufrgs.br/textecc/porlexbras/corpop/> Acesso em: 10 fev. 2026.

cuidados para o coronavírus ir embora, publicação digital infanto-juvenil (Leite, 2018). A partir da caracterização e do contraste com esses outros *corpora*, foram elaboradas questões para testes de compreensão leitora a serem realizadas com estudantes em sala de aula. 50 estudantes de último ano do Ensino Fundamental de escola pública municipal de Porto Alegre - RS responderam os testes impressos, com a professora de Português, enquanto acessavam o livro digital em *chromebooks*[11] da escola. Os resultados mostraram que os alunos, de modo geral, consideraram o texto fácil e claro. Entretanto, seus escores de acertos e erros nas questões específicas sobre o conteúdo revelaram uma série de dificuldades de compreensão. Por fim, foram validadas as hipóteses de que esse público tende a não compreender por completo o conteúdo do livro, embora o material se apresente como adaptado para público escolar infanto-juvenil da idade deles.

É importante ressaltar que compreendemos a noção de *corpus* como conjuntos de textos autênticos extensos, conforme os princípios da Linguística de Corpus. Porém, nosso trabalho utiliza textos de curta extensão, pois foca na caracterização desses textos, com objetivo de acessibilidade e acesso do conteúdo científico às camadas mais simples da população. Entre os *corpora* denominados *corpora* de contraste nesta pesquisa também estão *corpora* de tamanhos pequenos, como *corpus* de redações escolares, formado por 154 redações de alunos de escola pública. Tal *corpus* foi reunido em dissertação de mestrado (Silva, 2021) focando na descrição do vocabulário de uso real dessa população, com os dados disponíveis na época, cuidadosamente digitalizados e manualmente conferidos para que não houvesse modificação do original escrito à mão pelos estudantes, por isso não é um *corpus* extenso. E o outro *corpus* de contraste Somos Heróis, também de pequena extensão, foi utilizado como contraste por ser o mais próximo encontrado do objeto principal desse estudo.

Ainda que de curta extensão, os *corpora* utilizados nessa pesquisa se justificam pela minuciosidade da digitalização, limpeza, formatação e organização dos dados, que geraram análises, descrições e resultados importantes e que podem contribuir com novos estudos e pesquisas relacionadas.

Neste estudo, a Linguística de Corpus provou ser um método que identifica lacunas de acessibilidade que ferramentas puramente automáticas podem ignorar. Assim, propõe-se que novos materiais educacionais incorporem diretrizes de Leitura Fácil e sejam validados por métodos experimentais de Linguística de Corpus antes de sua distribuição.

1. 1 *Corpus* de estudo

O livro A*prendendo sobre vírus e vacinas* (Rodrigues, 2020), foi criado em um projeto de extensão de mesmo nome da UFCSPA, e objetiva auxiliar o aprendizado dos alunos dos anos finais do Ensino Fundamental no entendimento sobre vírus e vacinas utilizadas na sua prevenção (Rodrigues, 2020, p. 65). Desde o início do projeto de doutoramento, o único livro da coleção de obras simplificadas da Editora da UFCSPA direcionado para a idade dos alunos da professora-pesquisadora (12+) é este, por isso foi escolhido para

[11] *Chromebooks* são computadores portáteis pequenos, menores que os *notebooks*, que não possuem memória, portanto, não baixam nem salvam arquivos. Utilizam um sistema de armazenamento na nuvem, por isso, oferecem melhor desempenho quando conectados à internet.

análise e testagem. Este livro digital tem 72 páginas, está disponível para *download* gratuito e foi publicado pela Editora da UFCSPA no ano de 2020.

2. Referencial Teórico

A pesquisa fundamenta-se em eixos principais que conectam a linguagem ao impacto social: a Linguística de Corpus, a Terminologia, a Acessibilidade Textual e Terminológica, a leitura e a compreensão e as questões éticas do estudo.

O trabalho está fundamentado nas teorias e pressupostos da Linguística de Corpus (LC), que explora a linguagem de forma empírica, por meio de dados reais, que se realizam na língua, via computador (Berber Sardinha, 2004). Foram utilizadas as frequências e concordâncias para identificar padrões de uso, relação entre os *corpora* e seus vocabulários e para identificar padrões de simplicidade/complexidade textuais (Anthony, 2019). Nesse sentido, a análise foi complementada pelo Processamento da Linguagem Natural (PLN), que não deixa de ser um ramo da Inteligência Artificial (IA) que lida com o processamento automático das línguas. O PLN (Caseli, Nunes, 2024) ajudou a estimar a complexidade textual do *corpus* de estudo e as ferramentas computacionais utilizadas ajudaram a relacionar os *corpora*.

A pesquisa também utiliza a Teoria Comunicativa da Terminologia (TCT) para sustentar a ideia de que a informação especializada precisa ser focada no usuário, evitando o distanciamento entre a ciência e a população (Cabré, 2005). A TCT defende que o uso e o contexto determinam se uma palavra assume caráter de termo, dependendo do seu comportamento real em contextos comunicativos. A Terminologia Experimental (Finatto, Zilio, 2005) complementa a TCT como uma prática necessária para validar empiricamente se os achados de pesquisa cumprem sua função social, utilizando leitores reais para validar as questões de pesquisa.

Acessibilidade Textual e Terminológica (ATT) foca na "ponte" entre o texto e o leitor, buscando potencializar a compreensão, agindo sobre o léxico, a sintaxe e a arquitetura do texto (Finatto, 2019). Relaciona-se com a Linguagem Simples (*Plain Language*) (Fischer, 2017), que reconhece o direito do cidadão de entender informações relevantes para seu cotidiano, e com a Leitura Fácil, que é mais focada nos fins didáticos, incluindo o uso funcional de imagens.

A leitura é vista como uma atividade complexa, que envolve processos visuais, fonológicos e semânticos (Kleiman, 2005). Diferencia-se a alfabetização (domínio do código) do letramento (uso social da língua) (Jouve, 2002). O foco é a compreensão leitora, que envolve a habilidade de extrair significado e interpretar o texto, indo além da simples decodificação (Fulgêncio, Liberato, 1992).

Além disso, com relação à ética dos dados, como a pesquisa realizou testes de compreensão leitora com seres humanos, foi necessária a apreciação e validação pelo Comitê de Ética e Pesquisa da Universidade, para que fosse aprovada e pudesse ser realizada com a devida autorização. Os dados colhidos foram todos anonimizados, para que não influenciasse nos resultados e respeitasse os direitos dos participantes.

3. Materiais e métodos

A metodologia seguiu uma abordagem quanti-qualitativa apoiada em ferramentas computacionais. O primeiro passo foi a organização do *corpus:* o material digital (PDF) foi convertido para texto simples (TXT) e revisado manualmente para remover interferências de conversão e de diagramação (por exemplo: boxes, cabeçalhos, números de página, etc.). O *corpus* de estudo resultou em 3.516 palavras (*tokens*) e 920 palavras diferentes (*types*), com uma variação vocabular de 26%.

Após esse primeiro passo, foram realizados testes preliminares com auxílio de ferramentas de simplificação. A ferramenta utilizada foi a MedSimples[12], que gera sugestões de reescrita para textos de saúde, identificando termos técnicos e palavras complexas. Essa ferramenta é um sistema baseado em PLN/IA.

Também foi utilizado o *software* AntConc[13] para contrastar o livro com três *corpora* de referência. O primeiro foi um *corpus* de redações escolares (produção real dos jovens), 154 textos de alunos de escolas públicas do RS, para verificar o vocabulário ativo dos jovens. Esse primeiro contraste serviu para representar o vocabulário escrito de estudantes de Ensino Fundamental de duas escolas públicas do Rio Grande do Sul. Esse contraste foi importante para tentar mostrar repertórios de palavras do livro que, porventura, os alunos desse público não costumassem escrever em suas redações. Tratando-se de livro da área da saúde, sobre vírus e vacinas, era esperada a presença de terminologias, diferentemente do *corpus* de redações, pois os assuntos diversos dos textos não incluíam saúde ou outras temáticas propensas ao uso de terminologia.

O segundo contraste, com o CorPop, um *corpus* de referência do português popular brasileiro escrito culto. Esse *corpus* foi organizado por Pasqualini (2018) e reúne textos de literatura simplificada e de jornais populares.

E o terceiro contraste foi realizado com um outro *corpus*, mais próximo tematicamente do texto do nosso livro em estudo: a publicação digital *Somos Heróis – os cuidados para o coronavírus ir embora*, de Pedro Leite (2020). Essa publicação foi selecionada porque apresenta vários pontos em comum com a obra de estudo: é texto adaptado; busca atender leitores a mesma faixa etária; tem acesso livre, disponibilidade para *download,* gratuidade; publicado somente em formato digital; trata da temática da COVID-19.

A partir da caracterização do *corpus* de estudo, dos testes preliminares e dos contrastes com os diferentes *corpora*, foram elaborados testes de compreensão leitora com 10 questões de múltipla escolha e 5 perguntas abertas. A população foi composta por 50 estudantes de 9º ano e turmas de Correção de Fluxo (CF) de uma escola municipal da zona norte de Porto Alegre, com idades entre 13 e 16 anos. Os testes foram aplicados em sala de aula, com os alunos acessando o livro via *chromebooks* enquanto respondiam ao questionário impresso. Dessa forma, a coleta de dados aconteceu em contexto digital, refletindo o cotidiano da inclusão digital nas escolas.

[12] MedSimples: ferramenta automática de auxílio à simplificação de textos sobre temas de Saúde (Apoio: Prêmio LARA 2019 – Empresa Google – *Latin America Research Awards* 2019). Disponível em: http://www.ufrgs.br/textecc/acessibilidade/page/cartilha/. Acesso em: 25 jun. 2026.

[13] Disponível gratuitamente em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 25 jun. 2026.

4. Resultados

A análise do *corpus* de estudo mostrou que, embora o livro seja apresentado como simplificado, contém termos técnicos e palavras difíceis, que não aparecem no vocabulário escrito pelos estudantes. Nos testes iniciais com auxílio da ferramenta MedSimples, foram apontados muitos termos e palavras que seriam de difícil compreensão, considerando o público. Como o livro é um material científico da área da saúde, muitas vezes não é possível eliminar os termos dessa escrita. Porém, é possível explicá-los, inclusive com exemplos. Conforme os preceitos da Linguagem Simples, essa é a maneira que temos para lidar com a terminologia em textos que pretendem ser simples.

Nos estudos pilotos, foram utilizados três diferentes *corpora* como referências de um repertório vocabular e lexical de linguagem potencialmente mais acessível para um público jovem e adolescente. No primeiro estudo pilotado realizado, a ideia era constatar possíveis diferenças de vocabulário escrito entre o *corpus* de estudo e o *corpus* de redações escolares. Entre as palavras coincidentes encontramos, como é natural, os elementos que perfazem a gramática da língua: artigos, preposições e conjunções (como, por exemplo: de, com, para, a, e, o etc.). Porém, entre as diferenças de vocabulário, há muitas terminologias e palavras vindas das experiências de todos com as vivências da pandemia – assintomático, imunológico, fronteiras, protocolo. Mas, além disso, também foi possível observar a presença de verbos (possuir, focar, fazer), de desvios da normal culta (achão) e de usos da linguagem da internet − seja abreviação, como tá e oq, ou uso de palavras estrangeiras, como free - como um diferencial nessa comparação. Esse primeiro contraste mostrou palavras que se tivessem sido utilizadas no livro, talvez fossem compreendidas mais facilmente pelo público-alvo.

Neste segundo estudo, foi realizado contraste entre o *corpus* de estudo e o repertório de palavras do CorPop (Pasqualini, 2018). Focando nas diferenças de vocabulário entre os *corpora*, são exemplos: anual, audição, protocolos, plataforma. No primeiro exemplo, a palavra "anual" não aparece no CorPop, mas "ano" sim. Então, quem sabe se a frase fosse reformulada para "Cuidar a campanha contra a gripe que acontece todos os anos!", isso a tornaria mais compreensível para o público pretendido? A palavra "audição" também não aparece presente no vocabulário do CorPop, já os itens "ouvido" e "ouvir", sim. Dessa forma, também para um insumo em testagens, teríamos esta opção: "esclarecimentos sobre como está a saúde dos olhos e dos ouvidos". Assim, o contraste entre o texto do livro e o acervo do CorPoP serviu para apontar pistas sobre como seria, em tese, uma formulação mais simples.

O terceiro *corpus* de contraste foi a publicação digital *Somos Heróis – os cuidados para o coronavírus ir embora*, de Pedro Leite (2021), que foi selecionado porque tem vários pontos em comum com o *corpus* de estudo. A principal questão, acreditamos, é: por que determinadas palavras são usadas num material que pretende ser simples, para jovens e adolescentes? Conforme Pires *et. al* (2023):

> Conhecer o público-alvo de seu texto é uma das características mais importantes de um escritor ou de um adaptador de texto para Leitura Fácil, afinal de contas, você precisará tomar muitas decisões com base no conhecimento prévio dos leitores, no objetivo do texto, nas habilidades de leitura de seus leitores, em seus potenciais e em suas dificuldades. (Pires *et al.* 2023, p.11)

Na tentativa de alguma sugestão que pudesse tornar o texto mais acessível, fomos atrás dos contextos de algumas dessas palavras. A ferramenta do AntConc que auxiliou nessa busca foi o *Concordance*. Os exemplos, tanto anteriores quanto seguintes, mostram que não há preocupação, por parte dos autores, com o público que vai ler o texto, contrariando os princípios tanto da Linguagem Simples quanto da Leitura Fácil. Percebemos que há uso de linguagem técnica que os profissionais da saúde dominam. Mas, e os alunos de Ensino Fundamental, como o livro salienta, será que compreendem a utilização das palavras "ancorada", "permanente", "produzida" e "proliferam"? Quem sabe uma "proteína fixa" ou "apoiada" fosse mais próxima da realidade desse público, assim como um "vírus eterno", "durável" ou "contínuo". A vacina poderia ser "feita" ou "fabricada" pelo Instituto Butantan. E os linfócitos poderiam "aumentar", "crescer" ou "se multiplicar".

Sabemos que evitar a terminologia por completo é algo difícil quando se trata de material da área da saúde. Então, na tentativa de auxiliar no entendimento desses termos que aparecem bastante ao longo do texto, os autores criaram, ao final, um glossário. Porém, ao analisarmos o glossário, percebemos que muitos termos são explicados exatamente com a mesma frase que aparece dentro do texto do livro, e outros são explicados com ainda mais terminologia. Ou seja, acabam por não explicar e, ainda, fazem o leitor ir até o final para depois retornar páginas para continuar sua leitura. Nesse sentido, nesse material, o glossário, aparentemente, não cumpre seu papel, sendo mais um elemento a se considerar nas testagens.

Os testes de compreensão leitora foram, então, construídos baseados nos achados dos testes preliminares, estudos pilotos e caracterização do *corpus* de estudo. Buscamos, ao longo das 10 questões, tratar do que mais chamou atenção: o uso de palavras potencialmente difíceis, uso de termos, glossário ao final do livro, falta de explicações claras e acessíveis, relação das imagens com as informações e também retomada pronominal. Os alunos tiveram desempenho mediano nas questões de marcar. Notamos que alguns relacionaram informações do livro e questões dos testes com seu conhecimento de mundo. E outros contribuíram com as suas opiniões nas questões abertas. Ainda assim, acreditamos que os testes comprovaram que há, sim, dificuldade desse público em compreender as informações do livro de forma completa. Houve questões em que confundiram as palavras difíceis e os termos; houve questões em que não relacionaram a imagem com as informações; houve questões com muitas respostas diferentes entre o grupo de 50 alunos pesquisado.

Nas questões abertas, a maioria deles considerou a leitura fácil (13 alunos), porém, não acertou as questões de múltipla escolha. Apenas 01 aluno considerou o texto "mais ou menos" e 06 alunos consideraram a leitura difícil. Eles relataram o uso das palavras desconhecidas como o principal fator que dificulta a leitura e o entendimento do livro, confirmando uma hipótese da pesquisa.

Além disso, sugeriram um dicionário maior e, também, no início do livro, com explicações mais fáceis. Todas as repostas que citaram o glossário vieram de encontro à quarta hipótese deste estudo, que justamente questiona o glossário, apostando que, como foi construído e onde está no livro, não auxilia na leitura. As opiniões dos estudantes confirmaram essa hipótese e o nosso olhar enquanto professora desse público e estudiosa dos princípios da ATT.

5. Considerações finais e impactos sociais

A pesquisa demonstra que a transformação digital na educação em saúde exige mais do que somente o acesso à tecnologia, pois ela necessita de reflexão e simplificação textual. A Linguística de Corpus se mostrou um método eficiente para identificar lacunas de acessibilidade que ferramentas puramente automáticas podem ignorar. O estudo baseou-se na possibilidade de impacto social, no qual o direito de entender é a base da cidadania. Visando esse objetivo, idealiza-se que novos materiais digitais, sobretudo de educação em saúde, incorporem diretrizes de Leitura Fácil e sejam validados por métodos da LC antes de sua distribuição em larga escala.

Confirmamos, por meio das testagens e da caracterização do *corpus*, a hipótese principal: o livro analisado não está plenamente acessível para estudantes de escolas públicas periféricas na faixa de 13 a 16 anos. O uso de linguagem técnica sem maiores explicações, o vocabulário complexo, as explicações circulares do glossário e as imagens puramente estéticas dificultam a compreensão do leitor.

Sabemos que realizar pesquisas com testagens carregam limitações e dificuldades. Ainda assim, acreditamos que a pesquisa contribui em muitos pontos: com estudos sobre o público com essa idade, com estudos sobre ATT em materiais para essa faixa etária, com estudos sobre ATT e com estudos sobre testagens, sobretudo testagens para público menor de idade e de camadas mais periféricas da população. Assim, nossa discussão busca contribuir com as pesquisas sobre leitura e compreensão de estudantes que são a grande maioria da população brasileira: os estudantes de escola pública. Representamos essa faixa da população que carece de mais discussões, estudos, pesquisas, para que haja investimento na educação pública, bem como nas aprendizagens de leitura e compreensão, sem as quais não há como ler ou compreender quaisquer outros assuntos, sem as quais não há como interpretar o mundo.

Como professora imersa nessa realidade e regente de turmas de anos finais do Ensino Fundamental há 14 anos letivos, penso que a pesquisa, sobretudo a pesquisa feita na universidade pública e a pesquisa associada a cursos de licenciatura, precisa contribuir para uma educação pública de qualidade, que busque a igualdade de uma sociedade hoje tão desigual.

Questões éticas

No que diz respeito às testagens de leitura realizadas pelos alunos da escola pública, este estudo foi apreciado pelo Comite de Ética e Pesquisa da UFRGS e foi aprovado conforme o parecer número 7.008.035 e CAAE: 81455224.6.0000.5347. Além disso, antes da realização dos testes, os pais e responsáveis pelos estudantes autorizaram as suas participações no estudo, por meio de assinatura no TCLE (Termo de Consentimento Livre e Esclarecido). Os estudantes assentiram suas participações, assinando o TALE (Termo de Assentimento Livre e Esclarecido). A inteligência artificial se fez presente neste estudo, na medida em que foi utilizado o PLN para estimar a complexidade/simplicidade do *corpus* de estudo e as ferramentas computacionais foram utilizadas para contrastar os *corpora*. O estudo foi todo realizado por uma única autora.

Referências

Abreu, K.; Hora, K.; Pinheiro, A. C. Compreensão leitora, consciência sintática e metacognição sob a abordagem da psicolinguística educacional: um estudo com o 7º ano do ensino fundamental. *Revista Pró Língua*, v. 15, n. 2, p.35-54, 2020.

Anthony, L. *AntConc* (Versão 3.5.8) [*Software*]. Tóquio: Waseda University, 2019.

Berber Sardinha, T. *Lingüística de Corpus.* Barueri: Manole, 2004.

Cabré, M. T. La terminologia, em disciplina em evolución. *Debate Terminológico*, Porto Alegre, 1:1-14, 2005.

Caseli, H. M.; Nunes, M. G. V. (org.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português.* 3 ed. BPLN, 2024.

Finatto, M. J. B.; Motta, E. Terminologia e Acessibilidade: novas demandas e frentes de pesquisa. *REVISTA GTLEX*, v. 2, p. 316-356, 2019.

Finatto, M. J. B.; Zilio, L. *Textos e Termos por Lothar Hoffmann*. Porto Alegre: Palotti, 2015.

Fischer, H. M. P. *Clareza em textos de e-gov, uma questão de cidadania*. Monografia (Especialização). Rio de Janeiro: PUC-Rio, 2017.

Fulgêncio, L.; Liberato, Y. *Como Facilitar a Leitura*. São Paulo: Contexto, 1992.

Jouve, V. *A leitura*. São Paulo: Editora UNESP, 2002.

Kleiman, A. *Oficina de leitura: teoria e prática.* São Paulo: Pontes, 2005.

Leite, P. *Sofia e Otto: somos heróis* – os cuidados para o coronavírus ir embora. Porto Alegre: PGL, 2021.

Marengo, S. M. D. A. *et al.* Caminhos para uma Terminologia Experimental. *ORGANON*, v. 40, p. 01-42, 2025.

Pasqualini, B. F. *CorPop*: um corpus de referência do português popular escrito do Brasil. Tese (Doutorado). Porto Alegre: UFRGS, 2018.

Pires, V. O. D. *et al. Manual de Leitura Fácil para educadores.* Pelotas: Editora IFSUL, 2023.

Rodrigues, L. C. J. *Aprendendo sobre vírus e vacinas* [recurso eletrônico]. Porto Alegre: Editora da UFCSPA, 2020.

Silva, B. R. *Vocabulário escrito de estudantes de escolas públicas do Rio Grande do Sul.* Dissertação (Mestrado). Porto Alegre: UFRGS, 2021.

Bulas de medicamentos para o novo Corpop Saúde: curadoria e avaliação linguística de simplificações geradas por modelos de inteligência artificial

Maria José Bocorny Finatto[1], Arthur Scalercio[2], Bianca Franco Pasqualini[1]

[1] Programa de Pós-Graduação em Letras – Universidade Federal do Rio Grande do Sul – Porto Alegre – RS – Brasil, [2] Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brasil

mariafinatto@gmail.com, arthurscalercio@id.uff.br, biancapasqualini@gmail.com

Abstract. *This paper introduces the new CorPoP Saúde corpus, designed to support materials written in plain, accessible language on health topics for Brazilian citizens with limited literacy and education. In addition to the overall structure of the new corpus, we describe three subcorpora comprising the original patient leaflet versions and their simplifications generated by Artificial Intelligence (AI) models and revised by experts, following the protocol proposed by Scalercio et al. [2026]. The paper highlights data from 30 inserts covering 10 hypertension medications included in the 2025 Farmácia Popular national program.*

Resumo. *Apresenta-se o novo corpus CorPoP Saúde, concebido para apoiar a redação de materiais em linguagem simples e acessível sobre temas de saúde para cidadãos brasileiros com letramento e escolaridade limitados. Além da estrutura geral do novo* corpus*, descrevem-se os três* subcorpora *que abrangem versões originais de bulas para o paciente e simplificações geradas por modelos de Inteligência Artificial (IA) e revisadas por especialistas, seguindo o protocolo de Scalercio* et al. *[2026]. São destacados os dados de 30 bulas de 10 medicamentos para hipertensão do programa nacional Farmácia Popular de 2025.*

1. Introdução

No Brasil, a educação em saúde é um serviço clínico que inclui "diferentes estratégias educativas, as quais integram os saberes popular e científico, de modo a contribuir para aumentar conhecimentos, desenvolver habilidades e atitudes sobre os problemas de saúde e seus tratamentos" [Conselho Federal de Farmácia, 2016]. Essa educação, entretanto, necessita do apoio de servidores especializados e de materiais informativos e didáticos que sejam de fácil entendimento e uso, com linguagem simples e acessível e manipulação simples por parte dos usuários dos serviços de Saúde. Por isso, o novo *corpus* CorPoP Saúde é uma iniciativa de pesquisa no tema da promoção da acessibilidade textual e do oferecimento de insumos linguisticamente tratados para apoiar a apresentação de conteúdos escritos sobre temas de saúde em linguagem simples e fácil de entender. Esse

corpus visa atualizar e expandir um *corpus* original, o CorPoP[14], que já tem servido de referência desde 2018 para indicar possíveis padrões mais genéricos de um português popular escrito, mais fácil de entender para pessoas adultas com letramento e escolaridade limitados.

O novo CorPop Saúde[15] também foi planejado para subsidiar a crítica de materiais informativos escritos e multiformato, em tese, acessíveis, sobre diferentes temas de Saúde, aspecto não contemplado no *corpus* original. Disponível gratuitamente, com funcionalidades de IA acopladas à consulta e a buscas do usuário, visa ajudar a ampliar o Letramento e a Educação em Saúde (LS) no Brasil. Isso será feito via incorporação e tratamento de textos divulgativos especialmente selecionados sobre diferentes temas médicos de maior interesse popular, como guias e cartilhas institucionais sobre cuidados e prevenção de doenças, bulas de medicamentos, termos de consentimento (TCLEs) que preenchemos para realização de exames médicos rotineiros (como os compilados e estudados por Ponomarenko [2022], e Ponomarenko e Rebechi [2023]), além de materiais educativos diversos produzidos por órgão públicos e instituições de Saúde do Brasil.

Com ferramentas e técnicas de Linguística Computacional, de Processamento de Linguagem Natural, de Linguística de *Corpus* e Plataformas de Inteligência Artificial (PIA), o novo CorPoP Saúde já tem apoios da Fundação de Amparo à Pesquisa do RS e do Instituto Nacional de Ciência e Tecnologia em IA Responsável, o INCT-CNPq TILD-IAR[16], visando contribuir para: **a)** ampliar a oferta pública de *corpora* temáticos devidamente curados para o português do Brasil; **b)** fomentar o trabalho de curadoria linguística desses *corpora* frente a recursos deles gerados somente por sistemas de PIAs; **c)** subsidiar ações para ampliar o letramento em saúde que demandem dados simplificados confiáveis gerados com curadoria humana especializada; **d)** apoiar decisões de simplificação textual válidas para o cenário da comunicação em Saúde do Brasil.

O novo *corpus* também visa subsidiar recursos de interface humano-computador para promover a autonomia tanto dos agentes quanto dos usuários dos serviços do sistema SUS, ajudando a ampliar a leiturabilidade de textos e de informações sobre temas de saúde de maior interesse daquelas pessoas que somente têm acesso às unidades e serviços do sistema público. Indiretamente, esses esforços de pesquisa pretendem ajudar a fazer cumprir a Lei 15.263[17] e a Lei 14.874[18], que instituem a clareza textual e o direito de entender como princípios universais e básicos para todas as pessoas, especialmente nos cenários da comunicação entre órgãos públicos e o cidadão. Este empreendimento de pesquisa se justifica à medida que pesquisas e iniciativas em acessibilidade e adaptação de textos em temas de Utilidade Pública no Brasil deparam-se com um cenário social desafiador em termos de letramento linguístico em geral e de Letramento em Saúde. Dados do Indicador de Alfabetismo Funcional [INAF, 2024] mostram que 29% da população brasileira de 15 a 64 anos é classificada entre os níveis de analfabetismo total e rudimentar. Além disso, 36% da população atinge apenas o nível de alfabetismo elementar e somente 10% alcançam o nível de proficiência completa, o que conferiria plena capacidade de interpretar e correlacionar informações textuais complexas. Esse

[14] Disponível gratuitamente em: www.ufrgs.br/textecc/porlexbras/corpop/index.php. Acesso em 08 set. de 2026.
[15] Detalhes em: https://sites.google.com/view/novocorpopsaude20/p%C3%A1gina-inicial
[16] Veja mais detalhes sobre estudo com bulas para o CorPoP Saúde em: https://tildiar.dcc.ufmg.br/projetos/.
[17] https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15263.htm. Acesso em 08 set. de 2026.
[18] https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l14874.htm. Acesso em 08 set. de 2026.

expressivo contingente de pessoas com letramento básico ou rudimentar agrava demandas na esfera da Saúde, pois muitas informações que chegam ao cidadão, sobre diferentes temas, tendem a exibir terminologias e estruturas textuais complexas [Paraguassu, 2024].

No tema das informações sobre uso de medicamentos, a magnitude desse problema já foi mapeada na Pesquisa Nacional sobre Acesso, Utilização e Promoção do Uso Racional de Medicamentos (PNAUM), em estudo transversal de base populacional realizado em todas as regiões do país. Os resultados da PNAUM [Dal Pizzol *et al*., 2019] demonstraram que 54,1% dos brasileiros têm tanto interesse quanto dificuldades para compreender bulas, ressaltando a complexidade sintática e léxica desses documentos, incompatível com o nível de escolaridade e letramento de grande parte da população. Essa barreira específica com bulas – consideradas apenas a sua versão específica para o paciente – faz com que cerca de 25% das pessoas interrompam o uso de medicamentos devido a dúvidas e confusões geradas pela leitura de uma bula. E, justamente, para ajudar a mitigar esse problema, o novo CorPoP Saúde está incluindo bulas do acervo da Agência Nacional de Vigilância Sanitária (Anvisa), na versão específica para pacientes. Todas as bulas, por força de lei, estão disponíveis para acesso público no *site* da Anvisa. No CorPoP Saúde, essas bulas estarão acompanhadas de suas respectivas versões simplificadas, produzidas com apoio de recursos de IA, mas também de versões simplificadas validadas por especialistas da área da Farmácia e por especialistas linguistas. Com essas versões reunidas, a proposta final é gerar materiais educativos resumidos e simplificados para apoiar especialmente os farmacêuticos no momento da dispensação[19] dos medicamentos.

Para guiar a avaliação humana das simplificações geradas automaticamente, usamos, como balizador geral, o CorPoP original, tomado como referência lexical de linguagem simples, visto que continua sendo utilizado com sucesso em diferentes estudos, da Medicina ao Direito [Wives, Finatto, 2025; Finatto, 2025]. O CorPoP original tem cerca de 688 mil palavras/*tokens*, o que representa um vocabulário potencialmente familiar e estruturas sintáticas potencialmente compreendidas por pessoas adultas com letramento limitado, e inclui um acervo de jornais populares e de textos literários especialmente adaptados para leitores adultos que iniciam sua escolarização formal. Embora relativamente pequeno e desprovido de textos específicos sobre temas de Saúde, segue funcionando como um "termômetro lexical", auxiliando a identificar itens lexicais potencialmente distantes de um repertório cotidiano, os quais precisariam ser adaptados ou simplificados.

2. Referencial teórico-metodológico

A pesquisa ampara-se na Linguística de *Corpus*, conforme pioneiramente introduzida no Brasil por Berber Sardinha [2004], e em pesquisas sobre acessibilidade textual, e terminológica ATT) entendendo-se o processo de simplificação análogo à Tradução Intralinguística (TI) [Finatto, Tcacenco, 2021]. A acessibilidade terminológica refere-se à promoção da boa compreensão dos termos científicos ou técnicos, cujos significados precisam ser explicados de modo acessível, ao passo que a acessibilidade textual

[19]A dispensação de medicamentos é o ato profissional do farmacêutico de fornecer um ou mais medicamentos a um paciente, geralmente em resposta a uma receita médica ou de outro profissional autorizado. Esse ato envolve a devida orientação e todos os esclarecimentos necessários para que a pessoa faça o melhor uso do medicamento.

releciona-se às condições de compreensibilidade de um todo de texto.

Os trabalhos citados, entre tantos outros, têm proposto diretrizes para a reformulação de textos que tratam de temas de Utilidade Pública, com o propósito de facilitar a compreensão de suas mensagens por leitores com escolaridade e letramento limitados. Do ponto de vista epistemológico, fazem uma ponte direta com o pensamento de Paulo Freire [1967, 1989], que sempre nos lembra da importância de falarmos *com* as pessoas e não *para* as pessoas. Sob essa perspectiva, a linguagem simples (LS) é entendida não somente como um conjunto de diretrizes e técnicas de escrita, mas como ativação de evidências científicas e empíricas de estudos de longa data. Assim, a LS, que é somente uma parte de todo um processo multifacetado, não se resume a uma adaptação caridosa ou unilateral. Ela é, antes, um exercício de comunicação ativa que respeita o direito ao entendimento e ajuda a promover a inclusão via apropriação de conteúdos textuais em seus diferentes formatos. Afinal, no âmbito da Saúde, precisa-se ir muito além de "decifrar" textos. Por isso, no novo CorPoP Saúde, priorizam-se estratégias elaborativas e a participação dos usuários das informações, em vez de adaptações redutoras, como apenas a inserção de paráfrases após as terminologias, de sinônimos mais fáceis entre parênteses ou mesmo o valor absoluto de evitar a voz passiva.

Com o avanço da Inteligência Artificial (IA), o processamento computacional de textos sobre Saúde ganhou novas perspectivas. Wives e Finatto [2025], Scalercio *et al.* [2026] e Wilkens *et al*. [2026], por exemplo, demonstram a viabilidade do uso de Grandes Modelos de Linguagem (*Large Language Models*, LLMs) para gerar versões simplificadas e grafos de conteúdos de documentos médicos em português. No entanto, Wives e Finatto [2025], entre outros, mostram que esse processo, por melhor que possa ser ou parecer, nunca dispensará a supervisão humana. Afinal, textos gerados de forma automatizada demandam a validação de especializada pois envolvem riscos à vida. Além da precisão conceitual e científica há o perigo de "alucinações" e de desvios gerados por IA. Nesse ponto, também os Pequenos Modelos de Linguagem (*Small Language Models)* já mostram os seus valores, níveis de controle e melhores aplicabilidades.

No âmbito da Avaliação Automática de Leiturabilidade (AAL), a escassez de recursos linguísticos anotados em português brasileiro desafia o treinamento desses modelos computacionais mais robustos [Menezes *et al.*, 2023]. Frente a esse obstáculo, Menezes *et al.* [2023] propõem o uso de técnicas de aumento de dados textuais (por exemplo, com substituição de palavras polissêmicas por diferentes sinônimos possíveis e a retrotradução). Assim, geram-se classificadores de textos complexos e simplificados que combinam métricas de sistemas automáticos como o NILC-Metrix[20] com representações vetoriais contextualizadas. De nossa parte, em meio a desafios tão robustos, propomos um *corpu*s especialmente compilado e curado, com textos originais mais e menos complexos e pareados com simplificações automáticas devidamente tratadas e validadas por revisores especialistas em Saúde e linguistas.

3. Segmentos do *subcorpus* bulas da Anvisa e metodologia de validação humana de simplificações automáticas

O *subcorpus* de bulas divide-se em três segmentos específicos, descritos abaixo.

[20] NILC-Metrix: https://github.com/nilc-nlp/NILC-Metrix. Acesso em 08 de set. de 2026.

- ***Subcorpus* de bulas de Hipertensão** (bloco HT): composto por 30 bulas de anti-hipertensivos, correspondentes aos 10 medicamentos mais distribuídos no Programa Farmácia Popular do Brasil[21], representados por bulas de três fabricantes. Essas diferentes bulas foram colhidas visando-se registrar eventuais variações textuais do padrão da Anvisa, que são permitidas aos fabricantes. Esse *subcorpus*, composto pelas 30 bulas originais e respectivas versões simplificadas automaticamente, já passou por avaliação de linguistas e farmacêuticos associados ao projeto, tendo gerado 30 versões simplificadas finais devidamente validadas. Essas versões finais, devidamente revisadas e reescritas pelos especialistas, somam-se ao *subcorpus* geral na condição de bulas simplificadas validadas em versão final. A seguir, um exemplo do trabalho em uma das bulas do bloco HT (Quadro 1).

Quadro 1. Exemplo do trabalho em uma das bulas do bloco HT

Bula original 2. COMO ESTE MEDICAMENTO FUNCIONA? O Atenolol age preferencialmente sobre os receptores localizados no coração e na circulação, reduzindo a pressão arterial, quando usado continuamente. O Atenolol começa a ter uma ação significativa dentro de 1 hora após sua administração por via oral, atingindo seu efeito máximo em 2 a 4 horas. Esse efeito é mantido por no mínimo 24 horas.
Simplificação automática por IA 2. COMO ESTE MEDICAMENTO FUNCIONA? O Atenolol age principalmente no coração e nos vasos sanguíneos. Ele faz o coração bater um pouco mais devagar e com menos força. Com isso, a pressão do sangue diminui. O efeito começa em cerca de 1 hora depois que você toma o comprimido. O efeito máximo acontece entre 2 a 4 horas depois de tomar o comprimido. O efeito dura pelo menos 24 horas.
Versão simplificada após revisão final – linguística e técnica 2. COMO ESTE MEDICAMENTO FUNCIONA? O Atenolol age principalmente no coração e nos vasos sanguíneos. Quando o Atenolol é usado todo dia, ele relaxa os vasos sanguíneos e diminui os batimentos do coração. Com isso, a pressão do sangue diminui. O efeito começa em cerca de 1 hora depois que você toma o comprimido. O efeito máximo acontece entre 2 a 4 horas depois de tomar o comprimido. O efeito dura pelo menos 24 horas.

- ***Subcorpus* de bulas de Oncologia** (bloco ONCO): composto por 49 bulas de medicamentos de baixa e alta complexidade, usados em diferentes tipos de terapias farmacológicas para combater diferentes tipos de câncer. Esses textos passaram pelos mesmos processos de simplificação do bloco HT. Encontram-se atualmente na fase de validação linguística e passarão, posteriormente, pela validação técnica de farmacêuticos e pela etapa da redação final simplificada e consolidada.
- ***Subcorpus* de triagem geral** (*corpus* Likert): composto por 300 simplificações automáticas. Essas simplificações originaram-se de um lote aleatório de 150 bulas originais do bulário da Anvisa. Assim, o *corpus* Likert é uma amostra variada aleatória do universo de bulas, diferenciando-se por buscar espelhar a grande

[21] https://www.gov.br/saude/pt-br/composicao/sectics/farmacia-popular. Acesso em 08 set. de 2026.

variedade de medicamentos e/ou tratamentos envolvidos, bem como dos diversos fabricantes-autores dos textos. As avaliações humanas dessas simplificações, diferente dos blocos HT e ONCO, não envolvem reescritas ou comentários dos revisores especialistas, mas são feitas apenas por meio de pontuações atribuídas em uma escala Likert de 1 (muito ruim) a 5 (muito bom) em perguntas respondidas pelos revisores linguistas e pelos revisores técnicos farmacêuticos. Na avaliação linguística, há 3 perguntas sobre a qualidade da simplificação (léxico, estrutura, fluência) e 1 sobre a qualidade textual. Na avaliação técnica, há 2 perguntas sobre a qualidade geral da simplificação, preservação do significado principal e qualidade/adequação da remoção de conteúdos e/ou da inserção de informações novas, não contidas no documento original. Esse *subcorpus* de triagem serve para gerar um plano genérico das diferentes avaliações e percepções humanas e para observar compatibilidades, convergências ou discrepâncias entre diferentes julgamentos.

4. *Subcorpus* de bulas de medicamentos para hipertensão arterial (HT)

A Tabela 1 traz a caracterização lexical geral e estatísticas de superfície do *subcorpus* HT, o *corpus* CorPoP original, antes citado, é a referência lexical de simplicidade.

Tabela 1. Caracterização lexical geral e estatísticas do *subcorpus* HT

Corpus/*subcorpus*	Palavras (*tokens*)	Tipos (*types*)	TTR (*types/token ratio*)
***Corpus* original (referência simplicidade)**	688.788	32.311	0,0469
Bulas originais	77.306	3.116	0,0403
Bulas simplificadas por IA	41.991	2.279	0,0543
Bulas após validação + reescrita humana	48.532	2.513	0,0518

A Tabela 1 demonstra que ambas as versões simplificadas (por IA e pelos revisores humanos) reduziram substancialmente o volume textual em comparação à bula original, indicando um processo de diminuição textual relevante (de cerca de 37% de redução em termos de palavras no texto final). Entretanto, a versão validada por revisores apresentou um aumento tanto em *tokens* (48.532) quanto em *types* (2.513) em comparação à versão gerada por IA, que gerou 41.991 *tokens* e 2.279 *types*. Esse comportamento sinaliza, quantitativamente, o caráter elaborativo dos revisores: ao passo que a IA realiza um resumo restritivo priorizando cortes e substituições pontuais, os especialistas humanos optam por reintroduzir ou expandir trechos para inserir explicações didáticas e preencher lacunas conceituais.

5. Inteligência Artificial e padrões humanos de simplificação em HT

A análise quantitativa e qualitativa das escolhas textuais revelou contrastes nítidos entre

os limites do modelo de IA e a validação e reescritas simplificadas humanas. Usamos técnicas de PLN como recurso analítico para identificar e caracterizar os padrões da tomada de decisão revisória [Wives, Finatto, 2025]. Essas técnicas mostram que as simplificações por IA, ainda que de ótima qualidade, tendem a ser redutoras, baseadas em cortes de sentenças e em resumos que, às vezes, eliminam termos especializados e contextos sem explicá-los. Em contrapartida, os revisores adotam uma abordagem elaborativa, expandindo o texto para inserir paráfrases explicativas e preencher lacunas conceituais, conforme indicam as estatísticas de *tokens* e *types* detalhadas na Tabela 1. Com tais dados, mapeamos as decisões tomadas pelos revisores em três camadas específicas de análise:

- **camada lexical**: a terminologia especializada foi reescrita com o uso de termos de alta frequência identificados no *corpus* original CorPoP. Por exemplo, com a reformulação do termo "hipertensão" por "pressão alta", ou combinou-se o empréstimo da terminologia com uma paráfrase explicativa entre parênteses. Por exemplo, "excretado primordialmente por via renal" reformulado por "sai do corpo principalmente através da urina (pelos rins)";
- **camada discursiva e textual**: houve didatização das sentenças, substituindo-se a sumarização automática de palavras por explicações funcionais que tenderiam a facilitar a compreensão, por exemplo, ao reformular "bloqueando os receptores beta-adrenérgicos" por "ajuda o coração a bater mais devagar e com menos força";
- **camada pragmática**: substituído o registro passivo, característico da impessoalização típica das bulas originais, por uma estrutura que estabelece comunicação ativa e direta com o leitor, fazendo uso recorrente do pronome de tratamento "você", por exemplo, em "Podem ocorrer tonturas" por "Você pode sentir tontura".

A Tabela 2, a seguir, resume as estratégias de decisão humana no *subcorpus* HT por meio das camadas de análise antes citadas.

Tabela 2. Estratégias de decisão humana no *subcorpus* HT em camadas de análise

	Bula original HT	**Simplificação por IA**	**Validação/revisão humana**
CAMADA LEXICAL			
1	Este fármaco é excretado primordialmente por via renal.	O remédio sai do corpo pela urina.	O remédio sai do corpo principalmente através da urina (pelos rins).
2	Pacientes com disfunção renal requerem ajuste posológico.	Quem tem problema nos rins precisa de doses menores.	Pessoas com problemas nos rins podem precisar de doses diferentes.
CAMADA DISCURSIVA			
3	O fármaco atua bloqueando os receptores beta-adrenérgicos.	O remédio age no coração diminuindo a pressão.	O remédio ajuda o coração a bater mais devagar e com menos força.
CAMADA PRAGMÁTICA			
4	Podem ocorrer tonturas, fadiga e cefaleia.	Pode causar tontura, cansaço e dor de cabeça.	Você pode sentir tontura, cansaço e dor de cabeça.

5	A posologia deve ser estabelecida estritamente pelo facultativo.	Siga a dose que o médico mandou.	Siga exatamente a dose que o médico receitou.

Compilamos e validamos esse padrão elaborativo e centrado no leitor no *subcorpus* HT, usando-o no andamento do *subcorpus* ONCO. Embora as bulas oncológicas envolvam uma terminologia ainda mais complexa e tratamentos de alta gravidade, a validação humana tem buscado assegurar o mesmo compromisso com a acessibilidade textual e a segurança clínica, preenchendo eventuais lacunas deixadas pela simplificação por IA. A avaliação empírica em larga escala no *subcorpus* de triagem geral (Likert), ainda em andamento, permitirá comparar como os avaliadores humanos percebem ou qualificam as simplificações geradas automaticamente, sendo que elas foram consideradas, de um modo automático, como as melhores e as piores em termos apenas de escores de níveis de leiturabilidade e de similaridade semântica.

Na dimensão pragmática e discursiva, a tendência das reformulações feitas por pessoas foi preservar o caráter agentivo e demarcar a autonomia do leitor. Por isso, substituíram a designação passiva clássica de quem recebe o tratamento por "pessoa" em vez de "paciente", bem como as expressões que sugerem uma autoridade impositiva por ações de caráter profissional, como "prescrever" por "receitar" ou "orientar". Desse modo, tenderam a reconhecer o leitor-paciente como um ente ativo.

6. Índices quantitativos de leiturabilidade das simplificações em HT

Para avaliar o potencial de leiturabilidade das bulas do *subcorpus* HT, em suas diferentes versões, usamos a fórmula do Índice de Facilidade de Leitura de Flesch originalmente adaptada para o português brasileiro por Martins *et al.* [1996]. Mas, desta vez, com a medida calculada pela nova ferramenta Vidya Text[22]. Além do Índice Flesch revisado (IF-Vidya), traz uma adaptação da fórmula de leiturabilidade de Dale-Chall[23] [Chall, Dale, 1995] (Vidya-DC). A Tabela 3 detalha esses índices obtidos com a Vidya Text.

Tabela 3. Índices quantitativos de leiturabilidade das simplificações de bulas HT

Corpus/*subcorpus*	IF-Vidya	Interpretação Índice Flesch	Vidya-DC	Nível de escolaridade necessário para entender
Bulas originais HT	22,4	Muito difícil	10,08	Pós-graduação
Simplificação IA	40,9	Difícil	9,63	Universitário
Simplificação final - validação humana	42,8	Difícil	9,34	Universitário (inicial)

Na Tabela 3, o original das bulas HT situa-se em um nível de extrema complexidade (22,4), o que exigiria letramento de nível de pós-graduação. Embora a simplificação por IA tenha elevado o índice para 40,9 e a validação humana alcançado apenas 42,8, ambos os textos ainda se situam na faixa "difícil" (nível universitário). Essa persistência de

[22] Vidya: http://www.ufrgs.br/att/vidya. Acesso em 08 de set. de 2026.
[23] Estima a facilidade de entendimento de um texto com base na proporção de uso de um conjunto de palavras potencialmente difíceis de entender – palavras de baixa frequência – em relação ao número total de palavras do texto. Quanto maior o índice, maior a tendência de ser mais fácil de ler.

complexidade possivelmente decorre das restrições inerentes ao domínio farmacológico: a obrigatoriedade de preservar com exatidão dados numéricos, doses, efeitos adversos graves e restrições de uso talvez impeça reduções sintáticas ou lexicais mais drásticas e exija a manutenção de termos de baixa ocorrência no léxico comum.

Quanto ao Índice Flesch, a diferença de pontuação entre as bulas originais (40,9) e as simplificações humanas (42,8) é pequena. Isso ocorre porque as fórmulas de leiturabilidade clássicas têm como base o comprimento médio das frases e o número de sílabas por palavra. Assim, acabam penalizando as estratégias elaborativas que aumentam a extensão das sentenças para explicar termos especializados de baixa frequência. Apesar dessa limitação métrica, o *subcorpus* HT na versão revisada e finalizada por pessoas obteve melhorias robustas nos índices Dale-Chall, atingindo os menores valores de distância lexical (9,34 Vidya-DC), o que reflete substituições e mediação terminológica efetivas, com o uso de palavras de maior frequência.

Para mapear com mais detalhes essa adaptação lexical, realizamos uma análise comparativa da frequência relativa de itens de maior frequência nas bulas originais HT, nas simplificadas por IA e sua presença no *corpus* lexical CorPoP original. A Tabela 4 consolida a distribuição dos principais itens identificados no *subcorpus* HT.

Tabela 4. Distribuição dos principais termos de conteúdo identificados no *subcorpus* HT.

Palavra/ item	Frequência bula original (%)	Frequência simplificação por IA (%)	Frequência *corpus* de referência lexical simples (%)	Razão de sobrefrequência (simplificação humana/*corpus* ref. simples)	Diferença absoluta (simplificação humana e simplificação por IA)
médico	1,329%	1,263%	0,029%	43,9	-0,066%
pressão*	0,867%	0,938%	0,006%	161,4	+0,071%
você	0,914%	0,870%	0,420%	2,1	-0,045%
dose	0,953%	0,870%	0,001%	1197,8	-0,083%
medicamento	0,779%	0,826%	0,002%	569,1	+0,048%
coração*	0,562%	0,777%	0,052%	14,8	+0,215%
sangue*	0,572%	0,680%	0,011%	60,0	+0,108%
usar	0,564%	0,525%	0,014%	38,5	-0,039%
dia	0,641%	0,556%	0,172%	3,2	-0,084%
rins	0,367%	0,451%	0,001%	444,0	+0,085%
tratamento	0,398%	0,398%	0,020%	19,8	0,000%
problemas	0,181%	0,383%	0,049%	7,9	+0,202%

* No *corpus* de referência de simplicidade lexical, o CorPoP original, identificamos essas palavras predominantemente em contextos figurados ou metafóricos, como, por exemplo, “bom de coração” e “sangue bom”.

A distribuição de dados lexicais indica que revisores humanos tendem a aumentar deliberadamente a frequência de termos concretos e familiares em detrimento de terminologias abstratas ou imperativas da simplificação por IA. Os revisores empregam o vocábulo “problemas” mais do que o dobro de vezes (0,383%) em comparação com a

simplificação por IA (0,181%), com aumento de +0,202%. Esse comportamento tende a salientar uma escolha intencional por palavras de alta cobertura e familiaridade cotidiana, alinhadas ao repertório empírico do CorPoP original. De modo semelhante, itens biológicos como "coração" (+0,215%), "sangue" (+0,108%) e "rins" (+0,085%) apresentam elevações significativas na simplificação revisada e finalizada por pessoas.

7. Considerações finais e perspectivas

A versão final simplificada e revisada por linguistas e farmacêuticos no *corpus* amostral HT apresentou melhor leiturabilidade e, principalmente, estratégias didáticas. Essas estratégias não foram reconhecidas nem acionadas no protocolo de simplificação de bulas já elaborado por Scalercio *et al*. [2026]. Assim, tais estratégias devem estar contempladas nos comandos para simplificações automáticas em todo o *subcorpus*. Como encaminhamentos futuros, prevemos uma avaliação empírica da compreensão leitora de uma nova amostra de bulas revisadas e finalizadas por pessoas e as produzidas somente por IA, junto a cidadãos de escolaridade e letramento limitados. Além disso, daremos andamento às etapas de validação linguística e técnica do *subcorpus* de bulas ONCO e à consolidação das análises estatísticas do *subcorpus* de triagem (bloco Likert) integrando todos esses materiais e achados aos repertórios do novo CorPoP Saúde.

Questões éticas

Utilizamos a ferramenta de IA Gemini para ajustes de formatação e revisão gramatical deste texto. O uso do ChatGPT, versão 5.2 [OpenAI, 2026], para gerar uma simplificação inicial dos textos das bulas foi alertado aos revisores linguistas e farmacêuticos e será também alertado aos usuários finais desses textos quando acessados no novo CorPoP Saúde. A concepção geral deste artigo foi de Maria José B. Finatto. Bianca Pasqualini elaborou a análise dos dados dos revisores em camadas.

Agradecimentos

Agradecemos os apoios dos órgãos: CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), bolsa PQ 307088/2023-5; FAPERGS – Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, bolsa Pós-Doc PDTI1 Proc. 25/2551-0002702-0; FAPERJ - Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, processos SEI-260003/002930/2024, SEI-260003/000614/2023 e CAPES. Também aos Institutos Nacionais de Ciência e Tecnologia do CNPq, INCT-IAIA (bolsa 406417/2022-9), INCT-TILD-IAR (projeto "Bulas brasileiras e uso racional de medicamentos" e bolsa 408490/2024-1) e INCT-IAPROBEM (bolsa 408589/2024-8).

Referências

Berber Sardinha, T. (2004). *Linguística de Corpus*. Barueri: Manole.

Chall, J. S. and Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Cambridge: Brookline Books.

Conselho Federal de Farmácia (2016). *Serviços farmacêuticos diretamente destinados ao paciente, à família e à comunidade: contextualização e arcabouço conceitual*. Brasília: Conselho Federal de Farmácia.

Dal Pizzol, T. S. *et al.* (2019). Medicine package inserts from the users' perspective: are they read and understood?. *Revista Brasileira de Epidemiologia*, v. 22, p. e190009.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: instructions. *Educational Research Bulletin*, v. 27, n. 2, p. 37-54.

Finatto, M. J. B. e Tcacenco, L. M. (2021). Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. *TradTerm*, v. 37, n. 1, p. 30-63. DOI: https://doi.org/10.11606/issn.2317-9511.v37p30-63. Disponível em: https://revistas.usp.br/tradterm/article/view/168327.

Finatto, M. J. B. (2024). Da linguagem simples à acessibilidade textual e terminológica: um percurso com e na Linguística. In: Silva, A. H. P., Lagares, X. C. e Maia, M. (orgs.), *Linguagem simples para quem?*, Editora da Abralin. Disponível em: https://editora.abralin.org/wp-content/uploads/2024/11/Linguagem-simples-para-quem_PDF.pdf.

Freire, P. (1967). *Educação como prática da liberdade*. Rio de Janeiro: Paz e Terra.

Freire, P. (1989). *A importância do ato de ler: em três artigos que se completam*. São Paulo: Cortez.

Indicador de Alfabetismo Funcional - INAF (2024). *Relatório INAF Brasil 2024: letramento e competências digitais*. São Paulo: Ação Educativa; Instituto Paulo Montenegro.

Martins, T. B. F. et al. (1996). *Readability formulas applied to textbooks in Brazilian Portuguese*. Notas do ICMSC-USP, São Carlos.

Menezes, L. C., Paes, A. e Finatto, M. J. B. (2023). Abordagem baseada em aumento de dados para avaliação automática de leiturabilidade. *Domínios de Lingu@gem*, v. 17, p. e1721.

Paraguassu, L. B. (2024). *Linguagem Simples na Saúde: contribuições dos estudos da linguagem para o Letramento em Saúde no Brasil*. Tese de Doutorado em Letras, Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Ponomarenko, G. L. (2022). *Compreender para consentir: a importância da tradução intralingual em termos de consentimento da área médica*. Trabalho de Conclusão de Curso, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Ponomarenko, G. L. e Rebechi, R. (2023). Entender para consentir: Uma proposta de tradução intralingual em Termos de Consentimento Informado. *Letras & Letras*, Uberlândia, v. 39, p. e3904, p. 1–27. DOI: 10.14393/LL63-v39-2023-04. Disponível em: https://seer.ufu.br/index.php/letraseletras/article/view/68971.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing – ACL*.

Scalercio, A. et al. (2025). Annotation guidelines and challenges for automatic simplification of Portuguese drug leaflets. In: *Proceedings of Propor/Lang4Health*.

Disponível em: https://aclanthology.org/2026.propor-2.19/

Wives, L. K. e Finatto, M. J. B. (2025). Usando LLMs para simplificar e representar documentos médicos. In: *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados* (SBBD 2025), Porto Alegre, SBC, p. 415-425.

Wilkens, R., Caseli, H., Neris, V. and Villavicencio, A. (2026). The visible and the latent linguistic clues of mental health in Brazilian Portuguese textual posts. In: Souza, M. et al. (eds.), *Proceedings of the 17th International Conference on Computational Processing of Portuguese* (PROPOR 2026) - Vol. 2, Association for Computational Linguistics, p. 135-147. Disponível em: <https://aclanthology.org/2026.propor-2.21/>.

UD_Kadiweu-Unicamp: expandindo a diversidade tipológica da coleção Universal Dependencies com o primeiro *treebank* de uma língua da família Guaicuru

Leonel F. de Alencar[1,2], Maria Filomena S. Sandalo[1], Charlotte Marie C. Galves[1]

[1]Instituto de Estudos da Linguagem – Universidade Estadual de Campinas (Unicamp)
Campinas – SP – Brasil [2]Centro de Humanidades – Universidade Federal do Ceará
Fortaleza – CE – Brasil

leonel.de.alencar@ufc, {sandalo,galvesc}@unicamp.br

Abstract. *With 353 treebanks for 193 languages from 36 language families in release 2.18, Universal Dependencies constitutes the most widely adopted standard for morphosyntactic annotation. Nevertheless, the enormous linguis tic diversity of the Americas remains underrepresented. This article presents UD_Kadiweu-Unicamp, the first treebank for a Guaikuruan language in the UD collection. The treebank is derived from the Corpus Kadiwéu – Gramática Pedagógica, automatically annotated in the Penn Treebank style, but with an additional layer of morphemic segmentation. We discuss the process of automatic conversion and manual revision and present basic statistics together with automatic vali dation and evaluation results, which place UD_Kadiweu-Unicamp among the best-rated treebanks for Amerindian languages.*

Resumo. *Com 353 treebanks de 193 línguas pertencentes a 36 famílias linguís ticas na versão 2.18 da coleção homônima, o modelo Universal Dependencies constitui o padrão mais difundido para anotação morfossintática. No entanto, a enorme diversidade linguística das Américas ainda se encontra sub-representada. Este artigo apresenta o UD_Kadiweu-Unicamp, o primeiro treebank de uma língua da família Guaicuru na coleção UD. O treebank deriva do Corpus Kadiwéu – Gramática Pedagógica, anotado automaticamente no estilo do Penn Treebank, mas com uma camada adicional de segmentação morfêmica. Discutimos o processo de conversão automática e revisão manual e apresentamos estatísticas básicas, bem como resultados de validação e avaliação automáticas, que posicionam o UD_Kadiweu-Unicamp entre os treebanks de línguas ameríndias mais bem avaliados.*

1. Introdução

Com 353 *treebanks* de 193 línguas na versão 2.18, de 15/05/2026, somando 2,3 milhões de sentenças e 37 milhões de *tokens* (Zeman et al., 2026), a coleção Universal Dependencies (UD) constitui um empreendimento sem paralelo. Essa coleção foi inaugurada em 2015 com apenas 10 *treebanks* de 10 línguas majoritárias, perfazendo cerca de 185 mil sentenças e 3,5 milhões de *tokens* (Nivre et al., 2015). Ao longo de uma década e de sucessivas versões, expandiu-se em progressão geométrica, passando a abranger 36 famílias linguísticas. Esses números revelam a crescente adoção da teoria UD para a anotação morfossintática de *corpora* (Marneffe et al., 2021; Zeman, 2023).

Embora a América seja o continente com a maior diversidade linguística (Ciucci; Kilarski, 2025), UD incorporou a primeira língua ameríndia apenas na versão 2.4 (Thomas, 2019), enquanto o português estreou na versão 1.2. O número de línguas ameríndias aumentou nos últimos seis anos. Na *release* 2.18, há 25 treebanks para 25 dessas línguas. Apesar desses avanços, a família indo-europeia ainda predomina, com 193 treebanks de 80 línguas. A diversidade linguística das Américas permanece sub-representada. O Brasil possui cerca de 150 línguas indígenas ainda vivas (Storto, 2019). Rodrigues (1986) identifica no país 13 famílias linguísticas e 10 línguas isoladas.

Este artigo trata da construção, composição, anotação e avaliação do UD_Kadiweu-Unicamp, o mais recente treebank de língua ameríndia da coleção UD, o primeiro de UD para uma língua da família Guaicuru.

2. A língua kadiwéu e seus *corpora* na Plataforma Tycho Brahe

O kadiwéu (ISO-639: kbc) é uma língua polissintética falada no estado de Mato Grosso do Sul, a única representante da família linguística Guaicuru no Brasil (Sandalo, 1995). Está severamente ameaçada: de aproximadamente 1.500 pessoas da etnia kadiwéu, menos de 300 a falam (Pires, 2022).

As línguas polissintéticas caracterizam-se pela formação de palavras complexas correspondendo a sentenças inteiras em línguas como o português (Bußmann, 2002, p. 525), como em (1), do *Corpus Kadiwéu – Gramática Pedagógica* (Sandalo et al., 2024b) (doravante KadGramPed). A polissíntese tem-se revelado um desafio para o processamento computacional (Tyers; Mishchenkova, 2020; Gibert et al., 2025). As sentenças (2)-(4), do mesmo *corpus*, evidenciam outras facetas da complexidade gramatical da língua. Classificadores numerais têm como função principal individualizar nomes, como em (2), onde ABS(ENT) refere-se à raiz dêitica *ca*, que indica que o referente da expressão nominal está ausente ou fora da visão do locutor (Sandalo; Michelioudakis, 2016). Em vez de adposições, a língua possui diferentes sufixos de aplicativo. Um objeto de primeira ou segunda determina a ordem SOV (Sandalo; Galves, 2025, p. 2). Noutras configurações, porém, a ordem é livre. O verbo concorda com o sujeito ou o objeto, dependendo da hierarquia de pessoa 2>1>3. O kadiwéu licencia *pro-drop* tanto para sujeito quanto objeto. Como o possuidor precede ou subsegue o núcleo nominal que expressa o *possessum,* portador do genitivo, sentenças como (3) são estruturalmente ambíguas: *niganigawanigi* pode também constituir complemento de *libinienigi,* que funciona como predicativo em ambas as leituras.

(1) jipegitege
1ERG:ficar_perto:APL.3F
estou perto dela

(2) ica liGeladi ipegitegi
M:ABS 3GEN:casa 3ERG:ficar_perto:APPL.3M
'a casa dela/dele fica perto dele'

(3) ica liwigo niganigawanigi libinienigi
M:ABS 3GEN:espírito menino 3GEN:beleza:DIM:M
esta fotografia da criança é bonita (esta fotografia é a beleza da criança)

(4) akaami oko jipacaGataGagitiwaji
2 1p 1ERG:ficar_perto:PL:APPL.2:M:PL

'estamos perto de vocês'

O UD_Kadiweu-Unicamp baseia-se no KadGramPed, um dos dois *treebanks* de estrutura sintagmática do kadiwéu disponíveis na Plataforma Tycho Brahe (doravante PTB) (Veronesi, 2026). Em versões futuras, incorporará textos do *Corpus Kadiwéu* (Sandalo et al., 2024a) (doravante CorKad), composto de narrativas originárias. Esses corpora, atualmente em desenvolvimento, são anotados automaticamente conforme uma extensão do esquema do Penn Treebank (Galves et al., 2017; Sandalo; Galves, 2023).[24] Para tanto, aplica-se um *parser* de regras determinístico sobre as sentenças com etiquetas de classe de palavra provenientes de um léxico computacional. Todos esses recursos integram o projeto de pesquisa *Digitally annotated corpora of Brazilian Indigenous languages with automatic translations* (DACILAT).[25] Podem ser consultados on-line ou exportados em JSON, facilitando sua utilização em *pipelines* externos à PTB.

Quadro 1. Segmentação morfêmica e glosas de (2).

Morfema	Categoria	Glosa
i	Gnr	M
ca	Ncl	ausente
l	Gen	3
iGeladi	n	casa
i	Erg	3
pegi	v	ficar_perto
tegi	Apl	3M

O KadGramPed constitui-se de 206 sentenças elicitadas de falantes nativos entre 2025 e 2026, das quais 183 estão com status `FINALIZADO` no atual estágio de desenvolvimento do parser, que, na versão de 11:42:35 de 28/08/2026, alcança uma acurácia de 88,52% nesse subconjunto de sentenças. As árvores de 23 sentenças estão com status `REVISAR`. Várias dessas sentenças representam casos de ambiguidade estrutural, para as quais o parser gera uma árvore possível, mas que não corresponde ao estímulo da elicitação. A classificação de status foi feita manualmente, com frequência envolvendo adjudicação entre os autores deste artigo.

O KadGramPed embasará o desenvolvimento de um tradutor automático. A maior parte desse *corpus,* como (1)-(4), constitui-se de traduções do conjunto de dados de Alencar (2021), formado por predicações qualificativas do tipo copulativo em português, inglês e nheengatu, associadas a representações semânticas geradas por uma gramática computacional no formalismo Grammatical Framework (Ranta, 2011).

24 A escolha desse esquema de anotação decorre de histórica parceria entre a Unicamp e a Universidade da Pensilvânia, no âmbito da qual se desenvolveram o Tycho Brahe Parsed Corpus of Historical Portuguese e o respectivo *parser* (Faria; Galves; Magro, 2024), utilizado como base para o do kadiwéu.

25 Processo FAPESP n° 22/09158-5.

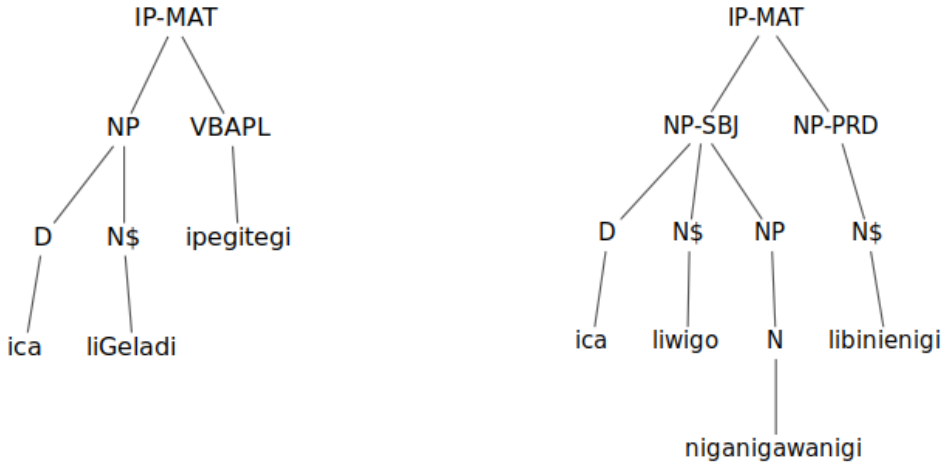


Figura 1. Árvores sintagmáticas de (2) e (3) geradas pelo *parser* da PTB.

O KadGramPed e o CorKad preservam a ortografia de palavras amalgamadas como *adakake*, formada por *aG* 'não' e *dakake* 'duro', mas as desmembram na árvore sintática. Diferentemente do Penn Treebank (Taylor; Marcus; Santorini, 2003), ambos comportam, além da estrutura de constituintes, a segmentação morfêmica de palavras complexas, fazendo jus à riqueza morfológica da língua (Galves et al., 2017). Esses dois processos de segmentação realizam-se automaticamente para as palavras registradas no léxico. O Quadro 1 exemplifica o segundo tipo.

3. Metodologia de construção do *treebank*

A disseminação do paradigma dependencial no *parsing* é mais recente que a utilização de modelos de estrutura de constituintes. O mais influente exemplo dessa última abordagem, o Penn Treebank, remonta ao início dos anos de 1990. Diversos conversores para o formato dependencial foram implementados, com maior ou menor grau de acurácia, nenhum, pelo que pudemos apurar, 100% exato (Johansson; Nugues, 2007). Como se limitam ao esquema de anotação do Penn Treebank, atrelado, por sua vez, a particularidades do inglês, essas ferramentas se revelaram impraticáveis para conversão de *treebanks* como o Icelandic Parsed Historical Corpus (IcePaHC), que, se seguem o formato geral do Penn Treebank, deste divergem significativamente a fim de acomodar as especificidades de uma língua com morfologia e sintaxe bastante diferentes do inglês. Isso levou Arnardóttir et al. (2020) a desenvolver um conversor próprio para adaptar o IcePaHC ao modelo UD. A mesma equipe aplicou a ferramenta para construir a versão em UD do FarPaHC, um treebank histórico do feroês nos moldes do IcePaHC.

Arnardóttir et al. (2020) ressaltam que não avaliaram formalmente seu conversor e os resultados do processo de conversão. Além disso, os dois *treebanks* de UD construídos com a ferramenta se encontram inválidos em UD 2.18: o do islandês apresenta 5.885 erros e 17.152 advertências, o do feroês, 332 e 7.655, respectivamente. Por outro lado, os esquemas de anotação dessas duas línguas nórdicas divergem significativamente do adotado pelos corpora do kadiwéu, que também possuem uma camada de segmentação morfêmica inexistente naqueles *treebanks*.

Como o KadGramPed (a exemplo do CorKad) e sua documentação estão em construção, com muitas análises sujeitas a revisão, decidimos construir nosso próprio

conversor para UD. No desenvolvimento do UD_Kadiweu-Unicamp e da ferramenta de conversão, adotamos alguns princípios norteadores. Em primeiro lugar, as diferentes camadas de anotação do *treebank* deveriam receber o atributo `manual native`, o mais valorizado na avaliação de UD pelo *script* `evaluate_treebank.pl`. Isso implica revisar manualmente cada aspecto da anotação, em vez de confiar num processo automático sujeito a erros. Em segundo lugar, o conversor deveria gerar um CoNLL-U válido mesmo a partir de informações incorretas ou incompletas na representação de entrada. Em terceiro lugar, deveria expressar em UD toda a riqueza das análises do KadGramPed, incluindo a segmentação de *tokens* multipalavras e a informação codificada na segmentação morfêmica. O Quadro 2 exemplifica a reformulação, em UD, das representações sintática e morfológica de (2) na Figura 1 e no Quadro 1. Não contempla o classificador numeral *ca* por se tratar de categoria inexistente no inventário de *features* da coleção UD, lacuna que pretendemos sanar na próxima *release*. A PTB não representa explicitamente os lemas, recuperados, pelo conversor, sobretudo a partir das raízes lexicais. A Figura 2 exibe a árvore dependencial correspondente.

Quadro 2. Oito primeiras colunas do formato CoNLL-U para (2).

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
1	ica	ica	DET	D	Gender=Masc Number=Sing PronType=Dem	2	det
2	liGeladi	Geladi	NOUN	N$	Gender=Masc Number=Sing Person[psor]=3	3	nsubj
3	ipegitegi	pegi	VERB	VBAPL	Mood=Ind Person[erg]=3 Person[obj]=3 VerbForm=Fin Voice=Appl	0	root
4	.	.	PUNCT	PUNCT		3	punct

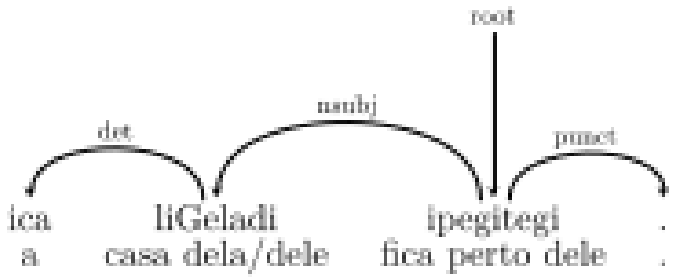


Figura 2. Árvore dependencial de (2) gerada pelo conversor.

Para orientar o desenvolvimento de um conversor em Python, o primeiro autor deste artigo, responsável pelo *treebank* de UD do nheengatu (Alencar et al., 2026), anotou manualmente seis das sentenças menos complexas, procurando reproduzir fielmente em UD a toquenização, a análise sintagmática e a segmentação morfêmica adjudicadas no *corpus* de partida. Essas sentenças constituíram a primeira versão do UD_Kadiweu

Unicamp, que passou a funcionar como padrão ouro. Uma versão inicial do conversor foi aplicada para gerar anotações em CoNLL-U para novas sentenças, utilizando as informações constantes na exportação em JSON do KadGramPed.

Em iterações sucessivas, o conversor foi aperfeiçoado submetendo sua saída ao validador de UD (`validate.py`) e corrigindo manualmente os problemas detectados. Periodicamente, sentenças completamente validadas foram extraídas do CoNLL-U gerado pelo conversor, revisadas manualmente e adicionadas ao padrão ouro. A partir daí, novas informações sobre relações sintáticas, lematização e traços morfológicos foram incorporadas ao conversor.

Em algumas sentenças, a aplicação do conversor ou a revisão de sua saída revelou anotações incorretas ou incompletas nos dados-fonte. Muitos desses casos foram corrigidos na PTB, implicando, frequentemente, mudanças nas regras do *parser*, e novas exportações em JSON das sentenças revisadas foram geradas para reaplicação do conversor. Outras questões, como a atribuição exaustiva de rótulos funcionais, como SBJ para sujeito, aguardam a elaboração (ou a reformulação) de regras do *parser*, a implementação de novas soluções tecnológicas ou o aprofundamento de aspectos da estrutura gramatical da língua.

4. Estatísticas básicas do UD_Kadiweu-Unicamp

Para cada treebank da coleção UD, o *script* `conllu-stats.pl` computa uma série de estatísticas não apenas sobre o tamanho do corpus, mas também sobre os diferentes níveis de anotação. Nesta seção apresentamos dados relativos à *release* 2.18.

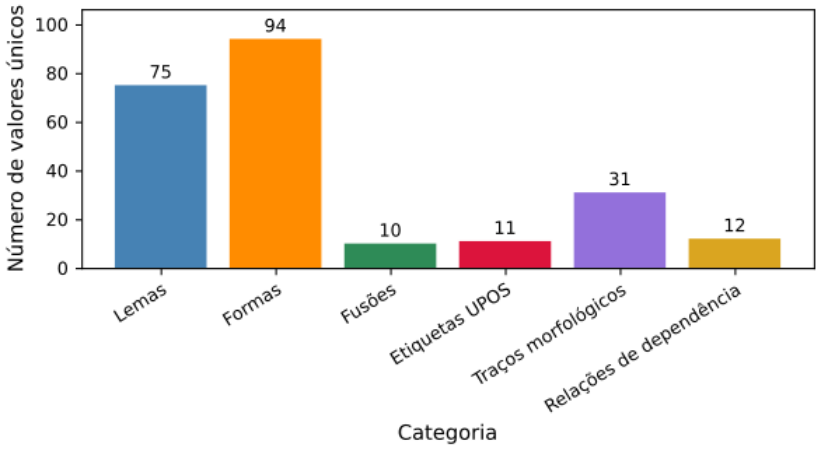


Figura 3. Frequências sem repetição de diferentes categorias de anotação.

O *treebank* contém 71 sentenças, totalizando 306 *tokens*. Desses, 12 são *tokens* multipalavras, cada um dos quais se desdobra em duas palavras sintáticas. A Figura 3 e a Figura 4 exibem as frequências, sem repetições, de lemas, formas, palavras amalgamadas, etiquetas de classe de palavras, traços morfológicos e relações de dependência.

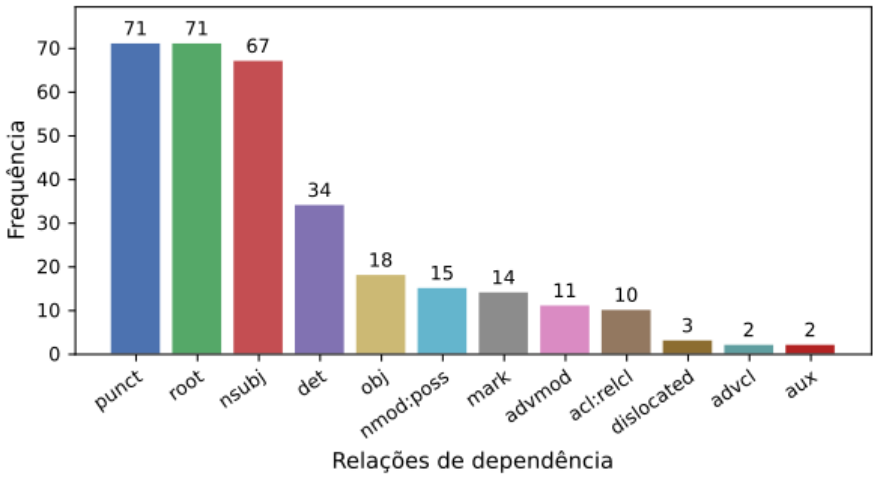


Figura 4. Frequências das relações de dependência.

5. Validação e avaliação

Para assegurar o funcionamento de aplicações baseadas na coleção UD, todos os *treebanks* são submetidos periodicamente a um procedimento de validação por meio dos *scripts* `check_files.pl` e `validate.py`, cujos resultados são consolidados num relatório on-line. A validação leva em conta aspectos formais como a estrutura de arquivos do repositório do *treebank* e a sanidade do formato CoNLL-U assim como a obediência a diferentes diretrizes de anotação. Nesse relatório, os *treebanks* são classificados em diferentes categorias de validação, que refletem diferentes graus de adesão aos requisitos do modelo (Tabela 1). A categoria CURRENT VALID engloba *treebanks* completamente válidos. O termo ERROR indica erros de validação, enquanto WARNING refere-se a desvios menos graves. Comumente, questões tratadas como desvios numa versão de UD tornam-se erros numa versão seguinte. Há um período de tolerância de cerca de quatro anos para *treebanks* com erros, após o qual esses corpora são excluídos da coleção, passando a integrar a categoria RETIRED, como foi o caso do UD_Mbya_Guarani-Dooley na *release* 2.18.

Tabela 1. Validação da coleção UD em 26/05/2026.

Categoria	*Treebanks*	Línguas
CURRENT VALID WARNING	98	59
CURRENT VALID	106	77
SAPLING EMPTY	56	48
CURRENT ERROR LEGACY	138	94
SAPLING ERROR	35	35
CURRENT ERROR BACKUP	2	2
RETIRED ERROR	10	9
CURRENT ERROR NEGLECTED	9	6
SAPLING VALID	1	1
TOTAL VALID/BACKUP/LEGACY/NEGLECTED	**354**	**194**

Conforme a Figura 5, apenas três *treebanks* de línguas ameríndias alcançam o status máximo de validação, entre os quais o do kadiwéu. Um total de seis se encontra na categoria imediatamente inferior, enquanto os restantes 16 se incluem na categoria CURRENT ERROR LEGACY por apresentarem erros de validação.

Além da validação, os *treebanks* da coleção UD passam, a cada *release*, por uma avaliação automática por meio do *script* `evaluate_treebank.pl`, que calcula, para cada *treebank*, um escore de 0 a 1, com base na qual atribui a cada corpus um número de 0 a 5 estrelas, com intervalos de 0,5. O cálculo do escore leva em conta os fatores e pesos das Tabelas 2 e 3, que refletem o tamanho, a riqueza e a qualidade da anotação, bem como a diversidade de gêneros textuais e a divisão do *treebank* em um conjunto de teste (TEST), um de desenvolvimento (DEV) e um de treino (TRAIN).

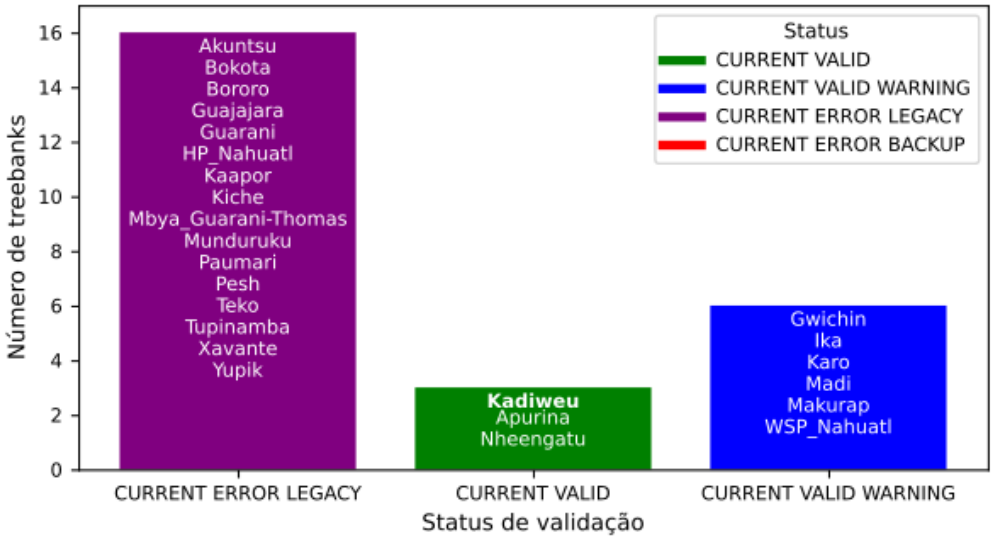


Figura 5. Validação dos *treebanks* de línguas ameríndias em 26/05/2026.

Tabela 2. Fatores multiplicativos dos diferentes tipos de anotação.

Tipo de anotação	Lemas	UPOS	*Features*	Relações
`manual native`	1.0	1.0	1.0	1.0
`converted with corrections`	0.9	0.9	0.9	0.9
`converted from manual`	0.8	0.8	0.8	0.8
`automatic with corrections`	0.5	---	0.5	---
outros valores / não especificado	0.4	0.1	0.4	0.1

O UD_Kadiweu-Unicamp alcança os escores máximos nos *features* e lemas. Zera, porém, o escore de tamanho e obtém o valor mínimo para divisão. Os demais escores baixos refletem a concentração em sentenças de um único gênero textual sem variedade morfossintática suficiente para abarcar mais que 64,71% e 32,43%, respectivamente, das etiquetas de classes de palavra e relações gramaticais de UD. O UD_Kadiweu-Unicamp também atinge a pontuação máxima de 1 nos critérios validação e disponibilidade do texto das sentenças, utilizados como fatores multiplicativos do escore total de 0,3371. Esse escore total resulta em 1,5 estrela, que lhe assegura a quarta posição entre os

treebanks de línguas ameríndias de UD, superando *treebanks* consolidados do português, como o UD_Portuguese-Bosque.

Tabela 3. Cálculo do escore total do UD_Kadiweu-Unicamp.

Fator	Peso	Escore do fator	Contribuição
Features	0,1111	1,0000	0,1111
Genres	0,1111	0,0556	0,0062
Lemmas	0,1111	1,0000	0,1111
Size	0,3704	0,0000	0,0000
Split	0,0741	0,0100	0,0007
Universal POS Tags	0,1111	0,6471	0,0719
Universal relations	0,1111	0,3243	0,0360

6. Considerações finais

Este artigo tratou do UD_Kadiweu-Unicamp, o primeiro *treebank* de UD para uma língua Guaicuru. Com isso, contribuímos para aumentar a cobertura tipológica dessa coleção multilíngue, que, embora inclua 25 línguas indígenas das Américas, ainda não faz jus à enorme diversidade linguística desse continente. Delineamos a composição e construção do *treebank* a partir da conversão automática das anotações de um *corpus* de estrutura sintagmática, seguida de revisão humana, que o inseriu na categoria máxima de validação de UD. Apesar de restrito a 71 sentenças e 306 *tokens* na *release* 2.18 de UD, os resultados da avaliação automática posicionam o UD_Kadiweu-Unicamp entre os mais bem avaliados das línguas ameríndias. Visando à *release* 2.19, o *treebank* vem sendo continuamente expandido e revisado, ampliando os inventários de classes de palavras, *features* e relações de dependência, acompanhando a evolução do *corpus* da PTB de que deriva. Paralelamente, o conversor tem sido aperfeiçoado. Entre os desafios que se apresentam está a modelação em UD dos classificadores numerais, ainda não contemplados no inventário de *features* do modelo.

Questões éticas

Os consultores indígenas integram o projeto DACILAT, pelo qual foram remunerados. Leonel F. de Alencar (LF) desenvolveu interativamente o conversor com auxílio do ChatGPT-5. LF implementou a infraestrutura do UD_Kadiweu-Unicamp nos repositórios de UD no GitHub, testou extensivamente o conversor, revisou manualmente os resultados e os submeteu aos *scripts* de validação e avaliação de UD. Filomena Sandalo (FS) e Charlotte Galves (CG) elaboraram o esquema de anotação do kadiwéu da PTB e as regras do respectivo *parser*. FS realizou a análise morfológica em conjunto com alunos, tanto indígenas quanto não indígenas. A análise dos dados e o texto do artigo, redigido inicialmente por LF e revisado por FS e CG, refletem inúmeras discussões entre os autores.

Referências

ALENCAR, L. F. de. Uma gramática computacional de um fragmento do nheengatu. *Revista de Estudos da Linguagem*, v. 29, n. 3, p. 1717–1777, 2021.

ALENCAR, L. F. de; SILVA, H. L. B.; GURGEL, J. L.; ALEXANDRE, D. M. Per aspera ad astra: improving the automatic evaluation of a Universal Dependencies treebank for

a low-resource language. *Revista de Estudos da Linguagem*, v. 34, n. 1, p. 8–50, 2026.

ARNARDÓTTIR, Þ. et al. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In: MARNEFFE, M.-C. de et al. (Ed.). *Proceedings of the Fourth Workshop on Universal Dependencies* (UDW 2020). Barcelona, Spain (Online): Association for Computational Linguistics, 2020. p. 16–25. Disponível em: <https://aclanthology.org/2020.udw-1.3/>.

BUßMANN, H. (Ed.). *Lexikon der Sprachwissenschaft*. 3. ed. Stuttgart: Alfred Kröner Verlag, 2002.

CIUCCI, L.; KILARSKI, M. Investigating the Indigenous Languages of the Americas: History and Prospects. In: BRADLEY, D. et al. (ed.). *Contemporary Linguistics*: Integrating Languages, Communities, and Technologies. [S. l.]: Brill, 2025. p. 379–392.

FARIA, P.; GALVES, C.; MAGRO, C. Syntactic annotation for Portuguese corpora: standards, parsers, and search interfaces. *Language Resources and Evaluation*, v. 58, n. 1, p. 301–346, 2024.

GALVES, C.; SANDALO, F.; SENA, T. A. de; VERONESI, L. Annotating a polysynthetic language: from Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, v. 59, n. 3, p. 631–648, set./dez. 2017.

GIBERT, O. D. et al. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In: MAGER, M. et al. (Ed.). *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas* (AmericasNLP). Albuquerque, New Mexico: Association for Computational Linguistics, 2025. p. 134–152. Disponível em: <https://aclanthology.org/2025.americasnlp-1.16/>.

JOHANSSON, R.; NUGUES, P. Extended constituent-to-dependency conversion for English. In: NIVRE, J. et al. (Ed.). *Proceedings of the 16th Nordic Conference of Computational Linguistics* (NODALIDA 2007). Tartu, Estonia: University of Tartu, Estonia, 2007. p. 105–112. Disponível em: <https://aclanthology.org/W07-2416/>.

MARNEFFE, M.-C. de et al. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255–308, 6 2021.

NIVRE, J. et al. Universal Dependencies 1.0. 2015. LINDAT/CLARIAH CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Disponível em: <http://hdl.handle.net/11234/1-1464>.

PIRES, V. *Palavras kadiwéu do mundo ancestral e do mundo novo*: palavras novas, palavras antigas, palavras humildes e palavras honorificadas. 44 p. Dissertação (Mestrado em Linguística) — Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2022.

RANTA, A. *Grammatical Framework*: Programming with Multilingual Grammars. Stanford, California: CSLI, 2011.

RODRIGUES, A. D. *Línguas brasileiras*: Para o conhecimento das línguas indígenas. São Paulo: Loyola, 1986.

SANDALO, F. et al. *Corpus Kadiwéu*. In: VERONESI, L. H. L.; GALVES, C. (Ed.). The Tycho Brahe Platform. Universidade Estadual de Campinas, 2024. Disponível em:

<https://www.tycho.iel.unicamp.br/home/corpora>.

SANDALO, F. et al. *Corpus Kadiwéu – Gramática Pedagógica*. In: VERONESI, L. H. L.; GALVES, C. (Ed.). The Tycho Brahe Platform. Universidade Estadual de Campinas, 2024. Disponível em: <https://www.tycho.iel.unicamp.br/home/corpora>.

SANDALO, M. F. A *Grammar of Kadiwéu*. 301 p. Tese (PhD Dissertation) — University of Pittsburgh, Pittsburgh, PA, 1995.

SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do Brasil: O problema de Anchieta. *Cadernos de Estudos Linguísticos*, v. 65, n. 00, 2023.

SANDALO, M. F. S.; GALVES, C. A brief description and analysis of tense, finiteness, and verbal embedding strategies in Kadiwéu. *Cadernos de Estudos Linguísticos*, Campinas, SP, v. 67, n. 00, p. e025021, 2025.

SANDALO, M. F. S.; MICHELIOUDAKIS, D. Classifiers and plurality: evidence from a deictic classifier language. *The Baltic International Yearbook of Cognition, Logic and Communication*, v. 11, n. 1, p. 1–40, 2016.

STORTO, L. R. *Línguas indígenas*: tradição, universais e diversidade. Campinas, SP: Mercado de Letras, 2019.

TAYLOR, A.; MARCUS, M.; SANTORINI, B. The Penn Treebank: An overview. In: ABEILLÉ, A. (Ed.). *Treebanks*: Building and Using Parsed Corpora. Dordrecht: Springer Netherlands, 2003. p. 5–22.

THOMAS, G. Universal Dependencies for Mbyá Guaraní. In: *Proceedings of the Third Workshop on Universal Dependencies* (UDW, SyntaxFest 2019). Paris, France: Association for Computational Linguistics, 2019. p. 70–77. Disponível em: <https://aclanthology.org/W19-8008>.

TYERS, F.; MISHCHENKOVA, K. Dependency annotation of noun incorporation in polysynthetic languages. In: MARNEFFE, M.-C. de et al. (Ed.). *Proceedings of the Fourth Workshop on Universal Dependencies* (UDW 2020). Barcelona, Spain (Online): Association for Computational Linguistics, 2020. p. 195–204. Disponível em: <https://aclanthology.org/2020.udw-1.22/>.

VERONESI, L. H. L. *A Plataforma Tycho Brahe*: um sistema para corpora sintaticamente anotados. 2026. 211 f. Tese (Doutorado em Linguística) - Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2026.

ZEMAN, D. *Cross-Language Harmonization of Linguistic Resources*. Prague: Institute of Formal and Applied Linguistics (ÚFAL), 2023. Habilitation thesis.

ZEMAN, D. et al. Universal Dependencies 2.18. 2026. LINDAT/CLARIAH CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Disponível em: <http://hdl.handle.net/11234/1-6149>.

Córpus Alienista: uma obra machadiana anotada com Universal Dependencies

Magali S. Duran[1], Ursula Puello Sydio[2], Lucelene Lopes[1], Thiago A. S. Pardo[1]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC), Instituto de Ciências Matemáticas e de Computação (ICMC), [2] Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH), Universidade de São Paulo (USP)

magali.duran@gmail.com, ursula.sydio@usp.br, lucelene@gemail.com, taspardo@icmc.usp.br

Abstract. *This paper describes the creation of the "Alienista", a publicly available literary corpus based on the book "O Alienista" by Machado de Assis. The corpus was automatically annotated using the Universal Dependencies framework and manually revised. The process prompted reflections on corpus annotation within the literary genre and resulted in 801 sentences annotated with lemmas, morphological features, morphosyntactic tags, and syntactic dependencies for each of its 20,728 tokens.*

Resumo. *Relata-se a criação do córpus literário "Alienista", publicamente disponível, com base na obra "O Alienista", de Machado de Assis. O córpus foi anotado automaticamente segundo a abordagem da iniciativa Universal Dependencies e revisado manualmente. O processo gerou reflexões acerca da anotação de córpus do gênero literário, e resultou em 801 sentenças anotadas com lemas, atributos morfológicos, etiquetas morfossintáticas e dependências sintáticas, para cada um de seus 20.728 tokens.*

1. Introdução

A Linguística de Córpus é uma grande provedora de recursos para o desenvolvimento do Processamento de Línguas Naturais (PLN). A ciência de compilar um córpus de textos que reflitam características de um autor, de um gênero, de um domínio ou de uma época fornece insumos fundamentais para que os métodos automáticos aprendam a distinguir tais características. Quando um córpus é enriquecido com anotações (análises de diferentes naturezas expressas por meio de etiquetas associadas a segmentos de texto), sua riqueza informacional aumenta significativamente para fins de PLN. A anotação de córpus, além de propiciar o aprendizado automático das análises humanas sobre o texto, beneficia os estudos linguísticos, pois permite que as etiquetas atribuídas sejam utilizadas como parâmetro de buscas.

Dentre as formas de anotação de córpus, uma que vem se destacando internacionalmente nos últimos quinze anos é a abordagem *Universal Dependencies* (UD) (Nivre et al., 2020; de Marneffe et al., 2021). Essa anotação possui um formato consensual que agrega várias análises: morfológica, morfossintática e de sintaxe de dependências. Uma vez que todas as línguas utilizam o mesmo formato e os mesmos conjuntos de etiquetas, os córpus anotados constituem um insumo valioso para estudos contrastivos entre línguas. Atualmente, quase 200 línguas já possuem córpus anotados segundo o formato da UD. No momento da escrita deste artigo, o português é representado na

plataforma UD[26] por sete córpus, sendo três exclusivamente de português brasileiro: o Petrogold (Souza et al., 2021), do gênero acadêmico; o Porttinari (Duran et al., 2023), do gênero jornalístico; e o DANTEStocks (Di Felippo et al., 2024), do gênero UGC (*User Generated Content*, ou seja, conteúdo gerado por usuários nas redes sociais). Além desses três, há o PortJur (Lopes et al., 2025), do gênero jurídico, disponibilizado no site do projeto POeTiSA[27].

Este artigo apresenta o córpus Alienista, constituído a partir do conto O Alienista, de Machado de Assis. Com essa iniciativa, pretende-se suprir duas lacunas: fornecer um córpus literário de autor brasileiro anotado com UD e fornecer uma anotação rica que permita conhecer características do léxico e da sintaxe machadiana, importantes, por exemplo, para estudos de atribuição de autoria.

O restante do artigo está organizado em cinco seções. A Seção 2 discorre sobre a abordagem de anotação utilizada (UD) e os progressos feitos em português dentro dessa abordagem. A Seção 3 discute a metodologia adotada para empreender a tarefa de construção do córpus. Na Seção 4 são comentados os resultados obtidos e, na Seção 5, apresentam-se as conclusões e ideias para trabalhos futuros.

2. Trabalhos Relacionados

A abordagem *Universal Dependencies* (Nivre et al., 2020; de Marneffe et al., 2021) é uma iniciativa multinacional com vistas a estabelecer diretrizes independentes de língua para anotar córpus. Como o próprio nome diz, a abordagem é baseada em sintaxe de dependências, mais precisamente no trabalho de Tesnière (2015). O modelo avançou significativamente nos últimos 20 anos, alcançando grande refinamento de conjuntos de etiquetas aplicáveis e apoiando a realização de diversas pesquisas multilíngues.

No que diz respeito ao português, são observados avanços também, especialmente no Brasil. O primeiro córpus de português anotado com UD foi o PUD (*Parallel Universal Dependencies*), um conjunto de 1.000 sentenças de um córpus paralelo que envolveu 23 línguas, criado para uma competição internacional de PLN em 2017 (Hajič & Zeman, 2017), mas não foi revisado por linguistas. A seguir, foi lançada a versão UD do córpus Bosque (Rademaker et al., 2017), resultado do mapeamento, para a UD, da anotação original do córpus Bosque (Afonso et al., 2002), a qual havia sido feita pelo parser Palavras (Bick, 2000) e revisada por linguistas. Esse foi o primeiro córpus de tamanho significativo para treinamento de parsers do português.

Em 2021, Souza et al. (2021) construíram um córpus do gênero acadêmico em português brasileiro, no domínio do petróleo, utilizando um parser treinado no córpus Bosque-UD e Pardo et al. (2021) anunciaram a construção de um córpus UD multigênero para o português brasileiro. No ano seguinte, foi lançada a versão UD do córpus de português europeu Cintil (Branco et al., 2022), majoritariamente de textos jornalísticos. Já em 2023, foi lançado o córpus Porttinari (Duran et al., 2023), um córpus jornalístico exclusivamente brasileiro anotado com UD, pela primeira vez contando com manuais de anotação para guiar a revisão feita pelos anotadores humanos (Duran et al., 2022). Esse córpus foi a base do treinamento do Portparser (Lopes & Pardo, 2024), hoje o parser com

[26] https://universaldependencies.org/
[27] https://sites.google.com/icmc.usp.br/poetisa

melhor desempenho em língua portuguesa.

Outros gêneros se somaram ao conjunto de córpus de português brasileiro em UD: o DANTEStocks (Di Felippo et al., 2024), de tweets e, portanto, do gênero UGC (*User Generated Content*), e o PortJur (Lopes et al., 2025), do gênero jurídico.

Esse cenário mostra que há vários gêneros ainda não anotados com UD em língua portuguesa e que o presente trabalho preenche uma lacuna ao anotar um córpus literário baseado na obra de um grande escritor brasileiro.

3. Metodologia

O córpus Alienista teve como texto base o conto O Alienista, de Machado de Assis. Segundo a Wikipedia, a obra foi publicada pela primeira vez em oito capítulos quinzenais, na revista carioca A Estação, de outubro de 1881 a março de 1882. Em 1882, passou a integrar o volume de contos machadianos "Papéis Avulsos". Como a obra está em domínio público, a versão de O Alienista utilizada para constituir o córpus encontra-se disponível em vários sites da web[28]. Uma vez que não houve dificuldade para a obtenção do texto, relatam-se, a seguir, as demais etapas de constituição do córpus.

3.1. Segmentação em sentenças

A primeira tarefa para constituir o córpus de anotação no formato UD é a segmentação do texto em sentenças, pois a UD utiliza a sentença como unidade de anotação. Cada sentença dá origem a um arquivo num formato chamado CoNLL-U, com uma linha para cada token e 10 colunas de informações (ver Duran et al., 2022, para maiores detalhes).

A segmentação de sentenças automática normalmente considera a ocorrência de uma pontuação sucedida de inicial maiúscula. Nas narrativas, porém, essa regra nem sempre funciona, pois, em segmentos que combinam narração com discurso direto, pode haver relação sintática entre os segmentos. A sentença a seguir, por exemplo, apresenta um período em que dois segmentos terminam com pontos de exclamação:

> - Não é patuscada, não, senhora; eles estão gritando: "Morra o Dr. Bacamarte! O tirano!" - dizia o moleque assustado. (SENT0425)

Se as frases fossem separadas automaticamente, poderiam ser geradas quatro sentenças para anotação, dependendo do parser utilizado:

1. - Não é patuscada, não, senhora; eles estão gritando:
2. "Morra o Dr. Bacamarte!
3. O tirano!"
4. - dizia o moleque assustado.

Com essa divisão, contudo, perderia-se a dependência sintática entre as partes, já que o verbo "dizia" tem como objeto direto[29] ("ccomp" ou oração subordinada objetiva direta) o trecho "- Não é patuscada, não, senhora; eles estão gritando: "Morra o Dr. Bacamarte! O tirano!"" e o verbo "gritando" tem como objeto direto "Morra o Dr. Bacamarte! O tirano!". Por essa razão, decidiu-se realizar a segmentação manualmente pois, assim, poderia-se aprender melhor essa tarefa no gênero literário.

[28] https://machadodeassis.net/texto/@id/27353

[29] O objeto direto na UD é anotado com a etiqueta "obj", porém, quando se trata de uma oração subordinada objetiva direta, a etiqueta a ser utilizada é "ccomp" (Duran et al., 2022).

Terminada a tarefa, obteve-se um total de 801 sentenças, as quais foram identificadas com um prefixo e um número sequencial: M_ASSIS_ALIENISTA_SENTNNNN (nesse formato, NNNN é o espaço reservado para a numeração sequencial).

3.2. Pré-anotação automática

O córpus foi tokenizado e pré-anotado automaticamente com a versão mais recente do Portparser (Lopes & Pardo, 2024). Esse parser foi escolhido pois é o estado da arte para o português (com valores de acurácia acima de 96%) e está disponível na web[30].

3.3. Revisão manual da anotação

A fim de que o córpus pudesse ser revisado e editado manualmente, criou-se um projeto no editor Arborator-Grew[31] (Guibon et al., 2020) e fez-se o *upload* do córpus pré-anotado. A ferramenta tem sido utilizada por outros pesquisadores e oferece um ambiente amigável, com recursos de anotação visuais e intuitivos. Além de permitir a correção de erros detectados na atribuição de etiquetas e de dependência sintática, o Arborator-Grew permite a edição de tokens (divisão, junção ou correção da forma).

A anotação automática foi revisada e editada por dois linguistas durante sessões virtuais, num esforço sinérgico a fim de alcançar a melhor interpretação das sentenças. Utilizaram-se as diretrizes de anotação do projeto POeTiSA (Duran et al., 2022).

3.4. Validação do córpus anotado

Após a revisão manual, foram utilizados: a ferramenta Verifica-UD[32] (Lopes et al., 2023) e o sistema de validação do próprio projeto UD, que apontaram inconsistências de anotação, as quais foram manualmente resolvidas.

4. Resultados

A seguir são fornecidos dados do córpus Alienista e comentados, em tópicos separados, aprendizados relevantes que poderão ser úteis para os leitores.

4.1. Dados do córpus

O córpus contém 801 sentenças e 20.728 tokens, ou, se forem desconsiderados os sinais de pontuação, 17.639 tokens (numa média de 25,88 tokens por sentença). Há muita variabilidade no tamanho das sentenças, sendo a menor de 2 tokens e a maior de 263 tokens (cerca de 10% das sentenças do córpus têm até 5 tokens e cerca de 50% têm mais de 20 tokens).

4.2. Revisão da tokenização

A revisão detectou poucos erros no processo de tokenização automática. Um deles foi a descontração indevida da palavra "desse" em dois tokens "de" + "esse" quando se tratava da forma da terceira pessoa do singular do verbo "dar" no subjuntivo (a divisão em dois tokens só é devida quando se trata da preposição "de" combinada com o pronome "esse"). Outro erro de tokenização do parser foi não fazer a descontração da forma "lho" em "lhe"

[30] https://huggingface.co/spaces/NILC-ICMC-USP/Portparser.v2
[31] https://arborator.grew.fr/
[32] https://huggingface.co/spaces/NILC-ICMC-USP/VerificaUD

+ “o”, provavelmente porque essa forma de contratação já não é usual no português contemporâneo (que serviu para treinar o parser utilizado).

Por fim, houve um problema que exigiu uma decisão de tokenização: a abreviação “etc.”, quando ocorre em final de sentença, tem um ponto final amalgamado ao ponto da abreviação. Como o ponto da abreviação integra o token, criou-se um novo ponto para receber a anotação de ponto final da sentença (vide Figura 1).

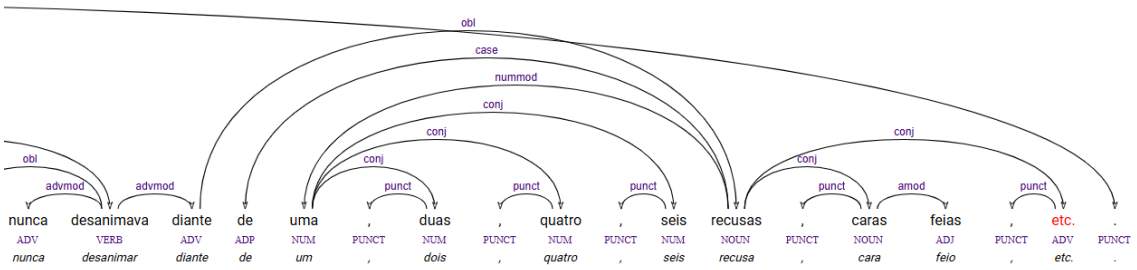


Figura 1. Excerto da sentença SENT0371.

4.3. Léxico

A riqueza lexical do córpus pode ser observada pelo baixo valor da *type/token ratio*: 23,21% (desconsiderando-se 3.089 tokens de pontuação, mas não desconsiderando *stopwords*), ou seja, 17.639 tokens para 4.094 *types*. Esses *types* reúnem-se em 2.797 diferentes lemas (alguns lemas, como os de verbos, podem apresentar diversos *types*, pois cada conjugação é um *type* diferente de um mesmo lema).

Não foram muitas as dificuldades do parser em relação ao léxico. Destaca-se o fato de não ter conseguido reconhecer “dous” como variante de “dois” e, por isso, não ter classificado o token com a etiqueta NUM. Para processar textos históricos, é importante que o léxico usado pelo parser preveja as alternâncias dos ditongos ou/oi: dous/dois, louro/loiro, dourado/doirado, doudo/doido, cousa/coisa, etc.

4.4. Etiquetas morfossintáticas

Das 17 etiquetas morfossintáticas da UD, as ocorrências (num córpus de 20.728 tokens) foram: NOUN (3.553), DET (3.154), PUNCT (3.089), ADP (2.493), VERB (2.269), PRON (1.365), ADJ (1.084), ADV (1.006), PROPN (802), CCONJ (745), SCONJ (408), AUX (512), NUM (216), X (22), INTJ (8) e SYM (2).

No gênero literário, quando há narrativa e inserção de trechos com discurso direto, observa-se uma frequência de tokens de pontuação muito maior do que aquela relatada na anotação de outros gêneros. O córpus contém doze *types* de pontuação que se repetem diversas vezes, totalizando 3.089 tokens, entre os quais: 1.500 vírgulas, 701 pontos finais, 311 travessões, 265 ponto-e-vírgulas, 77 pontos de exclamação, 76 aspas duplas, 60 dois-pontos, 43 pontos de interrogação, 34 reticências, 9 abre parênteses, 9 fecha parênteses, 2 aspas simples e 2 reticências com apenas dois pontos.

Também chama a atenção o fato de que os determinantes (DET) superam o número de preposições (ADP), o que não foi relatado em outros gêneros já anotados com UD (jornalístico, jurídico, acadêmico e UGC). Outro destaque é a presença muito mais frequente de pronomes (PRON). Tais características podem ser do gênero literário, ou do

português histórico, ou ainda do autor (só estudos comparativos com outros córpus anotados poderiam esclarecer isso).

4.5. A anotação sintática

Embora no geral o parser tenha tido um desempenho muito bom em todas as tarefas, foi na sintaxe que se concentrou a maior parte das edições de anotação. Alguns erros foram fáceis de corrigir, pois ficaram claros para os anotadores a partir da análise da lógica do texto. Um exemplo desse tipo é ilustrado a seguir. A Figura 2 mostra a anotação automática e a Figura 3 mostra a anotação corrigida (correções em vermelho) de uma sentença do córpus.

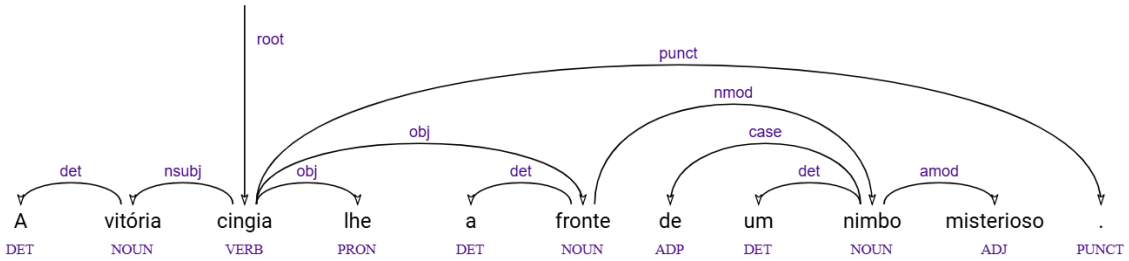


Figura 2 - Anotação automática da SENT0497.

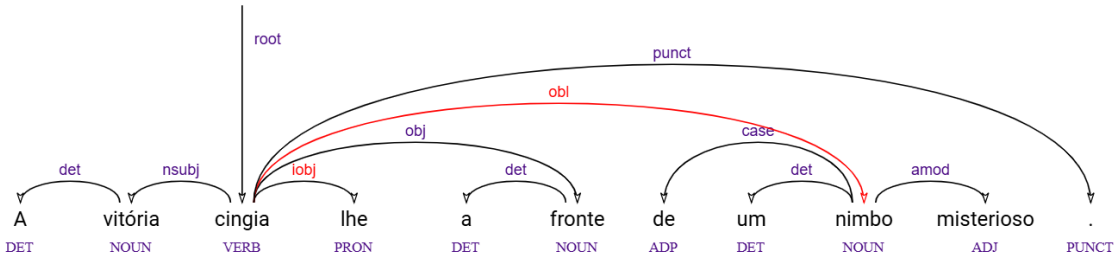


Figura 3 - Anotação corrigida da SENT0497.

O parser havia atribuído dois objetos diretos (obj), um deles ao pronome "lhe", que obviamente só pode ser objeto indireto (iobj). Além disso, "de um nimbo misterioso" é um adjunto adverbial (obl) de "cingia" e não um adjunto adnominal (nmod) de "fronte".

Houve, porém, casos que desafiaram os anotadores. Um deles foi a anotação da partícula "se". O "se" pode receber quatro possíveis etiquetas de dependência sintática: obj (objeto direto), iobj (objeto indireto), expl (partícula integrante do verbo) ou expl:impers (que serve tanto para "se" partícula apassivadora quanto para "se" índice de indeterminação do sujeito). A sentença a seguir ilustra essa dificuldade.

> Mas bem depressa a ilusão se desfez. (SENT0501)

No português, o verbo "desfazer" pode ter uso transitivo direto (alguém desfaz alguma coisa) ou uso intransitivo (sinônimo de "ruir", "desintegrar-se"). Na referida sentença, interpretou-se o verbo como pronominal (alternância incoativa do uso transitivo do verbo), atribuindo, portanto, a etiqueta "expl" para o "se"[33].

[33] O "se", contudo, poderia ser um "expl:impers" em uma construção como "Desfez-se a confusão", pois aí a interpretação seria "A confusão foi desfeita", ou seja, "A confusão" seria sujeito de uma passiva sintética, uma das formas de ocultar o agente de um verbo transitivo.

Curiosamente, foram encontradas muitas construções em que o pronome átono objeto indireto foi utilizado em lugar de um possessivo ou de um genitivo, respectivamente a primeira e a segunda paráfrases de cada excerto das sentenças a seguir:

SENT0050: "classificar-**lhe** os casos" = "classificar **seus** casos" = "classificar os casos **dela**"

SENT0103: "decepar-**lhe** as mãos" = "decepar **suas** mãos" ou "decepar as mãos **dela**"

SENT0107: "**lhe** descerrou os lábios" = "descerrou **seus** lábios" ou "descerrou os lábios **dele**"

SENT0208: "batiam-**lhe** no ombro" = "batiam em **seu** ombro" ou "batiam no ombro **dele**"

SENT0212: "perdoou-**lhe** a dívida" = "perdoou **sua** dívida" ou "perdoou a dívida **dele**"

SENT0218: "bater-**lhe** à porta" = "bater à **sua** porta" ou "bater à porta **dele**"

Isso mostra a preferência do autor pela primeira dessas três formas sintáticas de expressar um mesmo conteúdo semântico (que alternam os casos dativo, possessivo e genitivo). O uso possessivo é apontado por Moura Neves (2000, p. 455): "As formas de objeto indireto também podem ser empregadas para fazer referências do tipo possessivo". É interessante observar que, sintaticamente, tais pronomes são classificados como objeto indireto, embora os verbos a que se unem não sejam atualmente considerados transitivos indiretos nem bitransitivos. Embora tal uso ainda exista no português contemporâneo, é muito mais raro. Na obra, encontrou-se também um objeto indireto não pronominal codificando esse tipo de informação, o que pode reforçar a hipótese de que constitui preferência do autor (ou característica do português histórico):

> Crispim Soares saiu de casa espumando de cólera, e declarando às pessoas a quem encontrava que ia arrancar as orelhas **ao tirano**. (SENT0680)

Casos mais complexos exigiram decisões de anotação. A prosa machadiana apresenta muitas coordenações assindéticas e muitas justaposições. Na UD, existe uma etiqueta para anotar orações coordenadas sindéticas e assindéticas (conj) e outra para anotar orações justapostas (parataxis). As assindéticas se assemelham às justapostas, pois ambas carecem de conjunção. Para diferenciá-las, foi necessário observar se ambas as orações tinham o mesmo sujeito. Em caso afirmativo, a oração foi tratada como oração coordenada assindética e anotada como "conj". Em caso negativo, ou seja, se o sujeito das duas orações não era o mesmo, as orações foram tratadas como justapostas e unidas com a relação de dependência "parataxis". A Figura 4 ilustra ambos os casos: o sujeito de "deu" e "fez" é o mesmo, embora elíptico e por isso as duas orações foram unidas pela relação "conj"; já o sujeito de "petrificou" é "terror" e, portanto, diferente do sujeito de "deu" e "fez", por isso a oração foi unida pela relação "parataxis" ao primeiro verbo período, "deu".

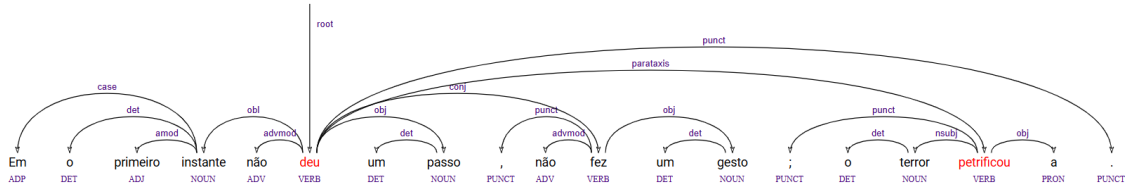


Figura 4. Distinção de oração coordenada assindética e oração justaposta (SENT0432).

Outra decisão de anotação, porém específica do léxico, foi tomada na anotação ilustrada

pela Figura 5. Nesse caso, a expressão "vá que" foi anotada como expressão fixa, com função semelhante à de um advérbio que expressa probabilidade (como "talvez") e por isso ligada ao verbo principal da sentença pela relação de dependência "adv".

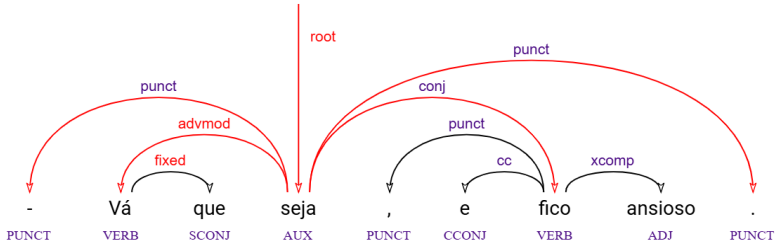


Figura 5. Decisão de anotação de expressão fixa (SENT0069).

5. Conclusão

A anotação de córpus em português brasileiro segundo a abordagem UD já acumula alguma experiência e, com isso, há um bom aparato disponível para novas iniciativas, como a que foi aqui relatada.

O córpus Alienista, embora não seja extenso, é constituído por uma obra literária completa, permitindo a exploração de análises que considerem a sequência de sentenças e não apenas sentenças isoladas. Pelo fato de o córpus ter sido manualmente segmentado em sentenças, pôde-se conhecer melhor essa tarefa no gênero literário e, em trabalhos futuros, pretende-se usar segmentadores automáticos e fazer revisão humana apenas nos trechos que contenham travessão e/ou dois pontos, marcas de combinação de discurso indireto e direto num mesmo enunciado.

A disponibilização de mais um córpus UD em português brasileiro enriquece o conjunto de recursos para estudos da língua. Linguistas interessados poderão usar as etiquetas de anotação como parâmetros para explorar o córpus, enquanto desenvolvedores de recursos de PLN poderão utilizar o córpus para treinar modelos classificadores sensíveis a contextos literários e históricos. O córpus está publicamente disponível no portal web do projeto POeTiSA, em https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Também se agradece ao apoio do INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação - https://tildiar.dcc.ufmg.br/) (processo CNPq/SECTICS/CAPES/FAPs #408490/2024-1). Ursula Puello Sydio agradece o financiamento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (processo CAPES #88887.920964/2023-00).

Referências

Afonso, S.; Bick, E.; Haber, E.; Santos, D. (2002). Floresta sintá(c)tica: a treebank for Portuguese. In the Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), pp. 1698-1703.

Bick, E. (2000). The parsing system Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Aarhus University Press.

Branco, A.; Silva, J.R.; Gomes, L.; Rodrigues, J.A. (2022). Universal Grammatical Dependencies for Portuguese with CINTIL Data, LX Processing and CLARIN support. In the Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pp. 5617-5626.

de Marneffe, M.C.; Manning, C.D.; Nivre, J.; Zeman, D. (2021). Universal Dependencies. Computational Linguistics, Vol. 47, N. 2, pp. 255-308.

Di Felippo, A.; Nunes, M.G.V.; Barbosa, B. (2024). A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach. In the Proceedings of the 15th Symposium in Information and Human Language Technology (STIL), pp. 192-201.

Duran, M.S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In the Proceedings of the 14th Symposium in Information and Human Language Technology (STIL), pp. 115-124.

Duran, M.S.; Nunes, M.G.V.; Lopes, L.; Pardo, T.A.S. (2022). Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. Domínios de Lingu@gem, Vol. 16, N. 4, pp. 1608-1643.

Guibon, G.; Courtin, M.; Gerdes, K.; Guillaume, B. (2020). When Collaborative Treebank Curation Meets Graph Grammars. In the Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC), pp. 5291-5300.

Hajič, J.; Zeman, D. (2017). Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.

Lopes, L.; Nunes, M.G.V.; Duran, M.S.; Pardo, T.A.S. (2025). A sintaxe no tribunal: apresentando e explorando um corpus jurídico em português anotado sintaticamente segundo o modelo Universal Dependencies. In the Proceedings of the XVI Symposium in Information and Human Language Technology (STIL), pp. 220-232.

Lopes, L.; Pardo, T.A.S. (2024). Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In the Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR), pp. 401-410.

Lopes, L.; Duran, M.S.; Pardo, T.A.S. (2023). Verifica-UD: a Verifier for Universal Dependencies Annotation for Portuguese. In the Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival (UDFest-BR), pp. 451-460.

Moura-Neves, M.H. (2000). Gramática de usos do português. São Paulo: Editora UNESP.

Nivre, J.; de Marneffe, M.C.; Ginter, F.; Hajič, J.; Manning, C.D.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D.. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In the Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC), pp. 4034-4043.

Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021). Porttinari - a large multi-genre treebank for Brazilian Portuguese. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pp. 1-10.

Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; Paiva, V. (2017). Universal Dependencies for Portuguese. In the Proceedings of the Fourth International Conference on Dependency Linguistics, pp. 197-206.

Souza, E.; Silveira, A.; Cavalcanti, T.; Castro, M.; Freitas, C. (2021). PetroGold – Corpus padrão ouro para o domínio do petróleo. In the Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology (STIL), pp. 29-38.

Tesnière, L. (2015). Elements of Structural Syntax. Tradução de OSBORNE, Timothy; Kahane, Sylvain. Amsterdam: John Benjamins.

TTWtools: uma ferramenta aberta e transversal para etiquetagem e análise em linguística de corpus

Luciana Aguiar de Oliveira[1,2], José Henrique Padovani[3]

[1] Departamento de Letras e Linguística - Universidade do Estado de Minas Gerais, [2] Programa de Pós-Graduação em Estudos Linguísticos – Universidade Federal de Minas, [3] ECrIS/PPGMUS/Escola de Música – Universidade Federal de Minas Gerais

luciana.oliveira@uemg.br, jhp@ufmg.br

Abstract. *This paper presents TTWtools, an open-source application that integrates morphosyntactic tagging, morphosyntactic feature counting per text, and statistical analysis into a single environment, available both as a Python package and as an executable graphical application. Offering access to multiple taggers and interactive data visualization, the tool is particularly suited for workflows commonly associated with register variation research. Within an open and reproducible framework, TTWtools aims to democratize access to advanced corpus analysis for researchers and educators.*

Resumo. *Este artigo apresenta o TTWtools, uma aplicação de código aberto que integra etiquetagem morfossintática, contagem de features morfossintáticas por texto e análise estatística em um único ambiente, disponível tanto como pacote Python quanto como aplicativo gráfico executável. Oferecendo acesso a múltiplos etiquetadores e visualização interativa dos dados, a ferramenta é especialmente adequada para fluxos de trabalho comumente associados à pesquisa em variação de registro. Em um framework aberto e reprodutível, o TTWtools busca democratizar o acesso à análise avançada de corpora para pesquisadores e educadores.*

1. Introdução

A Linguística de Corpus constitui-se enquanto campo e metodologia a partir da própria possibilidade de se analisar grandes coleções de textos por meio de ferramentas computacionais. A automação desses processos não apenas viabiliza, mas, em larga medida, define esse modo de fazer pesquisa linguística. Ao longo das últimas décadas, diversas ferramentas foram desenvolvidas para dar suporte a esse trabalho. Entre as mais difundidas, destacam-se concordanciadores de propósito geral como o *AntConc*[34] [Anthony, 2026], o *LancsBox X* [Brezina & Platt, 2024] e o *Sketch Engine* [Kilgarriff et al., 2014], que oferecem listas de frequência, análise de palavras-chave, linhas de concordância, etiquetagem morfossintática, construção de corpora, dentre outras funcionalidades. Essas ferramentas constituem referências fundamentais para o campo, embasando pesquisas teóricas e aplicadas, além de oferecerem recursos e análises que podem ser utilizados por professores e aprendizes de línguas, entre outros.

A pesquisa em variação de registro, em especial na tradição da Análise Multidimensional (AMD) [Biber, 1988] amplamente difundida pelos trabalhos de Biber

[34] A etiquetagem morfossintática de corpora no AntConc é realizada por meio de uma ferramenta complementar do mesmo desenvolvedor, o TagAnt [Anthony, 2024], distribuída e executada separadamente do concordanciador.

e colaboradores, tem dependido historicamente de um processo mais especializado: os textos precisam ser etiquetados por uma ferramenta dedicada, as contagens de features lexico-gramaticais computadas e as comparações estatísticas realizadas, frequentemente em ambientes separados. Ferramentas como o *MAT* (*Multidimensional Analysis Tagger*) [Nini, 2019], o *MFTE* (*Multi-Feature Tagger of English*) [Le Foll & Shakir, 2024], o *biberpy* [Sharoff, 2021] e o *BiberPlus* [Alkiek et al., 2025] foram desenvolvidas justamente para operacionalizar esse tipo de pesquisa, implementando a taxonomia de *features* proposta por [Biber, 1988] e tornando a AMD consideravelmente mais acessível[35].

Contudo, observa-se uma lacuna metodológica entre esses dois conjuntos de ferramentas. Por um lado, as ferramentas de propósito geral, embora ofereçam etiquetagem morfossintática, não foram projetadas para etiquetar corpora com taggers especializados na taxonomia de Biber (o que não constitui uma limitação dessas ferramentas, mas reflete simplesmente um escopo distinto de aplicação), tampouco permitem ao usuário escolher ou comparar entre múltiplos taggers dentro de um mesmo ambiente; além disso, são todas de código fechado, o que limita a transparência e a customização por parte dos usuários. Por outro lado, as ferramentas específicas relacionadas à AMD foram desenvolvidas exclusivamente para esse propósito e não oferecem outras funcionalidades exploratórias. O presente trabalho busca superar essa limitação ao propor uma ferramenta de código aberto que unifique etiquetagem morfossintática, extração de frequências e análise estatística em um só ambiente. Embora não tenha a pretensão de se comparar às ferramentas de propósito geral acima citadas, ainda bem mais limitada em termos de funcionalidades exploratórias, o diferencial do TTWtools é integrar bibliotecas de Processamento de Linguagem Natural (PLN) (como spaCy e Stanza) e etiquetadores específicos que implementam a taxonomia de features de Biber. Adicionalmente, o *TTWtools* gera tabelas automatizadas de contagem de etiquetas por texto e oferece outras funcionalidades analíticas que serão exploradas posteriormente neste trabalho. Por ser uma ferramenta de código aberto, o *TTWtools* favorece ainda a transparência metodológica e a replicabilidade das análises.

Após esta breve introdução, em que contextualizamos desafios que o *TTWtools* se propõe a enfrentar, o presente texto organiza-se da seguinte maneira: na segunda seção, apresentamos o desenho e a arquitetura da ferramenta, detalhando suas características gerais de uso, seus componentes (pacote Python e aplicativo gráfico) e as bibliotecas de linguística computacional e etiquetadores (*taggers*) implementados. Na terceira seção, exploramos as funcionalidades da aplicação em seu estágio atual, com destaque para as opções de análise estatística, exploração interativa de dados e as interfaces de visualização individual e comparativa de *corpora*. Por fim, na quarta seção, discutimos as limitações técnicas desta primeira versão e apontamos os aprimoramentos futuros previstos, visando expandir a aplicabilidade do software em contextos transversais de pesquisa e de ensino-aprendizagem.

[35]O *MFTE* disponibiliza um executável com interface gráfica para Windows, gera textos etiquetados em formato vertical e tabelas de contagens normalizadas por 100 palavras (além das contagens brutas e de uma normalização mista baseada em baselines linguisticamente significativos). O *BiberPlus* retorna *dataframes* com contagens de features por texto, sem exportação direta do corpus etiquetado acessível a não-programadores. O *biberpy* produz tabelas de contagens de features, mas não realiza etiquetagem, requerendo que os corpora sejam anotados por ferramentas externas (*spaCy*, *UDPipe*). O *MAT* possui interface gráfica para Windows, oferece contagens, *dimension scores* e exportação dos corpora anotados; contudo, não tem atualizações recentes.

2. TTWtools: desenho e arquitetura de desenvolvimento e uso

O conjunto de ferramentas desenvolvido no âmbito desta pesquisa, denominado *TTW* (*Think to Write*), compreende dois componentes: um pacote Python `ttwtools`, que pode ser utilizado programaticamente sem interface gráfica, e uma aplicação gráfica, *TTWtools*. Ao longo deste texto, o pacote Python será referido sempre em letras minúsculas (`ttwtools`) e o aplicativo gráfico será referido como *TTWtools*. Ambos os componentes são distribuídos como software livre, multiplataforma, e foram concebidos para atender a contextos diversos de pesquisa em linguística de corpus.

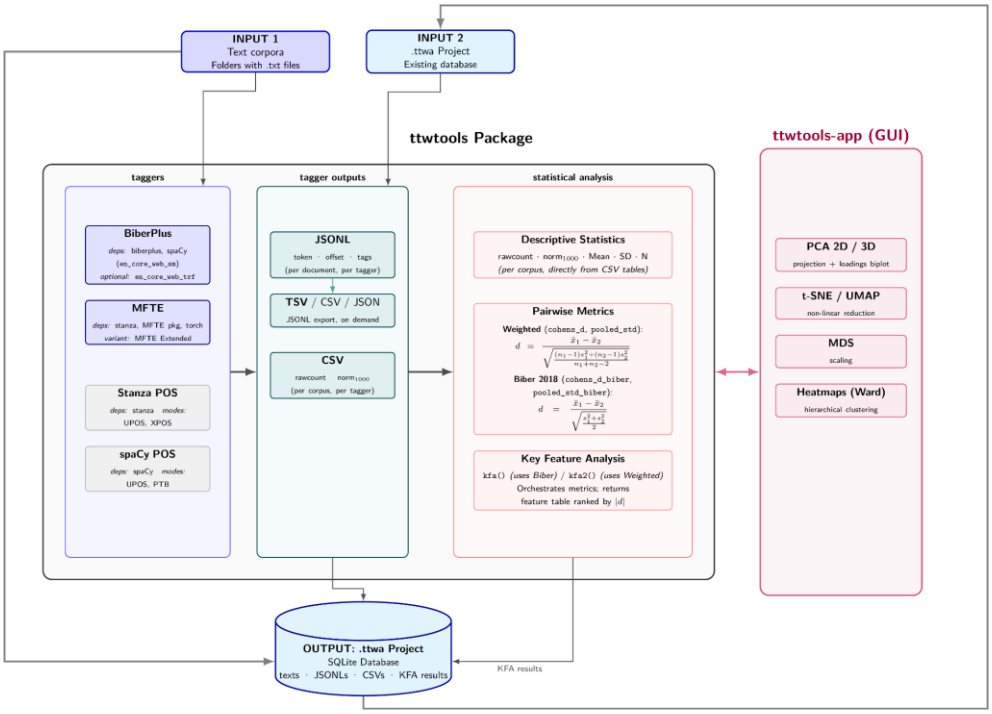


Figura 1. Arquitetura do TTWtools, ilustrando principais funcionalidades a integração entre o *backend* e o *frontend* (GUI).

A arquitetura do *TTWtools* (Figura 1) permite conduzir análises via *scripts* em Python, possibilitando o controle direto das funcionalidades do pacote `ttwtools` por meio da interface gráfica de usuário (GUI) desenvolvida em *HTML/JavaScript* que funciona tanto como aplicativo Desktop quanto como *frontend* Web caso o *backend* rode em um servidor local ou remoto. Essa versatilidade garante que o *TTWtools* possa ser distribuído como aplicativo em diferentes sistemas operacionais (Windows, Linux, macOS) ou que seja hospedado em um servidor web para acesso via navegador.

Os resultados de cada sessão de trabalho podem ser exportados em tabelas individuais ou consolidados em um arquivo de projeto com extensão `.ttwa` que funciona, internamente, como um banco de dados (*SQLite*), agregando múltiplas informações em arquivos/formatos diversos. Essa arquitetura garante o armazenamento seguro dos textos dos *corpora*, de múltiplas camadas de etiquetagem e de análises estatísticas, permitindo a retomada de projetos, o compartilhamento de seções de trabalho e a

retrocompatibilidade do arquivo com versões que porventura incorporem novos etiquetadores e mecanismos de análise estatística, por exemplo.

2.1. O pacote Python `ttwtools`

O pacote Python baseia-se em dois eixos funcionais principais: (1) interface com diferentes ferramentas e bibliotecas para etiquetagem morfossintática e (2) funções para análise estatística individual e comparativa entre *corpora*.

Para a etapa de etiquetagem, o `ttwtools` integra etiquetadores voltados à taxonomia de Biber [Biber, 1988] para análise de variação de registro: o *BiberPlus* (que utiliza opcionalmente a biblioteca Python *spaCy*) e o *MFTE* (que utiliza a biblioteca *Stanza*). Essas duas bibliotecas de PLN (*spaCy e Stanza*) também podem ser empregadas de forma autônoma no `ttwtools`, o que pode ser útil para inferir outras características dos *corpora* e das próprias ferramentas de etiquetagem. O conjunto de etiquetas disponíveis, assim como seu embasamento teórico e níveis de confiabilidade, advêm de cada etiquetador[36]. Na ferramenta *TTWtools,* o usuário pode visualizar a lista completa de tags disponíveis e escolhê-las conforme a necessidade da pesquisa. O processo de etiquetagem (*tagging*) gera arquivos JSONL internamente (com a posição e etiqueta de cada token), uma tabela de contagem absoluta de atributos (*raw count*) e uma tabela de atributos normalizados por 1000 palavras. Tais dados, que também são salvos no arquivo `.ttwta`, podem ser exportados individualmente.

Para a exploração quantitativa dos dados, o pacote conta com um módulo de estatística descritiva individual, que extrai a média e o desvio-padrão dos atributos em cada corpus. Apoiado nessas métricas base, o módulo de análise estatística comparativa implementa processos de *Key Feature Analysis* (KFA) [Egbert & Biber, 2023]. Ao calcular o *d* de Cohen para cada atributo em pares de corpora, a ferramenta ranqueia as *features* pela magnitude da diferença de uso. Além do método originalmente proposto por [Egbert & Biber, 2023], a ferramenta implementa a variante *kfa2*, que substitui a média simples das variâncias dos dois corpora pelo desvio-padrão combinado clássico, ponderado pelo número de documentos de cada corpus. Essa variante visa mitigar distorções em cenários onde há uma diferença expressiva no número de textos entre os corpora.

2.2. O aplicativo gráfico *TTWtools*

Utilizando o pacote `ttwtools` como *backend*, o aplicativo gráfico encapsula as capacidades analíticas da ferramenta em uma Interface Gráfica de Usuário (GUI) construída com *Vite.js*, *D3.js* e outras frameworks *JavaScript*, e empacotada via *PyInstaller*. Essa arquitetura permite distribuir um aplicativo independente e de fácil instalação, equipado com elementos visuais acessíveis e interativos para a exploração de dados. O intuito é, ao mesmo tempo, tornar acessível os processos implementados, democratizar o acesso a ferramentas de linguística de corpus e possibilitar a implementação gradativa de novas funcionalidades e desdobramentos da ferramenta e de

[36]Para consultar a documentação completa, a lista detalhada de etiquetas (*tags*) e as diretrizes de cada etiquetador, veja os respectivos sites e repositórios oficiais: BiberPlus (https://github.com/davidjurgens/biberplus); MFTE (https://github.com/mshakirDr/MFTE); spaCy (https://spacy.io/); e Stanza (https://stanfordnlp.github.io/stanza/).

seus componentes.

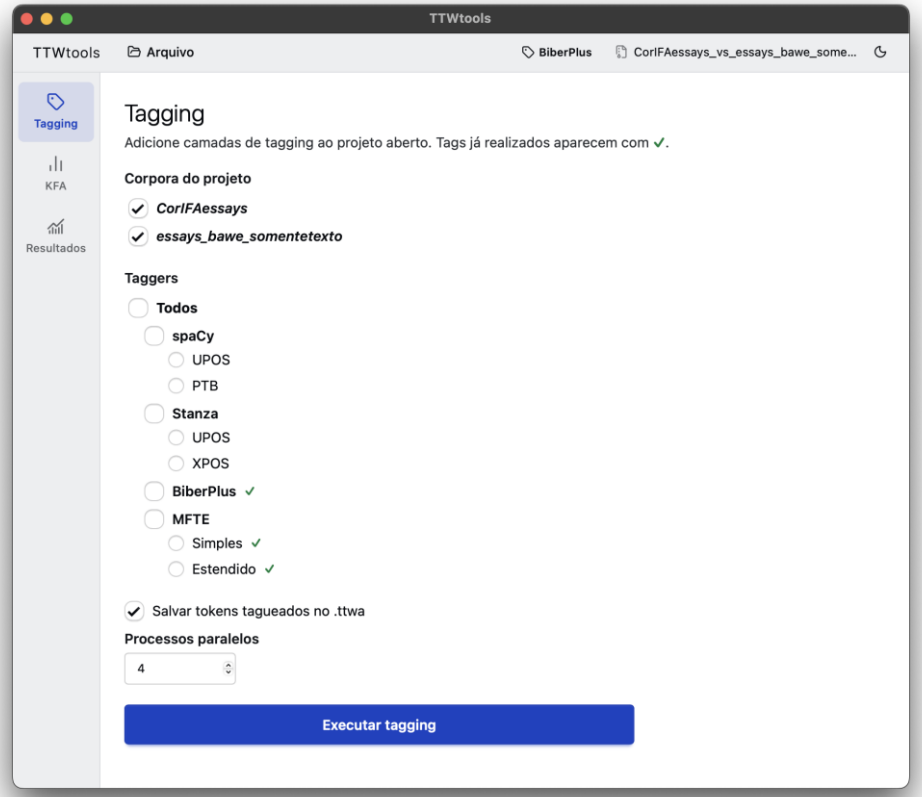


Figura 2. Opções de etiquetamento.

Tal como no uso do pacote Python, o usuário pode fazer o *upload* de um ou dois corpora (armazenados em pastas contendo arquivos de texto em formato `.txt`) e selecionar entre os diferentes tipos de etiquetamento disponíveis. É possível também configurar a aplicação para que os corpora sejam etiquetados por mais de um etiquetador simultaneamente (Figura 2). Neste caso, a ferramenta armazenará os múltiplos resultados em um único arquivo `.ttwa`, permitindo múltiplas camadas de anotação e análise. Após o etiquetamento, os corpora podem ser explorados por meio da ferramenta.

Além do processo de etiquetamento, a ferramenta oferece um conjunto de painéis para análise estatística, visualização e exploração interativa dos dados, descritos em detalhe na Seção 3.

3. Funcionalidades de visualização e exploração de dados

O *TTWtools* estrutura-se em torno de abas temáticas que acompanham o fluxo de investigação, partindo da visualização geral de características estatísticas extraídas a partir de cada camada de etiquetagem (Figuras 3 e 4), até a seleção, a partir de atributos/etiquetas específicos, de textos individuais.

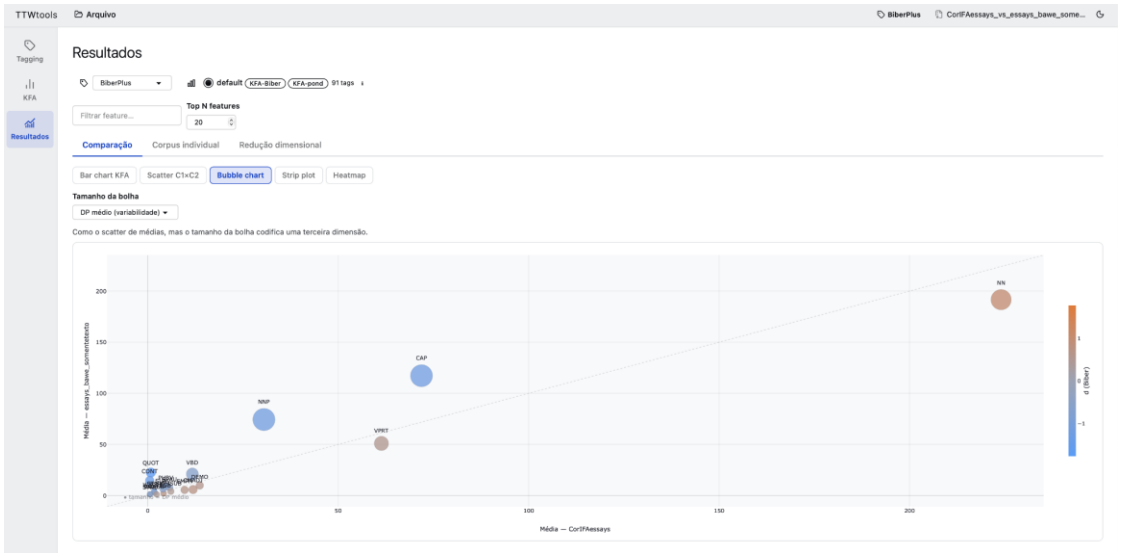


Figura 3. Visualização do desvio padrão comparado de dois corpora.

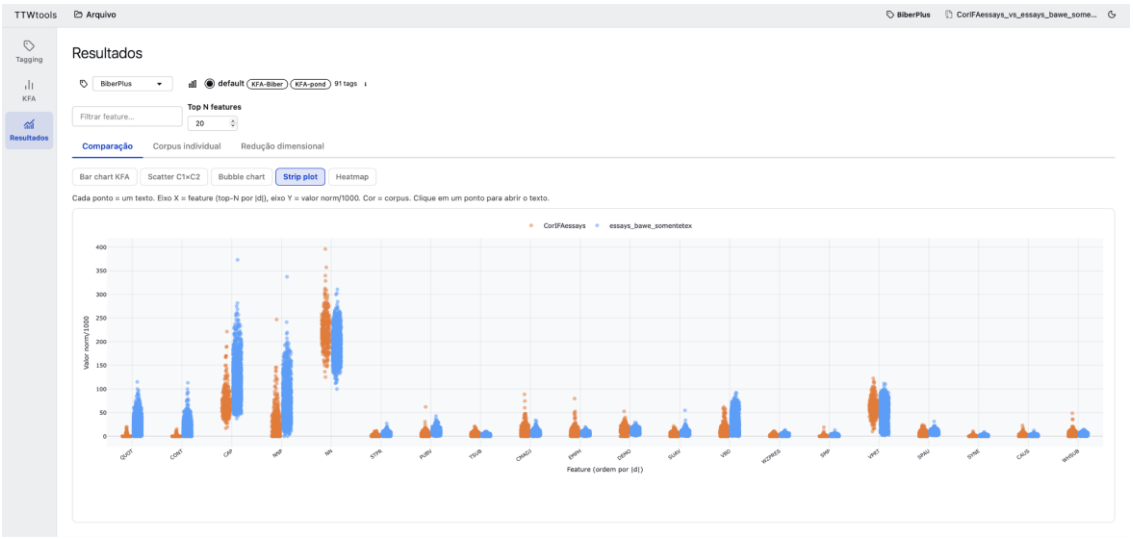


Figura 4. *Strip plot* para visualizar frequência (norm/1000) de atributos em um par de corpora.

3.1. Análise estatística comparativa: Key Feature Analysis

A aplicação oferece um painel de visualização dos dados estatísticos e de exploração interativa dos dados comparativos (Figura 5). Nessa visão comparativa, é possível organizar os *features* distintivos de cada *corpus* utilizando tanto métricas como a KFA (*d* de Cohen) e a KFA ponderada, ou utilizando métricas mais simples como média de ocorrências, desvio-padrão ou contagem absoluta (*raw count*). Ao clicar em qualquer um dos atributos nos gráficos (como aqueles da Figura 5, por exemplo), pode-se visualizar, em uma tela secundária (Figura 6), os textos que contêm aquela etiqueta. Essa visualização exibe os textos de maneira ordenada a partir da métrica selecionada .

Como já mencionado, o módulo de análises estatísticas implementado até o momento, automatiza os processos envolvidos no cálculo de Key-features, conforme proposto por [Egbert & Biber, 2023]. Primeiramente, a ferramenta possibilita escolher qual etiquetador será utilizado naquela análise e oferece a opção de filtrar as tags

desejadas. É possível ainda escolher entre as duas variações de cálculo de Cohen's *d* disponíveis, o método de junção (*union* ou *intersection*) e nomear a análise. Após executar a KFA, o pesquisador pode visualizar os resultados clicando na aba 'resultados'.

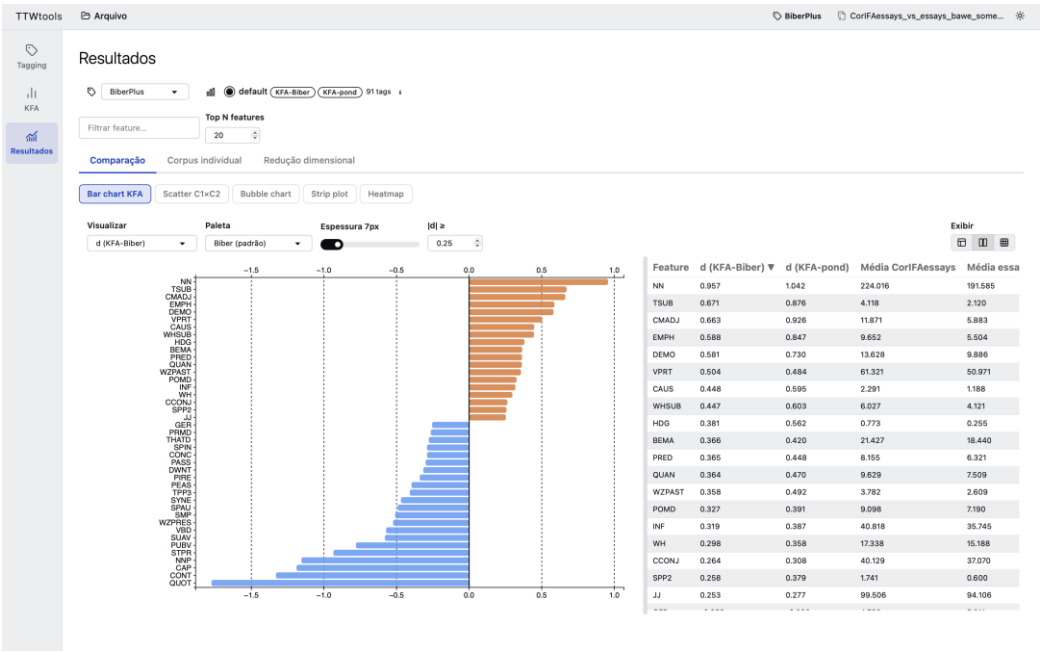


Figura 5. Visualização da *Key Feature Analysis* de dois corpora.

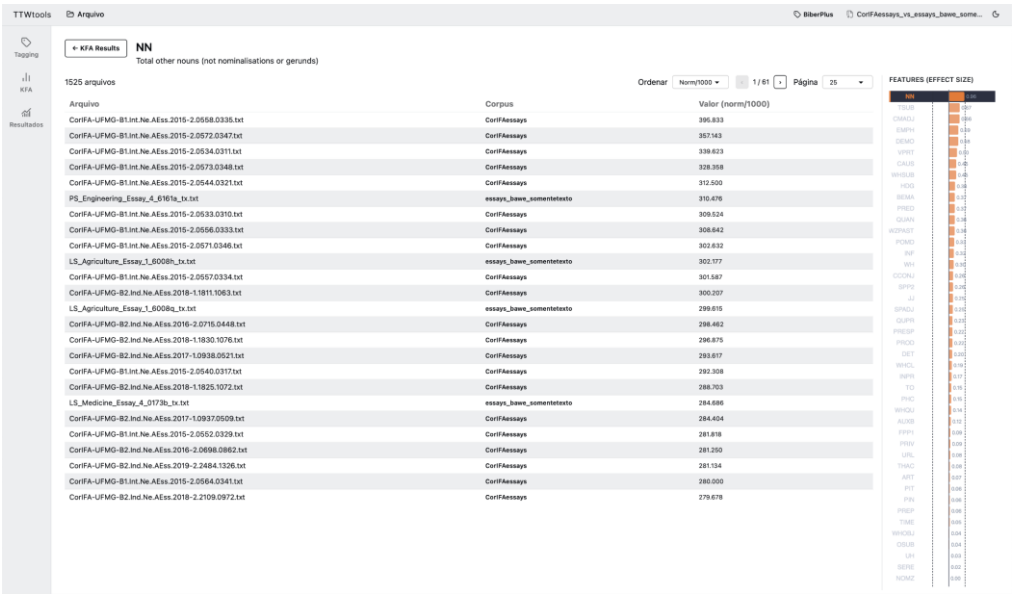


Figura 6. Textos ordenados por frequência de um atributo selecionado.

Por padrão, a aba *Resultados* apresenta, à esquerda, o gráfico de barras divergentes gerado pela análise e, à direita, as *features* correspondentes com suas respectivas métricas (Figura 5). Na parte superior da aba, é possível selecionar os resultados de acordo com o etiquetador desejado (caso mais de um tenha sido selecionado para a análise), filtrar uma *feature* individual ou definir quantas *features* serão exibidas (Top N *features*). Logo abaixo, o pesquisador pode alternar entre os diferentes tipos de visualização disponíveis. No menu *Visualizar*, que, por padrão, exibe o gráfico com os valores do *d* de Cohen, é possível selecionar qualquer uma das métricas disponíveis. Para o gráfico de barras

divergentes, há ainda opções de personalização: paleta de cores, espessura das barras e limiar de *d* a ser considerado na visualização (*d* 0,2, por exemplo). À direita da tela, as *features* analisadas e os valores de cada métrica são apresentados em formato de tabela, organizados conforme a métrica selecionada.

Ao clicar em uma etiqueta, seja na tabela, seja diretamente no gráfico, o pesquisador é direcionado a uma nova tela, na qual são listados os nomes de todos os arquivos (textos) dos corpora em que aquela etiqueta ocorre, a identificação do corpus ao qual cada arquivo pertence e a frequência. Os arquivos são ordenados em ordem decrescente por frequência da etiqueta, cujos valores podem ser exibidos em termos absolutos ou normalizados por mil palavras. Por fim, ao clicar no nome de um arquivo, o pesquisador é conduzido a outra tela, onde pode visualizar o texto integral.

3.2. Visualizador de texto com anotações interativas

A interface central para análise qualitativa renderiza cada texto como um documento com layout paginado. É possível selecionar as etiquetas de interesse, gerando visualizações que permitem identificar, em um texto específico do corpus, as etiquetas selecionadas. O painel lateral lista as *tags* do documento ordenadas por frequência, permitindo ao usuário selecionar as cores e os estilos de anotação desejados. *Tooltips* interativos fornecem definições das features diretamente sobre a palavra em que o cursor do *mouse* está, auxiliando na compreensão e interpretação das etiquetas (Figura 7).

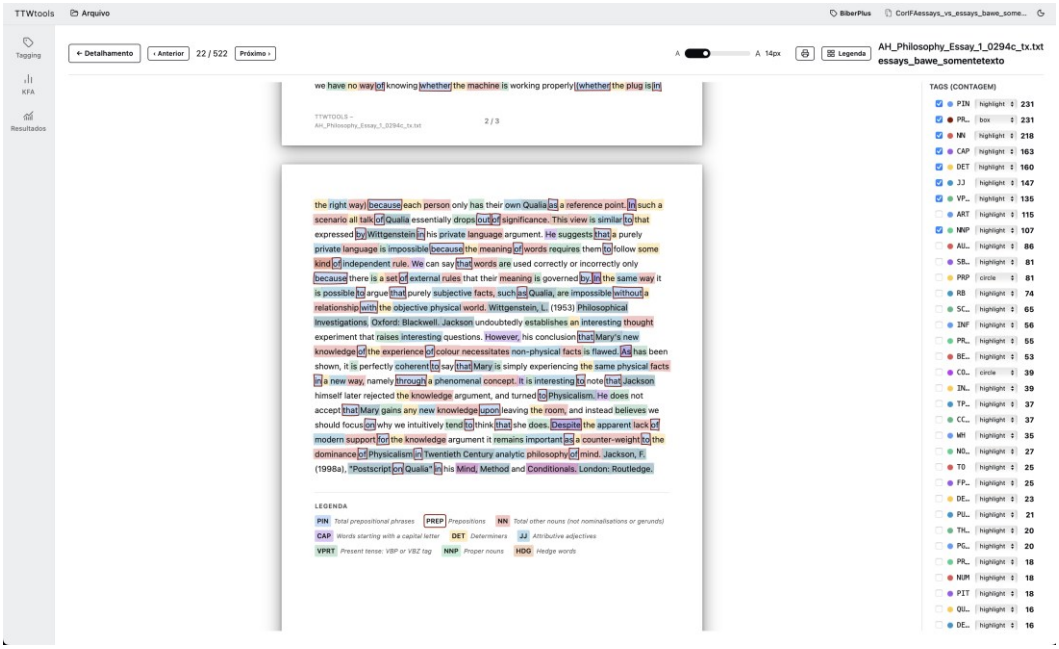


Figura 7. Visualização de um texto, em layout paginado, com seleção e marcação de etiquetas.

4. Conclusão e perspectivas

Neste texto, apresentamos o *TTWtools*, uma ferramenta de código aberto que integra etiquetagem morfossintática, extração de frequências e análise estatística em um único ambiente. A ferramenta compreende um pacote Python e um aplicativo gráfico. O

TTWtools tem o intuito de tornar mais acessíveis os fluxos de trabalho associados à pesquisa em variação de registro, buscando oferecer transparência metodológica, além da flexibilidade e da reprodutibilidade que caracterizam projetos de código aberto.

Atualmente a ferramenta limita-se ao processamento de textos em inglês. Essa restrição reflete o escopo do projeto de doutorado ao qual o *TTWtools* está vinculado, cujos corpora de interesse são compostos por textos nessa língua. Do ponto de vista técnico, no entanto, a arquitetura modular do *TTWtools* foi concebida de modo que novas capacidades possam ser incorporadas incrementalmente, sem necessidade de reestruturação profunda da aplicação. Isso se aplica à adição de novos etiquetadores que poderão, em versões futuras, ser organizados por língua, ampliando o alcance da ferramenta para além do inglês. Da mesma maneira, a estrutura da ferramenta permite implementar novas métricas estatísticas e de interfaces de visualização, tanto para análise comparativa de corpora quanto para a exploração de textos individuais.

Declaração sobre o Uso de Inteligência Artificial

Durante a realização desta pesquisa foram utilizados os modelos Gemini Pro 3.0, Claude Sonnet 4.7 e Deepseek Pro v4 (todos acessados entre maio e setembro de 2026) para tarefas de auxílio à programação e revisão de texto. Todas as sugestões geradas foram revisadas e validadas pelos autores, que assumem total responsabilidade pela precisão e integridade do conteúdo final do texto e da ferramenta nele descrita.

Referências

Anthony, L. (2026). AntConc. Waseda University.

Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press, 1 edition.

Brezina, V. and Platt, W. (2025). #LancsBox X. Lancaster University.

Davies, M. (2008). The Corpus of Comtemporary American English (COCA).

Egbert, J. and Biber, D. (2023). Key feature analysis: A simple, yet powerful method for comparing text varieties. Corpora, 18(1):121–133.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine. Lexicography, 1(1):7–36.

Le Foll, E. and Shakir, M. (2024). The Multi-Feature Tagger of English (MFTE): Rationale, description and evaluation. Research in Corpus Linguistics, 13(2):63–93.

Nini, A. (2019). The Multi-Dimentional Analysis Tagger. In Multi-Dimensional Analysis: Research Methods and Current Issues, pages 67–94. Bloomsbury Academic, London; New York.

Scott, M. (2020). WordSmith Tools. Lexical Analysis Software and Oxford University Press.

Sharoff, S. (2021). Genre annotation for the Web: Text-external and text-internal perspectives. Register Studies, 3(1):1–32.

FREIRE: Um Motor Computacional Adaptativo e Sensível à Lógica para Alfabetização de Adultos em Comunidades Tradicionais

Lucas Parreiras[1]

[1]Instituto Kunumi – Belo Horizonte – MG – Brasil

lucas.parreiras@kunumi.com

Abstract. *Large language models (LLMs) have demonstrated impressive capabilities in text processing, but their deployment in logic-controlled and socially contextualized applications remains uneven. This paper presents FREIRE (Fonemas e Relatos do Educando em Itinerário Regulatório Evolutivo), an adaptive computational engine designed for the literacy of non-literate adults in traditional fishing communities in Cumuruxatiba, Bahia. Built on FastAPI and utilizing gemini-2.5-flash, the system automates a synchronous 6-step pedagogical pipeline inspired by Paulo Freire's dialogic methodology. To ensure accessibility for non-readers, the platform integrates dynamic voice user interfaces (VUIs) with exclusive audio narrators synthesized for each stage. Additionally, the architecture includes a persistent cognitive record that maps mastered generative words to force a progressive scale of phonetic and syllabic complexity in subsequent sessions, alongside real-time methodological auditing via a multi-agent LLM-as-a-Jury approach. Preliminary evaluation in a development environment validates the system's architectural stability, highlighting the current stage of the project as it prepares for field validation with the local community.*

Resumo. *Os modelos de linguagem de grande escala (LLMs) têm demonstrado capacidades impressionantes no processamento textual, mas sua implantação tanto em aplicações controladas por lógica quanto em contextos socialmente motivados ainda é desigual. Este artigo apresenta o FREIRE (Fonemas e Relatos do Educando em Itinerário Regulatório Evolutivo), um motor computacional adaptativo projetado para a alfabetização de adultos não letrados em comunidades pesqueiras tradicionais em Cumuruxatiba, Bahia. Construído sobre o FastAPI e utilizando o modelo gemini-2.5-flash, o sistema automatiza um pipeline pedagógico síncrono de 6 passos inspirado na metodologia dialógica de Paulo Freire. Para garantir a acessibilidade de não leitores, a plataforma integra interfaces dinâmicas de usuário por voz (VUIs) com narradores de áudio exclusivos sintetizados para cada etapa. Além disso, a arquitetura inclui um prontuário cognitivo persistente que mapeia as palavras geradoras dominadas para forçar uma escala progressiva de complexidade fonética e silábica em sessões subsequentes, juntamente com auditoria metodológica em tempo real via uma abordagem multiagente baseada em LLM-*

as-a-Jury. A avaliação preliminar em ambiente de desenvolvimento valida a estabilidade arquitetônica do sistema, ressaltando o estágio atual do projeto que se prepara para a etapa de validação em campo com a comunidade local.

1. Introdução

Parte da comunidade de Processamento de Língua Natural (PLN) para o português brasileiro tem focado historicamente em tarefas de inferência de larga escala e similaridade semântica, a exemplo dos corpora ASSIN [Fonseca et al. 2016] e SICK-BR [Real et al. 2018]. Embora iniciativas mais recentes, como o InferBR [Bencke et al. 2024], explorem o uso de Modelos de Linguagem de Grande Escala (LLMs) na tentativa de avançar o estado da arte (SOTA) na geração de recursos linguísticos, tais modelos ainda apresentam dificuldades estruturais para compreender e processar distinções sensíveis à lógica estrita [Suzgun et al. 2024]. Para além dessa barreira técnica, persiste uma lacuna crítica na aplicação prática do potencial generativo dessas ferramentas em engenharia social ativa, particularmente no domínio da alfabetização de adultos em populações tradicionalmente marginalizadas.

Os métodos tradicionais de alfabetização de adultos frequentemente dependem de cartilhas textuais descontextualizadas que ignoram a realidade sociocultural do aprendiz e impõem uma barreira inicial íngreme devido à forte dependência de habilidades de leitura pré-existentes. Para preencher essa lacuna, este trabalho apresenta o desenvolvimento do FREIRE (Fonemas e Relatos do Educando em Itinerário Regulatório Evolutivo), um motor computacional de código aberto e focado em voz.

O sistema automatiza a transição da narrativa falada espontânea de um pescador não letrado para a autonomia gramatical estrutural. Ao transpor a pedagogia dialógica e a teoria de codificação da realidade de Paulo Freire [Freire 1968] em um pipeline de Inteligência Artificial (IA) síncrono e multimodal, o motor FREIRE contorna a barreira da dependência de texto por meio de guias de áudio independentes e escalonamento silábico dinâmico. Por se tratar de um trabalho em desenvolvimento, o foco atual reside na consolidação de sua robustez arquitetônica para subsidiar a iminente etapa de testes e validação em campo com a comunidade local.

2. Trabalhos Relacionados

A intersecção entre Inteligência Artificial, interfaces de voz e impacto social tem ganhado tração na literatura, embora aplicações focadas em métodos críticos de alfabetização de adultos ainda sejam raras. Historicamente, as iniciativas de PLN voltadas à educação e acessibilidade no Brasil concentraram-se na simplificação textual para pessoas com baixo letramento, a exemplo do pioneiro projeto PorSimples [Aluísio et al. 2008]. Embora essas ferramentas possuam alto valor social, elas ainda pressupõem uma proficiência leitora básica, impondo barreiras para adultos em estágio inicial de alfabetização.

Para transpor a exclusão textual, interfaces baseadas em voz (VUIs) têm sido exploradas para a inclusão digital de populações vulneráveis e comunidades tradicionais [Cuendet et al. 2013]. No entanto, a eficácia sociopedagógica dessas ferramentas depende diretamente da capacidade do modelo de compreender a identidade cultural e as variações

linguísticas do usuário. É neste cenário que se destacam as pesquisas em processamento de dialetos desenvolvidas na Universidade Federal da Bahia (UFBA). Os trabalhos liderados por Claro et al. na modelagem vetorial de variações regionais (como as abordagens de Dialect2Vec) [Mota et al. 2026] demonstram que mapear computacionalmente os regionalismos é um passo fundamental para evitar a alienação cultural em ferramentas de IA. O motor FREIRE dialoga diretamente com essa base empírica, operando via voz para processar o universo vocabular e o dialeto local de Cumuruxatiba sem impor uma norma culta descontextualizada no início do processo de aprendizado.

3. Arquitetura do Sistema

O motor FREIRE adota uma arquitetura síncrona orientada a eventos de voz, projetada para mitigar a latência de execução e eliminar barreiras de interação. O sistema é dividido em três camadas interconectadas:

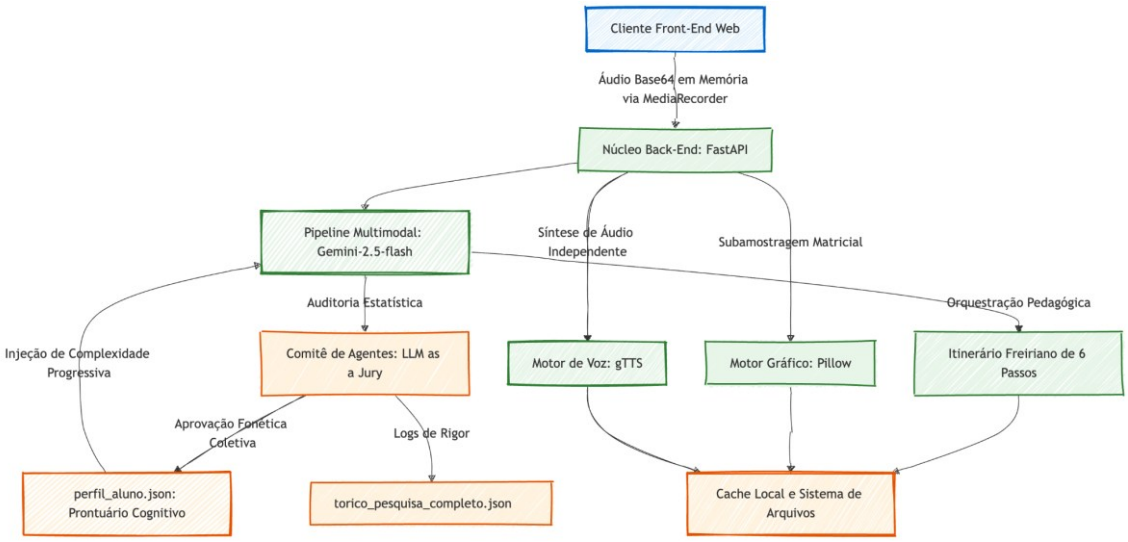


Figura 1. Diagrama arquitetônico do sistema FREIRE dividido em três camadas funcionais: (1) interface de acessibilidade e captura nativa de voz; (2) núcleo de processamento multimodal e orquestração do itinerário pedagógico; e (3) camada de governança de dados, auditoria cega e prontuário cognitivo adaptativo.

3.1. Interface e Acessibilidade de Voz (Front-End)

Operando sobre a API nativa *MediaRecorder*, suportada pelos principais navegadores web modernos (como Chrome, Firefox, Safari e Edge em contextos seguros), o sistema captura relatos falados sobre a rotina diária de trabalho do pescador. Os dados de áudio trafegam encapsulados em um fluxo Base64 diretamente na memória do servidor, evitando o *overhead* de E/S de disco e prevenindo falhas de limite de estado (`FAILED_PRECONDITION`). Para acomodar não leitores, a interface do usuário depende inteiramente de Interfaces de Usuário por Voz (VUIs), em que cada etapa computacional é emparelhada com um componente de reprodução de áudio claro e interativo.

3.2. Orquestração e Processamento Multimodal (Back-End)

A aplicação central é alimentada pelo *FastAPI*, gerenciando *pipelines* generativos concorrentes:

- **O Pipeline Pedagógico (gemini-2.5-flash):** O modelo consome o áudio bruto, realiza a transcrição imediata e extrai uma Palavra Geradora legítima do universo vocabular local. Em seguida, gera os metadados textuais estruturados via restrições de esquema (Pydantic) para preencher as etapas do itinerário educacional.
- **O Motor Gráfico (Pillow):** Renderiza programaticamente a matriz visual da "Ficha da Descoberta", cruzando linhas e colunas fonéticas de forma síncrona em um formato PNG de alta resolução.
- **O Motor de Voz (gTTS):** Gera arquivos de áudio independentes para cada fase pedagógica, transformando as instruções textuais em fala.

4. Fundamentação e o Itinerário Pedagógico de 6 Passos

A modelagem algorítmica deste sistema não fragmenta o conhecimento em módulos sintáticos tradicionais e mecânicos. O motor traduz computacionalmente a essência do método de alfabetização de Paulo Freire, estruturando a transição entre a investigação do universo vocabular, a codificação e a decodificação da realidade. Para evidenciar a rastreabilidade metodológica do sistema, a Figura 2 ilustra o mapeamento direto entre as fases teóricas originais e o *pipeline* computacional executado.

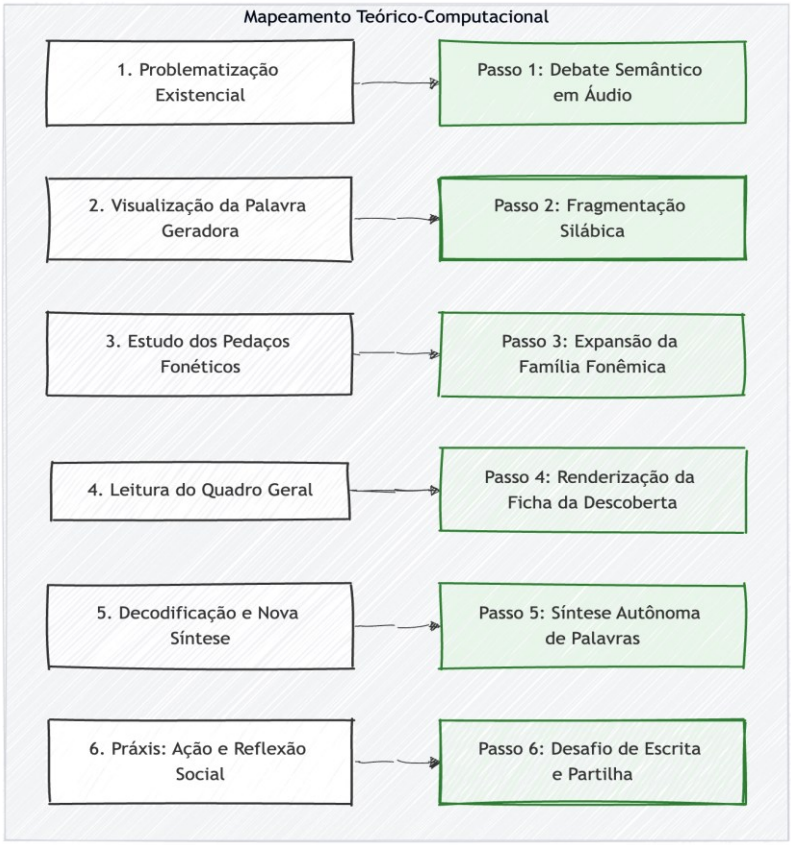


Figura 2. Mapeamento transversal entre a pedagogia freiriana (branco) e a execução computacional síncrona do motor FREIRE (verde).

Abaixo, detalhamos como a arquitetura operacionaliza cada uma dessas seis fases em um fluxo lógico e adaptativo:

1. **Debate Semântico e Contextualização:** Alinhado à fase freiriana de problematização de situações existenciais, um *prompt* de áudio curto e evocativo é gerado para provocar o debate crítico sobre a realidade prática do estudante (ex: discutindo o trabalho com a rede ou o período de defeso da pesca) antes que qualquer texto seja exibido.
2. **Decomposição Silábica:** A Palavra Geradora escolhida é isolada do contexto e dividida explicitamente em suas sílabas constituintes, apresentadas em letras maiúsculas e separadas por hífens (ex: MA-RÉ). Esta etapa estabelece a divisão fonológica inicial e permite ao educando a visualização gráfica exata das unidades silábicas. Na metodologia freiriana, é essa decomposição formal da palavra que substitui a abstração do vocábulo inteiro pela concretude de suas partes formadoras, preparando o terreno metodológico para a exploração das famílias fonêmicas.
3. **Expansão da Família Fonêmica:** O motor desdobra cada uma das sílabas em sua linha fonética completa, combinando as consoantes nativas com as vogais de apoio da língua (ex: MA-ME-MI-MO-MU e RA-RE-RI-RO-RU). Guias de áudio independentes gerados por síntese de voz permitem que o estudante treine a audição e vocalização de cada fonema isoladamente.
4. **A Ficha da Descoberta (Visão Global):** O módulo gráfico (Figura 3) reúne as linhas fonêmicas em uma grade visual bidimensional. Na teoria de Freire, este quadro permite ao alfabetizando descobrir autonomamente o mecanismo lógico das combinações linguísticas, percebendo como a união horizontal e vertical das sílabas cria a estrutura da língua.



Figura 3. Ficha da Descoberta apresentada na interface de usuário. A disposição matricial substitui a abstração do vocábulo inteiro pela visualização combinatória de suas sílabas constituintes.

5. **Síntese Autônoma e Criação:** O *pipeline* avalia e lista palavras reais em português que podem ser montadas a partir do cruzamento das sílabas da grade.

O aluno assume o papel de sujeito ativo, sendo incentivado a combinar as sílabas para formar e vocalizar novos termos como MORA ou REMA.

6. **Desafio de Escrita e Trabalho de Casa:** O ciclo é concluído com um exercício de fixação gráfica e motora guiado por áudio, incentivando o adulto a tentar a expressão gráfica livre em papel físico para compartilhar e partilhar no próximo círculo de cultura comunitário.

5. Avaliação e Métricas de Alinhamento: LLM as a Judge vs. Jury

Para validar cientificamente a execução metodológica e garantir a isenção estatística da pesquisa, o sistema implementa duas arquiteturas distintas de avaliação automatizada baseadas no estado da arte de alinhamento de modelos de linguagem:

5.1. *LLM as a Judge* (Avaliação Centralizada)

A abordagem de *LLM as a Judge* baseia-se na utilização de um único modelo de linguagem de alta capacidade (configurado com Temperatura zero, T = 0.0) atuando como um avaliador centralizado para classificar e criticar as saídas abertas de outros modelos [Zheng et al. 2023, Thakur et al. 2025]. No ecossistema FREIRE, esse agente consome as saídas textuais do *pipeline* pedagógico e atribui notas quantitativas em escalas Likert [Likert 1932] para Fidelidade Pedagógica e Contextualização Local (garantindo o alinhamento estrito com o socioecossistema de pesca de Cumuruxatiba), salvando uma análise descritiva rica diretamente em `historico_pesquisa_completo.json`.

5.2. *LLM as a Jury* (Avaliação por Comitê Multiagente)

Para mitigar os vieses de distribuição e o egocentrismo intrínseco de um único modelo avaliador, adota-se a arquitetura de *LLM as a Jury*. Esta técnica expande o paradigma anterior ao consolidar um comitê de múltiplos agentes com *personas*, instruções ou modelos heterogêneos. Na literatura recente, essa multiplicidade é aplicada tanto para promover debates estruturados que reduzem o viés individual [Chan et al. 2024], quanto para realizar votações cegas sobre a qualidade de uma saída, garantindo que a decisão final emerja de um consenso ou da aprovação por maioria [Li et al. 2024].

No motor FREIRE, o comitê opera de forma cega sobre a resposta fonológica do aluno para determinar a variável booleana `palavra_foi_dominada`. Somente se a maioria da banca validar o sucesso da assimilação semântica e gramatical do ciclo atual, a palavra de ordem é salva de forma persistente em `perfil_aluno.json`:

```
{
    "palavras_dominadas": ["REDE", "MARÉ"]
}
```

Na sessão seguinte, o motor analisa este arquivo e injeta uma restrição ativa no *template* de *prompt*. O modelo é forçado a elevar a dificuldade fonética, transitando de termos dissílabos simples para estruturas silábicas complexas, encontros consonantais ou palavras polissílabas mais longas (ex: movendo de RE-DE para RE-CI-FE ou CU-MU-RU-XA-TI-BA), garantindo uma curva de aprendizado personalizada, adaptativa e neurobiologicamente adequada.

6. Conclusões e Trabalhos Futuros

O motor FREIRE fornece um *framework* estruturado para a implantação de LLMs multimodais em ambientes socioeducativos acessíveis, auditáveis e sensíveis à lógica. Testes preliminares em laboratório mostraram que esquemas estruturados do `Pydantic`, biblioteca de observabilidade e validação de dados, vinculam com sucesso as saídas de texto em grades de interface robustas, enquanto o rastreamento persistente local permite uma progressão cognitiva consistente por meio de bancos de dados cognitivos, validando assim a arquitetura do *framework*.

Como este é um trabalho em desenvolvimento, a validação atual do sistema restringe-se a métricas automatizadas. Nosso próximo passo é incluir especialistas na etapa de avaliação. Conduziremos uma análise qualitativa sistemática das saídas geradas nas seis etapas do itinerário pedagógico, associada a uma meta-avaliação dos mecanismos de *LLM-as-a-Jury*. O objetivo desta etapa é validar a correlação entre os consensos gerados pelos modelos artificiais e o julgamento técnico de educadores e linguistas.

Com os mecanismos de avaliação devidamente calibrados, a pesquisa avançará para a experimentação prática territorial com os pescadores artesanais em Cumuruxatiba (BA). A fase de campo focará em coletar dados de usabilidade, mensurar o impacto pedagógico do motor e testar a robustez da conversão de fala em texto em contextos de variações dialetais e e com ruídos acústicos de fundo (como motores de barcos e ventos costeiros). Essa etapa será executada em parceria com laboratórios acadêmicos locais, visando a consolidação metodológica e a escalabilidade do sistema.

Referências

Fonseca, E. R., dos Santos, L. B., Criscuolo, M., e Aluísio, S. M. (2016). ASSIN: Avaliação de similaridade semântica e inferência textual. In *PROPOR*, páginas 1–8.

Real, L., Rodrigues, A., Vieira e Silva, A., Albiero, B., Guide, B., Thalenberg, B., Silva, C., Câmara, I. C. S., de Oliveira Lima, G., Souza, R., Stanojevic, M., e de Paiva, V. (2018). SICK-BR: A Portuguese corpus for inference. In *PROPOR 2018*.

Bencke, L., Pereira, F. V., Santos, M. K., e Moreira, V. (2024). InferBR: A natural language inference dataset in Portuguese. In *Proceedings of the LREC-COLING 2024*, páginas 9050–9060.

Suzgun, M., Gur, T., Bianchi, F., Ho, D. E., Icard, T., Jurafsky, D., e Zou, J. (2024). Belief in the machine: Investigating epistemological blind spots of language models. *Preprint, arXiv:2410.21195*.

Freire, P. (1968). *Pedagogia do Oprimido*. Rio de Janeiro: Paz e Terra.

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., & Fortes, R. P. M. (2008). Towards Brazilian Portuguese text simplification systems. In *Proceedings of the ACM symposium on Document engineering* (pp. 240–248).

Cuendet, S., Medhi, I., Bali, K., & Cutrell, E. (2013). VideoKheti: making video content accessible to low-literate and novice users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2833–2842).

Mota, L., Claro, D. B., Seno, E. R. M., e Araújo, R. C. (2026). dialect2vec: Um método baseado em vetores para transcrição dialetal do português a partir de questionários do ALiB. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)* - Vol. 1, pages 800–807, Salvador, Brazil. Association for Computational Linguistics.

Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., e Hupkes, D. (2025). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. *arXiv preprint arXiv:2406.12624v5*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1–55.

Chan, C.-H., et al. (2024). Challenging the LLM-as-a-Judge Paradigm: Mitigating Egocentric Bias through Multi-Agent Debate and Peer Review Systems. *arXiv preprint arXiv:2401.07124*.

Li, R., et al. (2024). LLM-as-a-Jury: Consensual Assessment of Generative Outputs via Multi-Agent Committees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ampliação e proposta de sistema de remissivas para um glossário multilíngue direcionado por corpus

Maria Eduarda Magalhães Furtado Dutra[1], Elisa Duarte Teixeira[1]

[1]Departamento de Línguas Estrangeiras e Tradução (LET), Universidade de Brasília

maddudutra2004@gmail.com, elisadut@unb.br

Abstract. *The Multilingual Glossary on Migration and Asylum [Furtado, 2019], TradMigra, was created to be a reference material for translators, interpreters and professionals who work with migrants and asylum seekers. The collection of entries was driven by the Multilingual Corpus on Migration and Asylum - COMMIRE [Furtado; Teixeira, 2022]. This paper describes the creation of new entries for TradMigra in the EN<>PT language pair by identifying Specialized Translation Units - UTEs [Teixeira, 2008] from the keywords of each corpus (using AntConc [Antony, 2024]); the identification of translational equivalents in the comparable target language corpus and the application of a new cross-reference system to the Glossary. The work resulted in 109 new entries, highlighted the need to add more texts to COMMIRE, and contributed to the implementation of Notion as the new platform for storing and filling terminological records.*

Resumo. *O Glossário Multilíngue sobre Migração e Refúgio [Furtado, 2019], TradMigra, foi criado como fonte de consulta para tradutores, intérpretes e profissionais que atuam no acolhimento de migrantes e refugiados. A coleta de suas entradas foi direcionada pelo Corpus Multilíngue sobre Migração e Refúgio - COMMIRE [Furtado; Teixeira, 2022]. Este trabalho relata a criação de novas entradas para o TradMigra no par EN<>PT, passando pelo levantamento de Unidades de Tradução Especializadas - UTEs [Teixeira, 2008] a partir das palavras-chave de cada corpus (via AntConc [Anthony, 2024]), pela identificação de equivalentes tradutórios no corpus comparável da língua de chegada e pela aplicação de um novo sistema de remissivas ao Glossário. Resultaram 109 novas fichas, evidenciando a necessidade de ampliação do COMMIRE e contribuindo para a implementação da plataforma Notion no armazenamento e preenchimento das fichas terminológicas.*

1. Introdução

O deslocamento forçado de indivíduos tem gerado a necessidade de melhoria no acolhimento de refugiados. O último relatório de *Tendências Globais* da ACNUR, publicado em 12 de junho de 2025, informa que 123,2 milhões de pessoas passaram por deslocamento forçado até o fim de 2024. As barreiras linguísticas estão entre os obstáculos mais desafiadores do refúgio, tanto para os deslocados quanto para os profissionais envolvidos em sua recepção. Dessa realidade surgiu, na Universidade de Brasília, o grupo de pesquisa MOBILANG (Mobilidades e Línguas em Contato), que abriga a linha de pesquisa "Terminologia da Mobilidade", desenvolvida em parceria com

o grupo de pesquisa TermiTraDiCo (Terminologia e Tradução Direcionadas por Corpus), ao qual esta pesquisa se filia.

O Glossário Multilíngue sobre Migração e Refúgio [Furtado, 2019], renomeado TradMigra, foi criado para auxiliar na superação dessa barreira linguística. Ele foi elaborado a partir do Corpus Multilíngue sobre Migração e Refúgio - COMMIRE [Furtado; Teixeira, 2022], que conta com textos originalmente escritos em português brasileiro, inglês, francês e espanhol, totalizando cerca de 1,5 milhão de palavras por língua. O corpus foi compilado com base em resultados de pesquisas no Google sobre a obtenção do refúgio — potencialmente os mais acessados por refugiados e profissionais que atendem esse público, como tradutores e intérpretes.

Este artigo relata um trabalho de Iniciação Científica (ciclo 2025/2026) vinculado ao projeto "Terminologia da Mobilidade: criando recursos para tradutores e intérpretes", cujo objetivo foi ampliar e aprimorar o TradMigra por meio do acréscimo de novas entradas no par EN<>PT e da criação de um sistema de remissivas adequado às unidades terminológicas do glossário. Em seguida, apresenta o referencial teórico — Linguística de Corpus e Terminologia, com destaque para os conceitos de Unidades de Tradução Especializadas (Teixeira, 2008), ficha terminológica e sistema de remissivas —, descreve a metodologia de coleta de dados no Corpus, a elaboração de fichas terminológicas, a busca de equivalentes e a aplicação do novo sistema de remissivas aos verbetes e, por fim, apresenta e discute os resultados alcançados.

2. Referencial Teórico

2.1. A Linguística de Corpus

Segundo Berber Sardinha (2004), corpus é um conjunto de textos naturais, em formato eletrônico e legível por computador, coletados como objeto de estudo linguístico e que busca ser representativo da língua ou área especializada estudada. O uso de corpora para investigar padrões linguísticos a partir de dados eletrônicos deu origem à Linguística de Corpus [Barros, 2004]. Essa abordagem empírica permite analisar dados quantitativos e qualitativos do domínio estudado em seu ambiente de ocorrência natural [Tagnin, 2015], possibilitando adquirir conhecimento sobre o léxico especializado da área — sua terminologia —, principalmente por permitir observar a recorrência e co-ocorrência de palavras em contexto (padronização da língua).

Corpora podem ser monolíngues, bilíngues ou multilíngues; quando contêm mais de uma língua, podem ser paralelos (textos e suas traduções) ou comparáveis (textos originalmente produzidos em duas ou mais línguas, de mesma área e, preferencialmente, mesma tipologia) [Tagnin, 2015]. O COMMIRE [Furtado; Teixeira, 2022] é um corpus comparável de textos da área de migração e refúgio, com tipologias similares (formulários, cartilhas, manuais, relatórios etc.), originalmente escritos em cada uma das línguas do projeto (português brasileiro, inglês, francês e espanhol), totalizando cerca de 1,5 milhão de palavras em cada um dos 4 subcorpora.

Na exploração de corpora eletrônicos, ferramentas de processamento permitem gerar uma lista de palavras-chave comparando um corpus de estudo (como o COMMIRE) a um corpus de referência (geralmente de língua geral). O resultado é uma lista de palavras peculiares ao corpus de estudo e pouco típicas no de referência [Tagnin, 2015] — candidatos a termo e um ótimo ponto de partida para compreender a área estudada

[Bowker; Pearson, 2002]. A seguir, será apresentada a Terminologia, ciência que estuda os termos e essencial no contexto da tradução especializada.

2.2. A terminologia na tradução especializada

O contato entre nações tem favorecido a troca de conhecimento produzido em diversas línguas, tornando a tradução especializada imprescindível, por possibilitar o intercâmbio de informações produzidas em diferentes contextos linguísticos. Dessa realidade surgiu a necessidade de se criarem obras de referência especializadas, que possibilitam a consulta a conjuntos terminológicos em contextos monolíngues, bilíngues e multilíngues [Krieger; Finatto, 2004].

O Glossário TradMigra busca tornar a padronização da língua em textos da área de migração e refúgio mais acessível a tradutores, intérpretes e demais usuários que precisem de equivalentes interlinguais. No contexto da Linguística de Corpus como abordagem teórico-metodológica para aquisição de conhecimento linguístico de uma área de especialidade, textos comparáveis são um recurso valioso para familiarização com o domínio e sua terminologia [Bowker; Pearson, 2002]. Os termos de uma área designam seus conceitos especializados [Barros, 2004], e no processo tradutório e de interpretação é pertinente verificar se os vocábulos que costumam acompanhá-los, entendidos aqui como Unidades de Tradução Especializadas (Teixeira, 2008), são recorrentes e, portanto, relevantes para o universo de discurso analisado.

2.3. Unidades de Tradução Especializadas (UTEs)

Unidades de Tradução Especializadas (UTEs), conceito introduzido por Teixeira (2008), são "(...) unidades que tipificam o uso da língua nos textos técnicos" (p. 11). UTEs foram os elementos levados em consideração para constituir as entradas do TradMigra, por estarem inseridas em textos de domínios específicos do saber e constituírem partes do discurso especializado recorrentes e relevantes para tradutores e intérpretes. Para cada nova UTE levantada nos subcorpora do COMMIRE, foi criada uma ficha terminológica.

Fichas terminológicas são o local de registro dos dados relativos aos termos que constituirão as entradas de uma obra de referência [Krieger; Finatto, 2004], podendo seus campos variarem conforme os objetivos do projeto [Barros, 2004]. Um glossário é uma obra que pode envolver uma ou mais línguas e que apresenta termos e seus equivalentes na(s) língua(s) contemplada(s) [Barros, 2004] — UTEs, no caso do TradMigra. Em sua elaboração, deve-se levar em conta campos que aprimorem e facilitem a consulta da obra pelas pessoas usuárias, dentre os quais figura o sistema de remissivas, explicado a seguir.

2.4. Sistema de remissivas

O TradMigra contava inicialmente com um único campo para registrar UTEs relacionadas, "com núcleo semelhante ao da entrada" [Furtado, 2019], que se mostrou ambíguo durante o preenchimento das fichas, conforme relata Gomes (2020). A organização de um sistema de remissivas visou, portanto, dar ao consulente do Glossário um recurso para recuperar as relações semântico-conceituais entre as UTEs.

Barros (2004) afirma que a aplicação de um sistema de remissivas coerente exige critérios quantitativos e qualitativos que determinem as relações a serem priorizadas. De modo geral, as remissivas podem sinalizar variações, informações complementares, disparidades de forma ou conteúdo entre as entradas e quaisquer outros aspectos relevantes para o público-alvo [Barros, 2004].

Tendo apresentado brevemente os principais conceitos das áreas centrais envolvidas neste projeto, a seguir será apresentada a metodologia de trabalho usada para ampliar o glossário e para criar e implementar um sistema de remissivas.

3. Metodologia

3.1 Criação de novas fichas terminológicas no par de línguas EN<>PT

A coleta de novas entradas e o preenchimento das fichas seguiram o procedimento descrito em Gomes (2020). As listas de palavras-chave dos subcorpora de inglês e português brasileiro foram extraídas no início do projeto com o Sketch Engine [Kilgarriff et al., 2004], utilizando os corpora de referência de língua geral já disponíveis na plataforma (Família de Corpora TenTen) e ordenadas por chavicidade decrescente. A partir daí, identificaram-se as palavras-chave não exploradas por outros membros do projeto, cujas UTEs correspondentes foram localizadas no COMMIRE, via consulta semi-automática com o AntConc [Anthony, 2024], e inseridas no TradMigra.

O processo de identificação de candidatos a UTE no corpus também foi realizado conforme descrito em Gomes (2020), seguido por uma avaliação quantitativa para estabelecer se eles possuíam, no mínimo, 3 ocorrências (indicado pela sigla "*freq*" no AntConc) em pelo menos 2 arquivos diferentes do corpus (indicado por "*range*"). Caso o candidato a UTE atendesse, também, a critérios qualitativos, i.e., possuir contextos pertinentes e minimamente elucidativos para o consulente, a UTE era então registrada como ficha terminológica, tal como descrito em Furtado (2019) e Gomes (2020). Os principais campos que compõem a microestrutura dessas fichas são:

UTE – entrada do verbete;
UTEs relacionadas – campo que recebeu o novo sistema de remissivas, discutido no item 3.3;
contexto – exibe dois contextos de uso retirados do corpus;
pt; en; es; fr – equivalentes da entrada nas outras línguas do projeto;
obs – observações sobre especificidades de uso da entrada.

3.2 Identificação de equivalentes tradutórios

O COMMIRE é um corpus comparável que reflete a linguagem natural empregada no âmbito especializado que abrange, mas, diferentemente de um corpus paralelo, ele não fornece equivalentes. A identificação destes depende, portanto, de um percurso investigativo em que o corpus da língua de chegada confirma, ou não, equivalentes propostos por quem traduz ou que foram encontrados em outros materiais de consulta.

Partindo da tradução *prima facie* de uma UTE, era feita uma pesquisa externa na internet para confirmar se essa forma ocorria em textos especializados da área. Esse passo agilizava as buscas no AntConc, pois o navegador corrigia palpites errôneos e indicava a

forma mais usual, evitando o processo mais trabalhoso de busca direta no corpus quando o palpite inicial estava incorreto.

Em seguida, os candidatos a equivalente eram verificados no corpus da língua de chegada seguindo os critérios qualitativos e quantitativos mencionados. Aqueles que atendiam aos critérios e continham uma palavra-chave da língua de chegada, ainda que não explorada pela equipe, se tornavam fichas de equivalente.

3.3 Proposta e aplicação de um sistema de remissivas

Furtado (2019) havia definido o campo "UTEs relacionadas" enquanto local em que seriam colocadas UTEs de núcleo semelhante ao da entrada (Figura 1, imagem da esquerda). Conforme mencionado, tal definição se mostrou ambígua e um contratempo no processo de preenchimento do campo [Gomes, 2020].

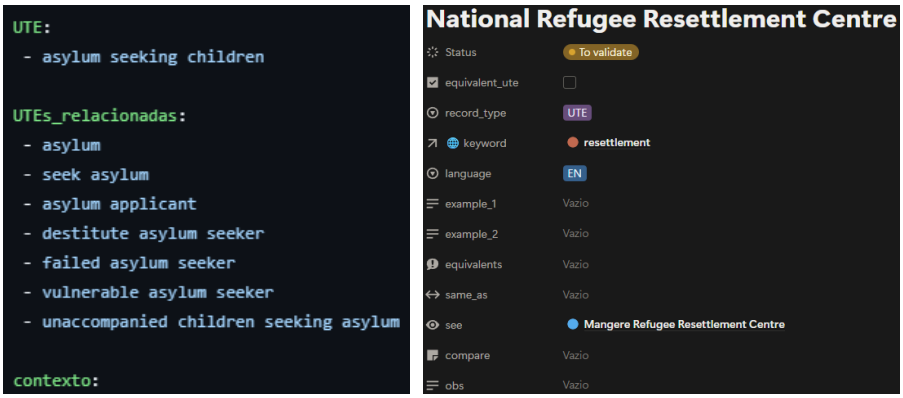


Figura 1. À esquerda: aplicação do antigo campo "UTEs relacionadas". À direita: aplicação da remissiva "vide", que aparece na figura como o campo "see", em novo layout da ficha na plataforma Notion. Fonte: elaboração própria

A obra *Curso Básico de Terminologia* [Barros, 2004] apresenta critérios essenciais a serem levados em conta antes da aplicação de um sistema de remissivas em uma obra terminográfica. Na elaboração da proposta para o TradMigra, foram mantidos em mente as principais relações de significação que as entradas do glossário mantêm entre si, conforme sugerido por Barros (2004). Assim, após avaliação do grupo de pesquisa, ficou definido que as categorias mais pertinentes seriam:

m.q. (mesmo que): usado para indicar sinonímia, como no caso de entradas que possuem mais de uma forma de designação para um mesmo conceito (Figura 2), como nomes por extenso e suas respectivas siglas e/ou formas variadas de se referir ao mesmo órgão, documento, lei, etc. A relação de sinonímia foi averiguada a partir da verificação dos contextos de uso de possíveis termos sinônimos;

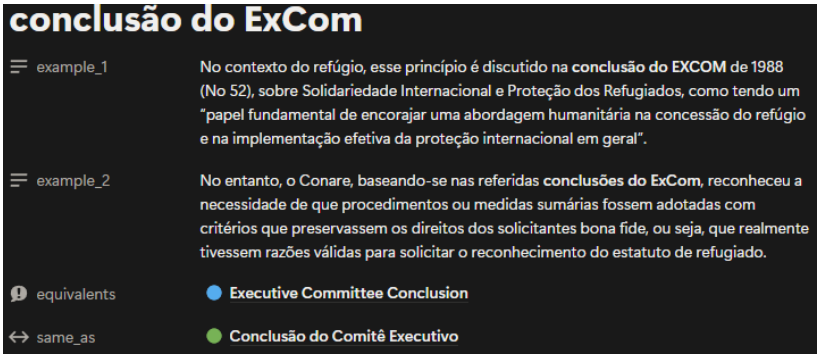


Figura 2. Exemplo de aplicação da remissiva "m.q.", que aparece na figura como o campo "same_as". Fonte: elaboração própria

vide: nos casos em que uma das formas sinônimas não contenha palavra-chave e/ou não possua frequência mínima no corpus, constituirá "ficha vazia", apenas com o campo **vide** preenchido, direcionando o consulente para a forma equivalente que foi registrada como entrada (Figura 1, imagem da direita);

C.f. (confrontar com): utilizada em fichas de UTEs que possuem relação de conteúdo complementar ou contrastante e que precisam ser comparadas para um entendimento completo (Figura 3).

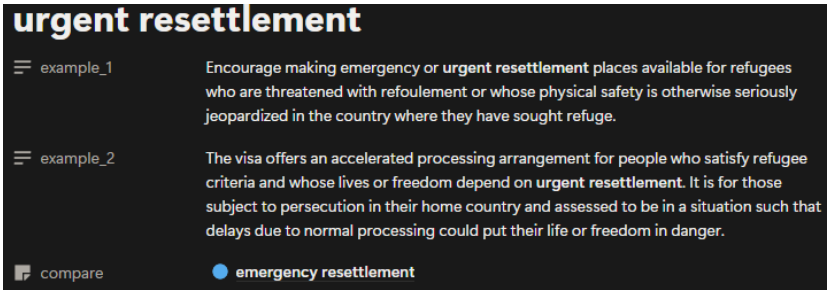


Figura 3. Exemplo de aplicação da remissiva "C.f.", que aparece na figura como o campo "compare". Fonte: elaboração própria

Tal proposta foi testada nas fichas criadas neste trabalho, o que auxiliou no estabelecimento do papel definitivo de cada remissiva.

Na seção a seguir, serão apresentados e discutidos alguns dos resultados obtidos neste ano de execução do Projeto e apontadas questões que surgiram durante a execução prática dos trabalhos,exigindo uma reflexão teórica mais aprofundada e a busca por soluções adequadas.

4. Resultados e Discussão

Foram acrescentadas 109 novas fichas ao TradMigra, 41 para o português e 68 para o inglês. Dessas, 29 foram fichas de equivalente, com 8 delas sendo equivalentes em EN de fichas em PT que já haviam sido feitas por outros pesquisadores do projeto em anos anteriores (Quadro 1).

Quadro 1. Palavras-chave das quais novas fichas e seus respectivos equivalentes foram criados. Fonte: elaboração própria

palavra-chave	língua de partida	ocs. no corpus	fichas geradas na LP	fichas geradas na LCh
apatridia	PB	870	15	5
comissariado	PB	474	6	1
IMDH	PB	176	2	0
conclusão	PB	266	4	2
resettlement	EN	1545	52	13

A palavra-chave "*resettlement*", por exemplo, rendeu 52 novas UTEs para o Glossário, das quais apenas 13 tiveram equivalentes que atendessem aos critérios para criação de nova ficha terminológica. O Quadro 2 mostra as 3 UTEs com mais de 50 ocorrências no corpus de inglês criadas a partir dessa palavra-chave, os respectivos equivalentes identificados e as frequências no subcorpus de português brasileiro.

Quadro 2. Amostra de UTEs da palavra-chave "resettlement" com mais de 50 ocorrências no COMMIRE que constituíram fichas de equivalente. Fonte: elaboração própria

UTE	Frequência da UTE no corpus	Equivalente	Freq. do equivalente no corpus
refugee resettlement	148	reassentamento de refugiados	66
resettlement country	84	país de reassentamento	39
resettlement programme	85	programa de reassentamento	147

O banco de dados do glossário e o site por onde era possível acessá-lo eram hospedados gratuitamente no Google Cloud. A plataforma passou a cobrar pela hospedagem e, por isso, o site está atualmente fora do ar. Como o novo local de armazenamento do TradMigra ainda era indefinido no início desta pesquisa, seguiu-se o que havia sido descrito em etapas anteriores, com a utilização do Bloco de Notas e o formato .TXT para preenchimento das fichas. A partir de 2026, a plataforma Notion passou a ser adotada. Além de oferecer armazenamento gratuito, forneceu uma interface mais rápida, intuitiva e organizada para o preenchimento das novas fichas.

A ausência de ocorrências suficientes no COMMIRE impossibilitou a criação de várias fichas de UTEs pertinentes, com inúmeras ocorrências em buscas online, como é o caso de "*projected resettlement needs*" (20 ocorrências somente em um arquivo do Corpus) e "*statelessness determination procedure*" (apenas 2 ocorrências, em 2 arquivos). Os critérios de preenchimento de fichas, que exigiam a ocorrência de pelo menos dois contextos de uso para cada UTE, por vezes não puderam ser atendidos, como nos casos de "apatridia *de jure*", "*ad hoc resettlement*" e "*Resettlement Admission Targets*", por exemplo, cujos contextos disponíveis eram idênticos ou pouco elucidativos (Figura 4).

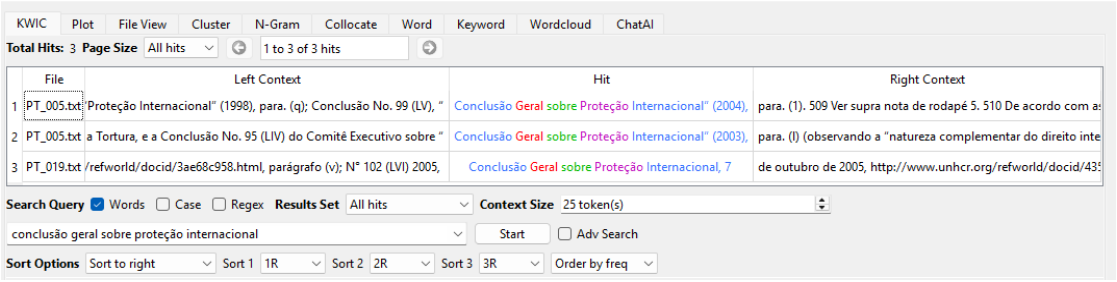


Figura 4. Exemplos de contextos não elucidativos em busca feita no AntConc.
Fonte: elaboração própria

Das 109 fichas criadas, 16 correspondem a equivalentes na direção PT>EN, e 13 na direção EN>PT. Mesmo quando encontrados, alguns equivalentes não geraram ficha por ausência de ocorrências suficientes no corpus, como foi o caso de "*statelessness determination procedure*", "*incidence of statelessness*" e "programa de reassentamento do ACNUR". Todos esses fatores revalidaram a necessidade de ampliação e melhora no balanceamento do COMMIRE, conforme discutido em Furtado e Teixeira (2022).

O sistema de remissivas apresentado na subseção 3.3 foi facilmente aplicado nas fichas terminológicas, confirmando a adequação dos critérios estabelecidos ao tipo de entrada do glossário e ao contexto linguístico da área de migração e refúgio.

5. Considerações finais

A criação de 109 novas fichas terminológicas, a identificação de UTEs equivalentes com auxílio do AntConc e a implementação de um sistema de remissivas, que possibilitou a recuperação de relações léxico-semânticas entre UTEs, ampliaram e enriqueceram o TradMigra, tornando-o uma ferramenta de consulta mais útil para tradutores e intérpretes da área de migração e refúgio. A nova plataforma de armazenamento (Notion) trouxe benefícios práticos e organizacionais que tornarão a continuidade do projeto mais fluida. Embora o COMMIRE necessite de ampliação enquanto fonte de extração de dados, de modo a se tornar mais representativo da área, os resultados obtidos superaram a proposta inicial, beneficiando futuros pesquisadores e consulentes. Por fim, é importante apontar a necessidade futura de aplicação do sistema de remissivas em fichas feitas fora deste trabalho, coletadas em projetos anteriores, de modo a fazer com que o sistema contemple a totalidade do TradMigra, bem como a urgência de deixá-lo disponível novamente na internet, para que beneficie outros usuários para além dos projetos de pesquisa envolvidos.

Questões éticas

O trabalho não envolveu questões éticas. Não houve uso direto de Inteligência Artificial (IA) em nenhuma das etapas descritas. Como aluna bolsista remunerada pela Universidade de Brasília do Programa de Iniciação Científica, Maria Eduarda Magalhães Furtado Dutra executou as etapas teóricas e práticas sob orientação da Prof.ª Dr.ª Elisa Duarte Teixeira, responsável também pela revisão do presente artigo. Anna Beatriz Dimas Furtado atuou como consultora ao longo do projeto, responsável pela transição do banco de dados para o Notion, sem contribuir para a escrita deste artigo.

Referências

ANTHONY, L. *AntConc*: A freeware corpus analysis toolkit for concordancing and text analysis. Versão 4.3.1. Tóquio, Japão: Waseda University, 2024. Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 19 abr. 2026.

BARROS, L. A. *Curso Básico de Terminologia*. São Paulo, SP, Brasil: Edusp, 2004.

BERBER SARDINHA, T. *Lingüística de corpus*. São Paulo: Editora Manole, 2004.

BOWKER, L.; PEARSON, J. *Working with Specialized Language: A practical guide to using corpora*. 1ª. ed. London; New York: Routledge, 2002.

FURTADO, A. B. D. *Glossário Multilíngue Online sobre Migração e Refúgio: uma proposta para tradutores e intérpretes*. Trabalho de Conclusão de Curso (Bacharelado em LEA-MSI) – Universidade de Brasília, Brasília, 2019. Disponível em: https://bdm.unb.br/handle/10483/23240. Acesso em: 18 abr. 2026.

FURTADO, A. B. D.; TEIXEIRA, E. D. Multilingual Corpus on Migration and Asylum (COMMIRE): planning, compilation, and overall content. *Texto Livre*, v. 15, p. 1–18, 2022.

GOMES, G. R. DA S. *Ampliação do Glossário Multilíngue Online Sobre Migração e Refúgio: O procedimento de criação de novas fichas.* Trabalho de Conclusão de Curso (Bacharelado em LEA-MSI) - Universidade de Brasília, Brasília, 2020. Disponível em: https://bdm.unb.br/handle/10483/27167. Acesso em: 19 abr. 2026.

KILGARRIFF, A. et al. *The Sketch Engine*. University of Brighton: Information Technology Research Institute Technical Report Series, jul. 2004.

KRIEGER, M. DA G.; FINATTO, M. J. B. *Introdução à Terminologia: Teoria e Prática*. 1ª. ed. São Paulo: Contexto, 2004.

TAGNIN, S. E. O. A Linguística de Corpus na e para a Tradução. In: *Corpora na Tradução*. São Paulo: HUB Editorial, 2015.

TEIXEIRA, E. D. *A Lingüística de Corpus a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual*. 2008. Tese de Doutorado - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2008. doi:10.11606/T.8.2008.tde-16022009-141747. Acesso em: 19 abr. 2026.

UNHCR. *Global Trends*. Disponível em: <https://www.unhcr.org/global-trends?_gl=1>. Acesso em: 18 abr. 2026.

Compilação e anotação de um corpus paralelo inglês–português de legendas para identificação de expressões idiomáticas

Rozane R. Rebechi[1], Diego Alves[2], Isabela Moreira de Oliveira[1], Christopher dos Santos Filho[1]

[1] Universidade Federal do Rio Grande do Sul (UFRGS), Brasil
[2] Universität des Saarlandes, Alemanha

rozane.rebechi@ufrgs.br, diego.alves@uni-saarland.de,
demoreirapraoliveira@gmail.com, crissfilho@gmail.com

Abstract. *Idiomatic expressions (IEs) challenge Translation Studies and Natural Language Processing because of their non-compositionality and context dependence. This study investigates the identification of IEs and their translation equivalents in an English–Portuguese parallel subcorpus from OpenSubtitles. The openai/gpt-oss-20b LLM analyzed 10,000 subtitle pairs using an instruction-based prompt with a task demonstration. Three annotators manually validated the results. Preliminary findings reveal difficulties in distinguishing literal from figurative uses, assessing conventionality, and handling heterogeneous subtitle quality, while highlighting the potential of parallel corpora for research on translation and figurative language.*

Resumo. *Expressões idiomáticas (EIs) desafiam os Estudos da Tradução e o Processamento de Linguagem Natural devido à não composicionalidade e à dependência contextual. Este estudo investiga a identificação de EIs e de seus equivalentes tradutórios em um subcorpus paralelo inglês–português do OpenSubtitles. O LLM openai/gpt-oss-20b analisou 10.000 pares de legendas por meio de um prompt baseado em instruções e em um exemplo demonstrativo da tarefa. Os resultados foram validados manualmente por três anotadores. Os resultados preliminares evidenciam dificuldades na distinção entre usos literais e figurados, na convencionalidade e na qualidade heterogênea das legendas, mas apontam o potencial de corpora paralelos para pesquisas sobre tradução e linguagem figurada.*

1. Introdução

As expressões idiomáticas (EIs) são combinações de duas ou mais palavras que funcionam como uma unidade semântica relativamente fixa e convencionalizada. Em geral, apresentam diferentes graus de opacidade semântica, de modo que seu significado não pode ser plenamente inferido a partir da soma dos significados individuais de seus constituintes lexicais [Alisoy 2025]. Além da não composicionalidade, essas expressões tendem a apresentar certo nível de estabilidade estrutural, já que alterações lexicais ou sintáticas podem comprometer o reconhecimento de seu sentido idiomático.

As EIs constituem fenômenos complexos para os estudos linguísticos e tradutórios não só devido a sua natureza não composicional, mas também à forte dependência de fatores culturais e contextuais. Assim, são desafiadoras tanto para sistemas de tradução

automática [Rebechi 2024] quanto para tradutores humanos [Rebechi et al. 2025]. Mais recentemente, os grandes modelos de linguagem (LLMs) vêm sendo empregados em tarefas relacionadas à linguagem figurada, incluindo identificação automática de EIs e extração de equivalentes tradutórios. Entretanto, ainda enfrentam dificuldades quanto à delimitação do que constitui uma EI, bem como à distinção entre usos literais e figurados [Sag et al. 2002; Donthi et al. 2025].

Utilizando um corpus paralelo de legendas extraído do OpenSubtitles v2024 [Lison e Tiedemann 2016], um amplo corpus multilíngue compilado a partir de legendas de filmes e séries, o presente trabalho tem como objetivo analisar as EIs identificadas pelo LLM openai/gpt-oss-20b e seus equivalentes tradutórios em português brasileiro, examinando, por meio de validação humana, a adequação das identificações realizadas pelo modelo.

2. Fundamentação teórica

A tradução de expressões idiomáticas (EIs) constitui um desafio para tradutores humanos e sistemas automáticos devido, entre outros fatores, à não composicionalidade e à convencionalidade que caracterizam esse tipo de unidade fraseológica [cf. Tagnin 2013]. Em muitos casos, o significado de uma EI não pode ser inferido a partir da interpretação isolada de seus constituintes lexicais, exigindo a consideração de aspectos contextuais, culturais e pragmáticos. Por exemplo, a EI inglesa *COMPARE apples and oranges* não pode ser traduzida literalmente como "COMPARAR maçãs e laranjas" sem perda do sentido idiomático, sendo uma possibilidade em português "MISTURAR alhos com bugalhos". Além disso, características formais, como a relativa fixidez e a ordem dos elementos lexicais, podem contribuir para o reconhecimento e a recuperação semântica das EIs, de modo que alterações estruturais podem dificultar a ativação de seu significado convencionalizado.

Neste estudo, a não composicionalidade é considerada em termos graduais, e não como propriedade absoluta das EIs. Para fins de anotação, são consideradas EIs as unidades poliléxicas convencionalizadas que, no contexto analisado, apresentam sentido figurado ou significado global não plenamente recuperável pela combinação dos sentidos mais imediatos de seus constituintes. A convencionalidade e a fixidez relativa funcionam como critérios complementares. Esses parâmetros permitem distinguir as EIs de colocações, marcadores discursivos e combinações livres, ainda que as fronteiras entre essas categorias nem sempre sejam facilmente identificadas. A interpretação contextual é especialmente importante para sequências que admitem tanto leitura literal quanto idiomática, como *white elephant*.

A convencionalidade não se restringe às EIs, mas caracteriza diferentes unidades fraseológicas e pode constituir uma dificuldade para a tradução. Embora não tratem diretamente de EIs, Goldschmidt et al. (2017) mostram que a correspondência palavra por palavra pode comprometer a convencionalidade de colocações, uma vez que combinações recorrentes em uma língua nem sempre são lexicalmente reproduzidas da mesma forma em outra. É o que ocorre, por exemplo, com o espanhol *tiempo completo*, cuja correspondência convencional em português é *tempo integral*, e não *tempo completo*. Assim, a proximidade excessiva com a estrutura lexical do texto de partida pode resultar em soluções semanticamente compreensíveis e gramaticalmente possíveis, mas pouco convencionais na língua de chegada.

Nesse sentido, corpora constituem importantes recursos para a tradução, pois permitem observar padrões linguísticos, combinações recorrentes e soluções convencionalizadas em textos autênticos [Stewart 2000]. Segundo Tagnin (2013), a convencionalidade manifesta-se nos níveis sintático, semântico e pragmático e desempenha papel importante no reconhecimento e na interpretação das unidades fraseológicas. A consulta a corpora permite, portanto, verificar empiricamente não apenas se determinada solução é possível na língua de chegada, mas também se ocorre de forma recorrente e em contextos comparáveis, contribuindo para a identificação de usos mais convencionais.

No caso específico das EIs, o problema da literalidade pode ocorrer em diferentes etapas do processo tradutório. Uma expressão empregada idiomaticamente pode receber uma tradução literal – inadequada – quando o tradutor, humano ou automático, não reconhece sua natureza idiomática no contexto. Mesmo quando a idiomaticidade é identificada, pode haver preferência por equivalentes formalmente mais próximos da expressão de partida, sobretudo quando a tradução literal corresponde também a uma expressão possível na língua de chegada. Rebechi et al. (2025) observaram que tanto tradutores em formação quanto sistemas de tradução automática tendem a privilegiar soluções mais óbvias e formalmente próximas. É o caso de *hot potato*, frequentemente traduzida por *batata quente*, solução formalmente correspondente e também convencional em português, em detrimento de outras possibilidades igualmente idiomáticas, como *pepino* ou *abacaxi [para descascar]*.

A proximidade formal, entretanto, pode resultar em quebra de convencionalidade quando a combinação correspondente não é convencionalizada na língua de chegada. Rebechi et al. [no prelo] observaram esse fenômeno na tradução automática de EIs relacionadas à alimentação. No trecho "*[...] he was as cool as a cucumber. He never forgot a line. And he never let anything ruffle him.*", os sistemas tenderam a preservar a imagem lexical da expressão inglesa, traduzindo *cucumber* por "pepino" e *cool* por "frio", "tranquilo" e "legal". Como resultado, observaram-se combinações não convencionalizadas em português. Nesse caso, o sentido idiomático poderia ser recuperado por uma expressão como "tranquilo como/feito/que nem um grilo".

A investigação dessas escolhas pode se beneficiar de corpora paralelos, que permitem observar como unidades fraseológicas são efetivamente traduzidas em contexto. Além disso, estudos baseados em corpus mostram que textos traduzidos podem apresentar regularidades próprias, frequentemente associadas ao chamado *translationese*. Rubino et al. (2016), por exemplo, demonstram que essas características podem ser identificadas por meio de métodos quantitativos aplicados a dados de tradução.

Neste estudo, um corpus paralelo inglês–português é utilizado para investigar a identificação e a tradução de EIs por sistemas de IA, com especial atenção às soluções produzidas em contexto. As traduções são analisadas segundo as seguintes categorias: equivalência idiomática e convencional; equivalência idiomática, mas não convencional; tradução por equivalente não idiomático (*trope change*, Chesterman, 1997); e erro de tradução. A análise permitirá, assim, examinar em que medida os sistemas reconhecem o uso idiomático das expressões em inglês e seus equivalentes em português.

3. Metodologia

Pesquisas baseadas em corpus indicam que as EIs apresentam distribuição desigual entre gêneros e registros. Segundo Biber et al. (1999), embora relativamente raras em textos escritos em geral, as EIs tendem a ocorrer com maior frequência em diálogos ficcionais, nos quais contribuem para efeitos de informalidade e representação da fala cotidiana. Com base nessas observações, este estudo utiliza o corpus OpenSubtitles [Lison e Tiedemann 2016], composto por legendas de filmes e séries no par de línguas inglês–português brasileiro, que contém mais de 115 milhões de pares de frases alinhadas e é formado tanto por fansubs, ou seja, legendas produzidas colaborativamente por fãs, quanto por profissionais. Portanto, essas traduções apresentam qualidade heterogênea, variando entre soluções idiomáticas adequadas e traduções excessivamente literais ou equivocadas [O'Hagan 2009; Dwyer 2012].

3.1. O corpus paralelo

Nesta etapa exploratória, cujo objetivo é avaliar a viabilidade de um procedimento automático para a identificação de EIs em inglês e de seus equivalentes tradutórios em um corpus paralelo de legendas, foi utilizado um recorte constituído pelos primeiros 10.000 pares do subconjunto inglês–português brasileiro. Esse procedimento não constitui uma amostragem probabilística e, portanto, não permite excluir possíveis vieses decorrentes da ordenação dos dados no corpus. As legendas potencialmente contendo EIs foram identificadas automaticamente pelo *openai/gpt-oss-20b*, modelo com aproximadamente 21 bilhões de parâmetros (OpenAI, 2025). Dos 10.000 pares analisados, o modelo identificou 928 legendas em inglês como contendo EIs.

Para a identificação e anotação automáticas, utilizou-se uma estratégia de prompting baseada em instruções e em um exemplo demonstrativo da tarefa, em que cada sentença foi apresentada individualmente ao modelo, acompanhada de instruções que incluíam uma definição de EI para fins desta pesquisa e exemplos ilustrativos. O modelo openai/gpt-oss-20b foi executado por meio do pipeline text-generation da Hugging Face, com os seguintes parâmetros: max_new_tokens=600 para o primeiro conjunto de dados e max_new_tokens=500 para o segundo; do_sample=False (greedy decoding, sem amostragem); temperature=0.0; top_p=1.0; e seed=42. O modelo foi carregado com dtype="auto" e device_map="auto", e as inferências foram realizadas em lotes de 20 exemplos. Cada par de legendas foi processado independentemente, sem recuperação de exemplos adicionais do corpus. O prompt completo utilizado na tarefa é apresentado no Quadro 1.

Quadro 1 – Prompt utilizado para identificação das EIs e de seus equivalentes

SYSTEM PROMPT = """You are a linguist expert in translation from English to Portuguese.

Objective: Your task is to analyse an English sentence (EN SENT) containing an idiomatic expression (IE) and check a Portuguese translated sentence to identify which tokens in the Portuguese sentence (PT SENT) correspond to the translation

of the idiomatic expression from English.

For group of sentence in English, idiomatic expression in English, translated sentence in Portuguese, you must respond with EXACTLY ONE LINE containing the words in the Portuguese sentence that corresponds to the translation of the idiomatic expression in English.

IMPORTANT RULES: - Your entire response must be a single line - Do not include any other text, explanations, or analysis

Example input: EN SENT: Watch your back

Watch your back. Right, K-6 were going to K-6.

Hate to say I told you so. IE: Watch your back

PT SENT: tome cuidado, certo, k-6, vamos para k-6, odeio diga que eu disse a você, bem, nós vamos conseguir para testá-lo imediatamente, saberemos em breve,

Example output: EI TRANSLATED TOKENS ARE: tome cuidado Remember: PLEASE, the translated token information must be given in a separate line, starting with ”EI TRANSLATED TOKENS ARE:”. For example, ”EI TRANSLATED TOKENS ARE: tome cuidado”. ”””

Além disso, o LLM também foi responsável por (i) atribuir um identificador ao item; (ii) identificar uma possível EI na legenda em inglês; e (iii) localizar, na legenda em português, o possível equivalente tradutório.

Os resultados produzidos pelo modelo foram posteriormente organizados para validação manual.

3.2. Anotação humana

A planilha resultante do processo descrito acima foi complementada com duas colunas destinadas à anotação humana. Para essa tarefa, contamos com três anotadores: dois com ampla experiência em tradução inglês–português e análise de linguagem figurada, e um com alta proficiência em língua inglesa, mas sem experiência específica em tradução. Os anotadores avaliaram se a expressão identificada pelo LLM correspondia efetivamente a uma EI (SIM ou NÃO). Além disso, foram incluídas as possibilidades de seleção da categoria INCONCLUSIVA, nos casos em que o contexto disponível não permitia uma decisão definitiva, e da categoria ERRO, utilizada quando o sistema identificava como EI uma sequência de palavras inexistente na legenda em inglês ou em português. Nas legendas classificadas como SIM, os anotadores avaliaram a tradução em português segundo as seguintes categorias: idiomática e convencional; idiomática, mas não convencional; não idiomática, mas convencional; incorreta; e erro. Os resultados dessa anotação serão discutidos na próxima seção.

4. Análise

4.1. Validação das EIs identificadas automaticamente

Conforme explicitado na seção anterior, a avaliação do desempenho do LLM na identificação de EIs e na análise de seus equivalentes tradutórios foi realizada por três

anotadores humanos. O grau de concordância entre os anotadores foi avaliado por meio do coeficiente Kappa, uma medida que estima o quanto a concordância observada excede aquela que seria esperada ao acaso. Na tarefa de identificação de EIs por um LLM, observou-se um nível de concordância moderado, com Kappa de Fleiss igual a 0,426. As concordâncias par a par variaram entre 0,383 e 0,474, sendo os maiores valores registrados entre os anotadores com maior experiência em tradução. Esse resultado reforça a complexidade inerente à delimitação do que constitui uma EI, bem como a subjetividade envolvida na avaliação de aspectos como convencionalidade, literalidade e adequação tradutória.

Além disso, a divergência parcial entre os anotadores evidencia que a identificação de EIs depende de fatores contextuais, culturais e pragmáticos, especialmente em corpora de legendas, nos quais o contexto disponível frequentemente é reduzido ou fragmentado. Observou-se, também, que os dois anotadores com experiência tradutória apresentaram maior convergência na avaliação das estratégias tradutórias. No entanto, são necessárias análises adicionais para que se possam estabelecer conclusões mais consistentes acerca da influência da experiência em tradução na identificação e interpretação de expressões idiomáticas.

Dentre as 928 frases contendo EIs, 772 (83,5%) apresentaram identificação correta dos *tokens* traduzidos (precisão), enquanto 75 (8,1%) foram classificadas como incorretas (falsos positivos) e 7 (0,8%) apresentaram erros de alucinação. Além disso, 71 EIs (7,6%) foram consideradas inconclusivas pelos anotadores. Por se tratar de um estudo preliminar, o foco principal foi na análise de precisão, sendo a ausência de avaliação de *recall* e *F1* uma limitação reconhecida, que será abordada em trabalhos futuros.

Foi possível agrupar os falsos positivos identificados automaticamente pelo LLM em algumas categorias principais:

1. Identificação incorreta de EIs. Por exemplo, o LLM identificou como EI a sequência *fired me* na legenda "*They fired me. Do you think it's related?*", desconsiderando que *me* funciona apenas como complemento pronominal do verbo *fire*, e não como constituinte de uma expressão;

2. Marcadores conversacionais recorrentes classificados como EIs. Algumas expressões frequentes na oralidade foram interpretadas como idiomáticas pelo modelo, embora funcionassem apenas como marcadores discursivos. É o caso de *all right* em "*That was so fun. All right. Here's another one*.";

3. Expressões potencialmente idiomáticas interpretadas como EIs mesmo em usos literais. Em diversos casos, o LLM classificou como idiomáticas expressões convencionalmente associadas a sentidos figurados, ainda que empregadas literalmente no contexto específico. Um exemplo é *white elephant* em "*The white elephant of the King of Siam, the sacred animal…*";

4. Colocações confundidas com EIs. O modelo também identificou como EIs combinações lexicais recorrentes que, embora convencionalizadas, não apresentam necessariamente sentido figurado. Por exemplo, *make a difference* em "*What difference does it make to him?".*

Além disso, os anotadores humanos também observaram que em diversos casos em que

o contexto não é conclusivo para se identificar a combinação de palavras como EI, o LLM o fez, muito possivelmente porque, em geral, essas combinações são, de fato, mais utilizadas em sentido figurado. É o caso, por exemplo, de *throw to the lions* em "*Throw them to the lions!*". Também observaram que combinações de palavras em caixa alta foram identificadas como EIs. Por exemplo, THE END.

4.2. Identificação automática dos equivalentes das EIs

Em relação à identificação dos equivalentes das EIs pelo LLM, os anotadores observaram:

> 1. Traduções idiomáticas e convencionais, ou seja, quando as legendas em português conseguiram manter a idiomaticidade da expressão, utilizando-se de equivalente convencional em português. Por exemplo, "*They can't get away with this*." por "Eles não podem escapar impunes".
>
> 2. Traduções idiomáticas, mas não convencionais. Nesses casos, a tradução preserva o caráter figurado da expressão original, mas recorre a uma combinação pouco usual ou sem ocorrência registrada em corpora de referência. Por exemplo, a legenda "*One good turn deserves another*" foi traduzida como "Favor com favor se paga", expressão sem ocorrência no Corpus do Português NOW (Davies 2016-), mas cuja interpretação pode ser inferida por analogia à expressão convencional "amor com amor se paga".
>
> 3. Traduções convencionais, mas não idiomáticas. Nesses casos, a tradução transmite adequadamente o significado da expressão original, mas sem preservar seu caráter figurado. Por exemplo, *for a start* em "*Well, for a start, you're the prettiest girl I've ever seen.*" foi traduzida como "para começar", solução convencional em português, porém menos idiomática do que alternativas como "pra começo de conversa", que manteria mais explicitamente a informalidade e a idiomaticidade da expressão original.
>
> 4. Traduções inadequadas. Conforme já mencionado, a qualidade das traduções presentes no corpus de legendas é heterogênea, uma vez que, potencialmente, qualquer usuário pode produzir ou sugerir legendas para determinadas obras. Além disso, não se pode descartar a presença de traduções automáticas sem pós-edição adequada. Um exemplo é a legenda em inglês "*Looks like he's not your guy*.", alinhada à legenda em português "parece que ele está não seu cara...", gramaticalmente inadequada em português. Ainda assim, o LLM conseguiu identificar corretamente a sequência "não seu cara" como "equivalente tradutório" da EI presente no original.

Em relação à avaliação das traduções extraídas do *OpenSubtitles*, a concordância geral foi substancialmente menor (Kappa de Fleiss = 0,148), indicando maior subjetividade ou dificuldade na tarefa. Nesse caso, duas das comparações par a par apresentaram valores baixos de Kappa (0,076–0,080), enquanto a maior concordância (0,465) foi novamente observada entre os avaliadores mais experientes, reforçando a influência da experiência na consistência dos julgamentos.

Interessante também observar que, em alguns momentos, o LLM conseguiu identificar o equivalente tradutório ainda que a legenda contivesse erro ortográfico. Por exemplo, "aia e noite", em vez de "dia e noite", como equivalente de *night and day*.

Apesar dessas limitações, os resultados preliminares sugerem que corpora

paralelos de legendas constituem material promissor para estudos sobre fraseologia, tradução e linguagem figurada.

5. Considerações finais

Este trabalho apresentou a compilação, organização e anotação de um corpus paralelo inglês–português extraído do OpenSubtitles v2024 com o objetivo de investigar a identificação automática de EIs em língua inglesa e de seus equivalentes tradutórios em português brasileiro. A partir dos dados apresentados por um LLM e da posterior validação por três anotadores humanos, foi possível observar potencialidades e limitações do uso de LLMs em tarefas envolvendo a linguagem figurada.

Os resultados preliminares indicam que corpora paralelos de legendas constituem material promissor para estudos em fraseologia, tradução e Linguística de Corpus, especialmente por reunirem linguagem coloquial e estruturas próximas à oralidade. A análise exploratória revelou casos em que o modelo identificou EIs e localizou seus equivalentes nas legendas em português, inclusive diante de traduções pouco convencionais ou linguisticamente problemáticas. Por outro lado, foram observados falsos positivos recorrentes, sobretudo em casos envolvendo colocações, marcadores discursivos e expressões potencialmente idiomáticas empregadas em sentido literal. A baixa concordância entre os anotadores na classificação dos equivalentes tradutórios, entretanto, exige cautela na interpretação e generalização desses resultados.

Algumas limitações devem ser consideradas. O estudo concentrou-se nos primeiros 10.000 pares de legendas do subconjunto analisado, procedimento que não constitui amostragem probabilística e não permite excluir possíveis vieses decorrentes da ordenação dos dados. Além disso, como a validação humana incidiu sobre as ocorrências previamente identificadas pelo LLM, o método empregado nessa etapa permite observar falsos positivos, mas não identificar sistematicamente falsos negativos. Outra dificuldade decorre do contexto reduzido das legendas, por vezes insuficiente para desambiguar usos literais e idiomáticos, bem como da qualidade heterogênea das traduções disponíveis no OpenSubtitles. Por fim, a concordância entre os anotadores, particularmente baixa na classificação dos equivalentes tradutórios, evidencia a necessidade de maior estabilização das categorias e dos critérios de anotação.

Como desdobramentos da pesquisa, pretende-se ampliar e diversificar o corpus analisado e refinar as diretrizes de anotação, submetendo-as a uma etapa piloto antes de uma nova avaliação da concordância entre os anotadores. Pretende-se, também, estabelecer um conjunto de referência que permita avaliar falsos positivos e falsos negativos e calcular métricas de desempenho apropriadas, além de comparar diferentes LLMs e estratégias de *prompting*. Finalmente, a incorporação de contexto anterior e posterior às legendas deverá contribuir para a desambiguação entre usos literais e idiomáticos e para a investigação das relações entre idiomaticidade, convencionalidade e qualidade tradutória.

Questões éticas

Os anotadores integram o grupo de pesquisa e, como não houve participação de sujeitos externos nem coleta de dados sensíveis, o estudo não foi submetido ao Comitê de Ética. Etapas futuras que envolvam participantes externos serão previamente submetidas ao Comitê. O ChatGPT 5.5 (OpenAI) foi utilizado como apoio à revisão textual e à

padronização formal do manuscrito, incluindo referências, identificação de inconsistências e sugestões linguísticas e estilísticas. Todas as decisões analíticas e revisões finais permaneceram sob responsabilidade dos autores. Diego Alves foi responsável pela preparação do corpus, elaboração do *prompt* e definição dos índices de análise; Rozane Rebechi, pela coordenação e realização da anotação e redação inicial do artigo; e Isabela Moreira de Oliveira e Cristopher dos Santos Filho, pela anotação e revisão do manuscrito. Todos os autores contribuíram para a versão final.

Referências

Alisoy, H. (2025) A taxonomic approach to structural and semantic dimensions in English phraseology. *Porta Universorum*, 1(4), 72-79.

Biber, D.; Johansson, S.; Leech, G.; Conrad, S.; Finegan, E. (1999) *Longman grammar of spoken and written English*. Harlow: Longman.

Chesterman, A. (1997) *Memes of translation*: the spread of ideas in translation theory. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Davies, M. (2016-) *Corpus do Português: NOW (News on the Web).* Provo, UT: Brigham Young University. https://www.corpusdoportugues.org/now/

Donthi, S.; Spencer, M.; Patel, O. B.; Doh, J. Y.; Rodan, E.; Zhu, K.; O'Brien, S. (2025) *Improving LLM abilities in idiomatic translation. In Proceedings of the Workshop on Language Models for Low-Resource Languages*, [S.l.: s.n.], 2025. p. 175–181.

Dwyer, T. (2012) Fansub Dreaming on ViKi: "Don't Just Watch But Help When You Are Free." *The Translator*, 18(2), 217–243.

Goldschmidt, M. A.; Rodrigues, R. R.; Degasperi, M. H. (2017) O uso de corpora comparáveis bilíngues como subsídio para a tradução de colocações. *Belas Infiéis*, [S.l.], 6(1), 9-24. Biblioteca Central da UNB. https://doi.org/10.26512/belasinfieis.v6.n1.2017.11416

Lison, P.; Tiedemann, J. OpenSubtitles 2016: extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC), 10, 2016. p. 923–929.

O'Hagan, M. (2009) Evolution of User-generated Translation: Fansubs, Translation Hacking and Crowdsourcing. *The Journal of Internationalization and Localization*, 1(1), 94-121.

OpenAI. *gpt-oss-120b / gpt-oss-20b model card*. 2025. https://huggingface.co/openai/gpt-oss-20b

Rebechi, R. R.; Moreira de Oliveira, I.; Tedesco, M. A.; Cardoso, I. F. Is MT the greatest thing since sliced bread? Human and Machine Translation of Food-Related Idioms. *Cadernos de Tradução* (no prelo). https://periodicos.ufsc.br/index.php/traducao/aheadofprintcadernostrad

Rebechi, R. R.; Marcon, N. O.; Faller, G. DE A.; Villavicencio, A. (2024) Machine Translation of Multiword Expressions: Double Dutch or Crystal Clear? *Proceedings of the New Trends in Translation and Technology Conference* - NeTTT 2024. 3-6 July 2024, Varna, Bulgaria, pp. 116-137.

Rebechi, R. R.; Marcon, N. O.; Faller, G. de A. (2025) Tradução automática e tradução humana de expressões multipalavras: descascando esse abacaxi. *Revel*, 23(44), 346-380, 2025. Disponível em: https://www.revel.inf.br/files/2608406c939eac801f1c0b08347cd490.pdf

Rubino, R.; Lapshinova-Koltunski, E.; van Genabith, J. (2016) Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies (NAACL-HLT). San Diego: Association for Computational Linguistics, p. 960–970.

Sag, I. A.; Baldwin, T.; Bond, F.; Copestake, A.; Flickinger, D. (2002) Multiword expressions: a pain in the neck for NLP. *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, 2002. Berlin: Springer, p. 1–15.

Stewart, D. (2000) Conventionality, Creativity and Translated Text: The Implications of Electronic Corpora in Translation. In M. Olohan (Ed.), *Intercultural Faultlines*, pp. 73–91. St. Jerome Pub, 2000.

Tagnin, S. E. O. (2013) *O jeito que a gente diz: combinações consagradas em inglês e português*. São Paulo: Disal.

E se Pellegrino Artusi fosse blogueiro? Tradução funcional de uma receita histórica na era da IA generativa

Stella E. O. Tagnin[1], Rozane R. Rebechi[2]

Universidade de São Paulo[1], Universidade Federal do Rio Grande do Sul[2]

seotagni@usp.br; rozane.rebechi@ufrgs.br

Resumo. *Este estudo investiga a tradução intralinguística de uma receita histórica de Pellegrino Artusi no contexto da formação de tradutores, com base na abordagem funcionalista, e discute o potencial da IA generativa para a reflexão tradutória. Estudantes de Tradução do Inglês adaptaram para um site culinário brasileiro contemporâneo a tradução publicada em português da receita Cuscussù, utilizando a versão inglesa como parâmetro. Foram analisadas adaptações relativas às convenções do gênero, ingredientes e medidas, procedimentos culinários e referências histórico-culturais. Os resultados revelam diferentes graus de intervenção, sobretudo na adequação às convenções de receitas on-line. O cotejo posterior das traduções com apoio do ChatGPT sugere que a IA pode favorecer a reflexão sobre escolhas tradutórias e suas implicações para a finalidade da tradução.*

Abstract. *This study investigates the intralingual translation of a historical recipe by Pellegrino Artusi in translator training, drawing on the functionalist approach, and discusses the potential of generative AI to support translation reflection. English Translation students adapted the published Portuguese translation of Cuscussù for a contemporary Brazilian cooking website, using the English version as a reference. Adaptations concerning genre conventions, ingredients and measurements, cooking procedures, and historical-cultural references were analyzed. The results reveal varying degrees of intervention, particularly in adapting the text to online recipe conventions. A subsequent comparison of the translations with the support of ChatGPT suggests that generative AI can foster critical reflection on translation choices and their implications for the purpose of the translation.*

1. Introdução

Traduzir é sempre uma tarefa complexa e, embora muitos acreditem que "receita qualquer um traduz" [Teixeira, 2004], esse gênero textual mobiliza conhecimentos técnicos, linguísticos e culturais que interferem diretamente nas decisões tradutórias [Rebechi, 2015]. No caso de receitas culinárias históricas, somam-se a essas questões as diferenças temporais entre os contextos de produção e de recepção, que podem envolver mudanças nas práticas gastronômicas, nos ingredientes, nas medidas, nos utensílios e nas próprias convenções do gênero. A receita *46. Cuscussù*, publicada por Pellegrino Artusi em 1891, é especialmente representativa dessa problemática, pois combina instruções culinárias, referências histórico-culturais, anedotas e descrições de práticas e utensílios pouco familiares ao leitor contemporâneo.

A adaptação de um texto dessa natureza para um novo contexto de recepção implica, portanto, decisões que ultrapassam questões relacionadas à equivalência

linguística. Ao transformar uma receita histórica em um texto destinado a um site culinário brasileiro contemporâneo, é necessário considerar, entre outros aspectos, sua exequibilidade para o leitor atual, as convenções do gênero no novo contexto de circulação e o grau de preservação de elementos históricos e culturais do texto de partida. Tais decisões tornam-se especialmente relevantes na formação de tradutores, uma vez que permitem discutir a relação entre finalidade, público-alvo e escolhas tradutórias, bem como os limites da adaptação diante das especificidades do texto de partida.

A crescente presença da inteligência artificial generativa (GenAI) acrescenta outra dimensão a esse cenário. Além de produzir traduções e reformulações, modelos como o ChatGPT podem ser empregados para identificar problemas, propor alternativas e discutir possíveis consequências das escolhas tradutórias. Em contexto pedagógico, interessa, portanto, investigar não apenas as soluções oferecidas por essas ferramentas, mas seu potencial como interlocutoras no processo de reflexão e tomada de decisão, desde que suas sugestões sejam submetidas à avaliação crítica do tradutor em formação [Cf. Penet et al., 2026].

Nesse contexto, esta pesquisa tem como objetivo investigar as principais adaptações realizadas por estudantes na tradução intralinguística [Jakobson, [1959] 2004] da versão publicada em português brasileiro da receita de Artusi (2009), tendo a tradução publicada em inglês [Artusi, 2003] como parâmetro, com base nos princípios da tradução funcionalista [Nord, 2006]. Busca-se, ainda, examinar as possíveis contribuições do ChatGPT para a identificação e a discussão dos problemas tradutórios envolvidos nesse processo, considerando seu potencial como ferramenta de apoio à reflexão sobre as escolhas realizadas.

2. Fundamentação teórica

Este estudo tem como base três principais eixos teóricos. Primeiramente, aborda-se a tradução de receitas culinárias, entendidas como textos fortemente marcados por aspectos culturais e por convenções próprias do gênero, que demandam do tradutor decisões que vão além da busca por equivalentes linguísticos. Em seguida, discute-se a perspectiva funcionalista da tradução, segundo a qual as decisões tradutórias são orientadas pela finalidade comunicativa do texto e pelas necessidades do público-alvo, perspectiva particularmente relevante para a tarefa de tradução analisada neste estudo. Por fim, a discussão volta-se para o uso da inteligência artificial generativa na formação de tradutores, considerando estudos recentes que examinam suas possibilidades, implicações e desafios no contexto pedagógico.

2.1. Tradução de receita culinária

Como sugere o próprio título da pesquisa de Teixeira (2004), *Receita qualquer um traduz. Será?*, a tradução de receitas culinárias é frequentemente percebida como uma tarefa simples, baseada sobretudo na substituição de ingredientes, utensílios e procedimentos por seus equivalentes em outra língua. Essa aparente simplicidade, no entanto, encobre dificuldades decorrentes do caráter culturalmente marcado da culinária. Ingredientes, utensílios, técnicas de preparo e formas de quantificação podem constituir marcadores culturais, vinculados a práticas específicas de determinada comunidade, de modo que a ausência de equivalência direta pode refletir não apenas uma lacuna linguística, mas

diferenças entre os próprios sistemas culinários das culturas envolvidas (Rebechi, 2015; Tagnin e Teixeira, 2004).

Além dos referentes culinários, as próprias convenções do gênero receita podem variar entre culturas. A organização do texto, o grau de detalhamento das instruções, as formas de quantificação e os conhecimentos pressupostos do leitor influenciam a maneira como uma receita é formulada e, consequentemente, traduzida.

Ao comparar receitas brasileiras contemporâneas originalmente escritas em português e em inglês, Rebechi e Silva (2017) identificaram diferenças, entre outros aspectos, no grau de detalhamento das instruções, evidenciando a necessidade de considerar as expectativas e práticas da cultura de chegada. No caso de uma receita histórica como a de Artusi (1891), à distância cultural soma-se a distância temporal: a tradução busca tornar práticas culinárias italianas do final do século XIX acessíveis a leitores brasileiros contemporâneos. Decisões relativas a ingredientes, medidas, utensílios e técnicas de preparo podem, portanto, envolver um conflito entre a preservação das especificidades histórico-culturais do texto de partida e sua adaptação às práticas atuais da cultura de chegada. Essas escolhas dependem, em grande parte, da finalidade atribuída à tradução e de seus destinatários, questão central à abordagem funcionalista discutida a seguir.

2.2. Tradução funcional

A abordagem funcionalista desloca a atenção da busca por equivalência entre os textos de partida e de chegada para a função que a tradução deverá desempenhar na situação comunicativa em que será utilizada. Na proposta de Nord (2006), as decisões do tradutor são orientadas pelo *skopos*, isto é, pelo propósito atribuído ao texto de chegada. Isso significa que um mesmo texto de partida pode dar origem a traduções distintas, igualmente justificáveis, dependendo de seus destinatários e das condições de produção. A função pretendida para o texto traduzido passa, assim, a constituir um critério central para a tomada de decisões tradutórias.

A definição dessa finalidade é explicitada no *brief* de tradução, que especifica aspectos relevantes da situação comunicativa de chegada, como público-alvo, finalidade, meio de circulação, tempo e lugar de recepção. É a partir desses parâmetros que se avalia a adequação das escolhas tradutórias: uma solução é adequada na medida em que contribui para que o texto de chegada cumpra a função prevista. Nessa perspectiva, portanto, adequação não se confunde com proximidade formal ao texto de partida. Em uma tradução de receita destinada à publicação em um site culinário brasileiro contemporâneo, por exemplo, a conversão de medidas, a explicitação de procedimentos ou a substituição de um utensílio pouco familiar podem ser justificadas quando favorecem a execução da receita pelo público previsto.

A orientação para o *skopos*, entretanto, não concede liberdade irrestrita ao tradutor. Nord (2006) complementa o princípio funcional com o conceito de lealdade, entendido como a responsabilidade do tradutor perante os diferentes participantes da interação tradutória, incluindo autor e destinatários. A lealdade estabelece, assim, limites éticos às adaptações realizadas em função do público e da finalidade da tradução. Essa distinção é especialmente pertinente para a tradução de uma receita histórica: tornar o texto de Artusi funcional para leitores brasileiros contemporâneos pode demandar adaptações, mas também requer considerar quais características do texto e de seu contexto

histórico-cultural devem ser preservadas. No presente estudo, o *brief* fornecido aos estudantes — produzir uma tradução funcional da receita para publicação em um site culinário brasileiro contemporâneo — constitui, portanto, o parâmetro a partir do qual suas decisões tradutórias são analisadas.

2.3. IA generativa e formação de tradutores

Os modelos de linguagem de grande escala (*Large Language Models* – LLMs) são sistemas de inteligência artificial treinados com grandes volumes de dados e capazes de identificar padrões linguísticos e gerar textos coerentes, o que possibilita sua utilização em diferentes tarefas relacionadas à linguagem [Naveed et al., 2024]. Entre as aplicações baseadas nesses modelos está o ChatGPT, sistema de inteligência artificial generativa (IAGen) desenvolvido pela OpenAI, que permite ao usuário interagir com a ferramenta por meio de instruções ou *prompts*. No campo da tradução, a crescente incorporação desses sistemas tem suscitado discussões não apenas sobre seu desempenho, mas também sobre as competências requeridas dos tradutores e as implicações éticas de seu uso. Nesse contexto, Alves et al. (2026) destacam a necessidade de uma formação que prepare os futuros profissionais para lidar de forma crítica com questões de responsabilidade, vieses algorítmicos e integridade profissional, desenvolvendo sua capacidade de julgamento diante das novas tecnologias.

A relação entre tradutor e IA não precisa ser compreendida exclusivamente em termos de automação ou substituição. Massey e Ehrensberger-Dow (2026) concebem a IAGen como uma tecnologia colaborativa capaz de complementar a atuação humana, desde que o tradutor mobilize seus conhecimentos, experiência e competências na interação com a ferramenta. A competência tradutória passa, assim, a envolver também a capacidade de formular e reformular *prompts*, analisar de forma crítica as respostas obtidas e decidir sobre sua pertinência, preservando a agência e a responsabilidade humanas sobre o processo tradutório.

No âmbito da formação de tradutores, Angelone (2026) propõe o emprego da IAGen como recurso para a prática deliberada, possibilitando ao aprendiz solicitar *feedback*, identificar dificuldades recorrentes e realizar atividades direcionadas a problemas específicos. Em vez de recorrer à ferramenta apenas para obter uma tradução pronta, o estudante pode utilizá-la para questionar escolhas, explorar e comparar alternativas e refletir sobre possíveis soluções, mantendo-se responsável pela avaliação da adequação das respostas recebidas. Desse modo, a IAGen pode funcionar não apenas como ferramenta de produção textual, mas também como recurso de apoio à reflexão sobre o processo tradutório e à tomada de decisões.

3. Metodologia

A tradução funcional da receita de *cuscussù* compreendeu duas etapas, que serão descritas abaixo.

3.1. Atividade por tradutores em formação

Primeiramente, os alunos da disciplina Tradução do Inglês II, do Bacharelado em Letras da UFRGS, foram apresentados pela professora aos princípios do funcionalismo em tradução [Nord, 2006]. Em seguida, receberam uma tarefa, a ser realizada em duplas ou

trios, cujo objetivo era aplicar esses princípios à tradução intralinguística de uma receita histórica para um contexto brasileiro contemporâneo. O texto selecionado foi a receita Cuscussù (Couscous), de Pellegrino Artusi, publicada originalmente em italiano em La Scienza in Cucina e l'Arte di Mangiar Bene (1891). Para a realização da atividade, foram fornecidas a tradução inglesa [Artusi, 2003] e a tradução brasileira [Artusi, 2009], ambas mais próximas da estrutura e das características do texto histórico.

A tarefa foi orientada por um *translation brief* elaborado especificamente para a atividade. O *skopos* estabelecido foi a publicação da receita em um site culinário brasileiro contemporâneo, tendo como público-alvo leitores não especialistas. A tradução deveria, portanto, desempenhar primordialmente uma função instrucional, ser clara e executável e seguir as convenções de receitas culinárias on-line brasileiras. Os estudantes foram orientados a considerar, entre outros aspectos, medidas, ingredientes, registro, estrutura do gênero e apresentação das instruções. Além da tradução, deveriam comentar o processo, justificando as principais adaptações realizadas a partir da tradução publicada em português, tendo em vista o público e a finalidade estabelecidos no *brief*.

Após a elaboração de uma primeira versão da tradução intralinguística, as soluções propostas foram discutidas coletivamente, com ênfase nas adaptações consideradas necessárias para atender ao *brief*. Os estudantes então revisaram suas traduções nos respectivos grupos, e 17 grupos entregaram uma versão final, acompanhada de comentários sobre suas decisões tradutórias. Para este estudo, foram analisadas essas versões finais, com especial atenção às adaptações relacionadas a: (i) estrutura e convenções do gênero receita; (ii) ingredientes e medidas; (iii) utensílios, técnicas e procedimentos culinários; e (iv) registro e referências histórico-culturais.

3.2 Interação dialógica com o ChatGPT

Em uma etapa exploratória da pesquisa, uma pesquisadora da USP realizou uma interação dialógica com o ChatGPT a partir da receita original de Artusi, em italiano. Diferentemente da atividade realizada pelos estudantes, que produziram uma versão funcional a partir da tradução brasileira publicada, utilizando a versão inglesa como texto de apoio, essa interação não teve como objetivo gerar uma tradução italiano–português. Buscou-se investigar de que maneira a ferramenta poderia atuar como interlocutora na identificação e na problematização de decisões envolvidas na adaptação de uma receita histórica para um contexto culinário contemporâneo.

A interação desenvolveu-se de forma progressiva. Inicialmente, solicitou-se à ferramenta que identificasse possíveis desafios de tradução presentes no texto. A resposta chamou atenção para aspectos como léxico de época e regional, sintaxe e estilo autoral, medidas imprecisas, técnicas culinárias implícitas, ingredientes e práticas alimentares historicamente marcados. A partir desse levantamento inicial, a pesquisadora direcionou o diálogo para problemas específicos, formulando perguntas e selecionando aspectos que considerava relevantes para o objetivo da tradução.

Um dos primeiros tópicos aprofundados foi a preservação do humor e da ironia de Artusi. A discussão concentrou-se, por exemplo, nos versos *Che non è impresa da pigliar a gabbo / Descriver bene questo grande intruglio, / Né da lingua che chiami mamma e babbo*. Em vez de solicitar diretamente uma tradução, a pesquisadora discutiu com a ferramenta a função dos versos no texto e as consequências de diferentes

estratégias. A interação levou à formulação do princípio de que, nesse caso, uma tradução funcional poderia priorizar a preservação do efeito humorístico em detrimento da literalidade, incluindo a possibilidade de traduzir *mamma e babbo* por "mamãe e papai", de modo a procurar reproduzir a infantilização irônica presente no original.

Em seguida, a pesquisadora explicitou que pretendia pensar em uma tradução funcional para os dias atuais, tornando a receita exequível para um leitor não especializado. A partir dessa definição, o diálogo passou a explorar sistematicamente três outros eixos. Para os utensílios, discutiu-se a possibilidade de preservar sua função culinária em vez de sua materialidade histórica, como no caso de *vaso di terra piano e molto largo*, *teglia di rame stagnata* e do sistema formado por *scodella*, *pannolino* e *spago*. Para as medidas, foram discutidas formas como *una presa* e *due dita (di bicchiere) scarse di acqua*, observando em que situações a vagueza poderia ser mantida ou substituída por referências mais adequadas às convenções contemporâneas. Finalmente, para tempos e gestos culinários, a interação explorou a possibilidade de combinar referências temporais com indicadores sensoriais de textura e consistência, especialmente diante de instruções como *macinatelo colla palma della mano per farlo divenir gonfio, grandioso e sciolto*.

A interação foi, assim, conduzida por meio de sucessivos ciclos de pergunta, resposta, seleção e aprofundamento, nos quais a pesquisadora orientava o percurso da discussão e avaliava de forma crítica as alternativas apresentadas. Ao final, os problemas discutidos foram sistematizados em quatro eixos — (i) utensílios; (ii) medidas; (iii) tempos e gestos culinários; e (iv) humor e estilo — que serviram de referência para o cotejo posterior com as soluções encontradas pelos estudantes.

4. Análise

4.1. Traduções humanas

A análise qualitativa das 17 traduções produzidas pelos estudantes revelou diferentes estratégias de adaptação ao *brief* de tradução, que foram agrupadas em quatro categorias principais: estrutura e convenções do gênero, ingredientes e medidas, utensílios e procedimentos culinários, e registro e referências histórico-culturais.

(i) Estrutura e convenções do gênero: as alterações mais recorrentes dizem respeito à adequação da macroestrutura às convenções de receitas culinárias on-line brasileiras. De modo geral, os estudantes reorganizaram o texto de Artusi, originalmente marcado pela alternância entre narrativa, comentários autorais e instruções culinárias, em seções como título, rendimento, ingredientes e modo de preparo, frequentemente com etapas numeradas. Alguns grupos acrescentaram ainda informações como tempo de preparo, grau de dificuldade, utensílios, dicas e informações adicionais. Em uma das versões, aparecem inclusive "Avalie esta receita" e "Deixe seu comentário", recursos característicos da circulação digital do gênero.

(ii) Ingredientes e medidas: as intervenções nesse âmbito foram mais heterogêneas e envolveram tanto explicitação quanto substituição. Ingredientes mencionados apenas durante o preparo no texto de partida, como pão, azeite, sal, pimenta, água e molho de tomate, foram frequentemente incorporados à lista inicial. Alguns grupos também procuraram tornar ingredientes menos familiares mais acessíveis: *coarse-grained semolina* foi traduzida como "sêmola de trigo", "cuscuz de trigo" e, em uma solução mais

domesticadora [Venuti, 1995], "farinha de milho flocada"; a vitela, por sua vez, foi mantida em algumas versões, mas em outras recebeu alternativas como frango, peito bovino ou patinho, mais comuns no contexto brasileiro.

(iii) Utensílios, técnicas e procedimentos culinários: também foram observadas adaptações destinadas a tornar a execução da receita mais clara e viável para o leitor contemporâneo. Utensílios específicos, como *earthenware dish* (tigela de barro) e *copper saucepan* (panela de cobre), foram generalizados como "tigela [...] resistente ao calor", entre outros, e "panela". Em um caso, o complexo sistema descrito por Artusi para cozinhar a sêmola no vapor foi adaptado com o uso de um "escorredor de macarrão", procurando reproduzir a função do utensílio com equipamentos disponíveis em uma cozinha atual. Outras soluções revelam estratégias distintas de explicitação: as almôndegas originalmente descritas como pouco maiores que avelãs foram apresentadas como tendo "mais ou menos 3 cm" em uma tradução e como "bolinhas pequenas (como se estivesse fazendo brigadeiro)" em outra.

(iv) Registro e referências histórico-culturais: por fim, houve uma tendência à simplificação do registro, com substituição de períodos longos e formas verbais menos usuais por imperativos diretos, frases mais curtas e linguagem mais próxima das receitas contemporâneas. As maiores diferenças entre os grupos, entretanto, dizem respeito à voz autoral de Artusi. Enquanto algumas versões preservaram os comentários sobre a origem do prato, as referências a Moisés e Jacó, a passagem bíblica, os versos presentes no original, entre outras digressões do autor, outras condensaram essas informações, transferiram-nas para seções como "Curiosidade", "Saiba mais" ou "Informações adicionais", ou simplesmente omitiram-nas ou as localizaram, como, por exemplo, a reformulação de "hoje em dia" para "no final do século XIX".

4.2. O potencial da interação com a IA para a reflexão tradutória

A comparação das traduções intralinguísticas dos estudantes com as questões e alternativas levantadas na interação posterior com o ChatGPT ampliou a análise de problemas já presentes na tradução publicada em português. Um exemplo encontra-se no preparo da sêmola. A versão publicada traz "dois dedos (de copo) escassos de água" e orienta a moagem do ingrediente "com a palma da mão" até ficar "inchado, grandioso e derretido", formulações pouco convencionais [Tagnin, 2013] para uma receita contemporânea. A versão inglesa já explicita parcialmente o procedimento, com *a few tablespoons of water* e *kneading it with the palm of your hand [...] until [...] loose and granular*. Alguns estudantes realizaram movimento semelhante. Um dos grupos, por exemplo, propôs adicionar água "uma colher de cada vez, massageando com as mãos até que a farinha inche e fique soltinha e granulada". Na interação com o ChatGPT, medidas e gestos culinários foram discutidos como problemas de adaptação funcional, e a ferramenta sugeriu combinar referências quantitativas com indicadores sensoriais de textura e consistência. O cotejo com o italiano — *due dita (di bicchiere) scarse* e *macinatelo colla palma della mano* — evidenciou, contudo, que explicitar o procedimento ou converter a medida implica uma decisão tradutória entre preservar a relativa imprecisão da prática histórica e tornar a instrução mais executável para o leitor contemporâneo.

Em relação aos utensílios e procedimentos, a tradução publicada instrui o leitor a colocar a sêmola em uma "taça de barro plana e muito larga, ou em uma torteira de cobre",

solução bastante próxima do italiano *vaso di terra piano e molto largo* e *teglia di rame stagnata*, enquanto o inglês emprega *a wide, flat earthenware dish*. Diante do *skopos* proposto, os estudantes frequentemente generalizaram esses utensílios, empregando, por exemplo, "tigela grande" e explicitando posteriormente o procedimento de cozimento no vapor, com a orientação de que o recipiente deveria ficar encaixado "como uma tampa, para que o vapor não escape". A própria reflexão de uma estudante registra que "taça de barro" e "torteira de cobre" provavelmente não estariam disponíveis na cozinha do leitor contemporâneo. Na interação com o ChatGPT, a discussão seguiu direção semelhante: diante de utensílios históricos, a ferramenta propôs considerar prioritariamente a função desempenhada por eles no processo culinário e sugeriu sua substituição por recursos contemporâneos capazes de realizar a mesma função.

A interação também abordou as referências histórico-culturais e a voz autoral, aspectos bastante preservados na tradução portuguesa publicada, mas frequentemente reduzidos nas versões funcionais produzidas pelos estudantes. Um dos grupos, por exemplo, retirou os versos presentes no original e reformulou a explicação sobre as restrições alimentares judaicas como "Curiosidade para o leitor", justificando que a passagem literária poderia confundir o público de um site de receitas. Em outra versão, criou-se uma seção intitulada "Sobre a receita", na qual a origem árabe e a tradição judaico-italiana são condensadas antes das instruções culinárias. Na conversa com o ChatGPT, humor, ironia e estilo autoral foram tratados como componentes da receita, e não apenas como elementos acessórios. A discussão dos versos *Che non è impresa da pigliar a gabbo / Descriver bene questo grande intruglio / Né da lingua che chiami mamma e babbo*, por exemplo, levou à consideração de estratégias que privilegiassem a manutenção do efeito humorístico, ainda que com afastamento da literalidade. O cotejo com o original italiano e com a versão inglesa reforçou, portanto, a necessidade de considerar as consequências de estratégias como manutenção, condensação, deslocamento ou omissão desses elementos diante do *brief* estabelecido.

5. Considerações finais

A análise das traduções intralinguísticas produzidas pelos estudantes evidencia diferentes formas de operacionalizar os princípios funcionalistas diante do *brief*, que previa a adaptação da tradução brasileira publicada da receita de Artusi para um site culinário contemporâneo. As intervenções foram mais recorrentes na estrutura e nas convenções do gênero, com a reorganização do texto em seções prototípicas de receitas on-line, mas também envolveram ingredientes e medidas, utensílios e procedimentos culinários, registro e referências histórico-culturais. As soluções encontradas revelam diferentes graus de intervenção: enquanto algumas buscaram sobretudo aumentar a clareza e a exequibilidade da receita, outras aproximaram ingredientes, utensílios e referências das práticas culinárias brasileiras. Os resultados mostram, assim, que a adequação ao *skopos* não conduz a uma solução única, mas exige negociar a função instrucional do novo texto com a preservação das especificidades culinárias, históricas, culturais e estilísticas da receita de Artusi.

A interação com o ChatGPT acrescentou outra dimensão a essa análise. Em vez de ser empregado para gerar uma tradução da receita ou determinar soluções consideradas ideais, o modelo foi utilizado como interlocutor para identificar problemas, explorar alternativas e explicitar possíveis consequências de diferentes decisões tradutórias. A discussão de aspectos como medidas imprecisas, utensílios históricos, procedimentos

culinários, humor e referências histórico-culturais tornou mais visíveis tensões já presentes nas soluções dos estudantes, como precisão e vagueza, funcionalidade e preservação histórica, ou função instrucional e dimensão literária. As sugestões do modelo constituíram, portanto, pontos de partida para a reflexão, cuja pertinência precisou ser avaliada pelas pesquisadoras em relação ao *brief* e aos diferentes textos cotejados.

Nesse sentido, a principal contribuição deste estudo para a discussão sobre IA generativa na formação de tradutores está em deslocar o foco da geração de traduções para a problematização das decisões tradutórias. Mais do que fornecer respostas, a ferramenta pode ser explorada para tornar explícitas alternativas que, de outro modo, poderiam permanecer pouco visíveis e para colocar em discussão os ganhos e as perdas associados a cada escolha. Seu potencial pedagógico reside, portanto, não na substituição do processo decisório do tradutor, mas na criação de um espaço adicional de reflexão em que as sugestões produzidas pelo modelo são confrontadas com o *brief*, o público-alvo, as convenções do gênero e as particularidades históricas e culturais do texto. Essa perspectiva preserva a agência do tradutor e atribui à avaliação crítica das respostas da IA papel central no processo formativo.

Como perspectiva futura, pretende-se incorporar a interação com a IA diretamente à atividade realizada pelos estudantes, após a elaboração de uma primeira versão da tradução. Esse novo método permitirá investigar de forma empírica se — e de que maneira — o diálogo com modelos generativos leva os tradutores em formação a manter, reformular ou justificar suas escolhas e quais tipos de intervenção da ferramenta efetivamente favorecem a reflexão sobre o processo tradutório. Será possível, assim, avaliar de forma mais sistemática a hipótese levantada neste estudo: a de que a contribuição mais relevante da IA para a formação de tradutores pode estar menos em sua capacidade de produzir uma tradução do que em sua utilização para questionar, comparar e justificar decisões tradutórias.

Questões éticas

Os dados analisados nesta pesquisa foram produzidos no contexto de uma atividade regular da disciplina Tradução do Inglês II, do Bacharelado em Letras da UFRGS, como parte das práticas pedagógicas previstas para a formação dos estudantes. Por se tratar de uma análise de produções decorrentes de atividade realizada em sala de aula, sem divulgação da identidade dos estudantes ou de quaisquer informações que permitam sua identificação, o projeto não foi submetido ao Comitê de Ética da universidade. No que se refere ao uso de inteligência artificial generativa, o ChatGPT, da OpenAI, foi empregado como ferramenta de apoio à pesquisa em dois momentos. Primeiramente, foi utilizado em uma interação dialógica de caráter exploratório, com o objetivo de identificar e problematizar aspectos linguísticos, culturais e culinários envolvidos na adaptação funcional da receita histórica. Posteriormente, foi utilizado como apoio ao cotejo e à discussão das soluções tradutórias analisadas. A ferramenta também auxiliou na revisão textual do manuscrito, incluindo a identificação de repetições e a padronização formal das referências. Todas as análises, interpretações e decisões finais permaneceram sob responsabilidade das autoras.

Referências

Alves, D., Esqueda, M. D., & Gonçalves, J. L. V. R. (2026). Ética nos Estudos da Tradução: Das encruzilhadas do passado à mega-encruzilhada da Inteligência Artificial. *Benjaminiana - Revista De Estudos Em Tradução e Imagem*, *2*(1), 69–108. https://doi.org/10.67017/breti.v2i01.77443

Angelone, E. (2026). Generative AI as a facilitator of deliberate practice in translator training. In J. C. Penet, J. Moorkens, & M. Yamada (Orgs.). *Teaching translation in the age of generative AI: New paradigm, new learning?* (pp. 27–47). Language Science Press.

Artusi, P. (1891). *La Scienza in Cucina e l'Arte di Mangiar Bene*. Firenze: Salvadore Landi.

Artusi, P. (2003). *Science in the Kitchen and the Art of Eating Well*. Toronto, Buffalo and London: University of Toronto Press.

Artusi, P. (2009). *A Ciência na Cozinha e a Arte de Comer Bem*. (M. O. Anabela Cristina Costa da Silva Ferreira, Trad.) Salto e Itu, SP: Associação Emiliano Romagnoli Bandeirante.

Jakobson, R. (1959 [2004]) "On linguistic aspects of translation". In: Venuti, L. (ed.). The Translation Studies Reader. London/New York: Routledge. p. 138-143.

Massey, G., & Ehrensberger-Dow, M. (2026). Translation competence in the age of generative AI: Debates, dilemmas, directions. In: J. C. Penet, J. Moorkens, & M. Yamada (Eds.), *Teaching translation in the age of generative AI: New paradigm, new learning?* (pp. 3–25). Language Science Press.

Naveed, H. et al. (2024). A Comprehensive Overview of Large Language Models. *arXiv*:2307.06435v8 [cs.CL] (20 Feb 2024), https://arxiv.org/pdf/2307.06435.pdf

Nord, C. (2006). Loyalty and fidelity in specialized translation. *Confluências – Revista de Tradução Científica e Técnica*, 4, 29–41.

Penet, J. C., Moorkens, J., & Yamada, M. (Eds.). (2026). *Teaching translation in the age of generative AI: New paradigm, new learning?* (Translation and Multilingual Natural Language Processing 25). Language Science Press. https://doi.org/10.5281/zenodo.17580856

Rebechi, R. R. (2015) *A Tradução da Culinária Típica Brasileira para o Inglês: um Estudo sob o Enfoque da Linguística de Corpus*. 2015. 393 f. Tese (Doutorado em Letras)—Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.

Rebechi, R. R.; Silva, M. M. Brazilian recipes in Portuguese and English: the role of phraseology for translation In: Mitkov, R. (ed.) *Computational and Corpus-based Phraseology*, ed.1. Cham: Springer, 2017, p. 102 - 114.

Tagnin, S. E. O. (2013). *O jeito que a gente diz: combinações consagradas em inglês e português*. Disal.

Tagnin, S. E. O.; Teixeira, E. D. (2004) Linguística de Corpus e tradução técnica: relato da montagem de um corpus multivarietal de culinária. *TradTerm*, São Paulo, v. 10, 313-358,

Teixeira, E. D. *Receita qualquer um traduz - será: a culinária como área técnica de tradução*. 2004. Dissertação (Mestrado) – Universidade de São Paulo, São Paulo, 2004.

Venuti, L. (1995). *The Translator's Invisibility*. London & New York: Routledge.

Ecolinguística: uso do IRaMuTeQ na análise do nódulo "animal" em corpus das Ciências Agrárias

Shirlene Bemfica de Oliveira[1]; Mariana Gonzaga Alves da Silva[2]; Deise Prina Dutra[2]; Paula Tavares Pinto[3]; Marjorie Lourdes Almeida Corrêa[2]

Instituto Federal de Minas Gerais[1]; Universidade Federal de Minas Gerais[2]; Universidade Estadual Paulista[3]

Autora correspondente: shirlene.o@ifmg.edu.br

Resumo. *Este estudo adota uma perspetiva linguística sustentável e ecológica para explorar as circunstâncias do uso do nódulo "animal" em um corpus de Ciências Agrárias (Couto, 2007). O nódulo "animal" foi escolhido devido à sua alta frequência no corpus de estudo e por fazer parte dos projetos de produção sustentável no contexto nacional, da segurança alimentar e da conservação ambiental (Santos, 2026). A reconstrução do contexto e significado do nódulo "animal" é apresentada e discutida por meio dos dendrogramas e nuvens de palavras geradas e os resultados apontam que embora a sociedade pressione por sistemas agrícolas integrados e sustentáveis, a linguagem científica presente neste corpus advindo de revistas de alto impacto nas Ciências Agrárias ainda reflete uma narrativa de dominação e utilitarismo.*

Palavras-chave: Ecolinguística; animal; Ciências Agrárias

Abstract. *This study adopts a sustainable and ecological linguistic perspective to explore the circumstances of the use of the node "animal" in a corpus of Agrarian Sciences (Couto, 2007). The "animal" node was chosen due to its high frequency in the study corpus and because it is part of sustainable production projects in the national context, food security, and environmental conservation (Santos, 2026). The reconstruction of the context and meaning of the "animal" node is presented and discussed through the dendrograms and word clouds generated, and the results indicate that although society is pushing for integrated and sustainable agricultural systems, the scientific language present in this corpus from high-impact journals in Agrarian Sciences still reflects a narrative of domination and utilitarianism.*

Keywords: Ecolinguistics; animal; Agrarian Sciences

1. Introdução

No cenário contemporâneo, a pressão por sistemas de produção alimentícia sintonizada com a conservação ambiental tem redirecionado as Ciências Agrárias para abordagens integradas e sustentáveis (Farias, 2021). No campo da Zootecnia, de acordo com o autor, o pastejo animal nesses arranjos deixa de ser uma atividade isolada e passa a desempenhar um papel estratégico no ecossistema agrícola, maximizando a reciclagem de nutrientes finitos na interface solo-planta-animal e diminuindo perdas ambientais. Essa busca por sustentabilidade e eficiência bioeconômica reposiciona o papel do animal nas pesquisas científicas, tornando-o um elo indispensável para a resiliência dos agroecossistemas e para a segurança alimentar global (Farias, 2021). Contudo, essa transição em direção a

uma visão integrada na produção agropecuária reflete e é construída por meio do discurso científico. Sob a perspectiva da Ecolinguística, a linguagem não funciona como um espelho neutro das práticas de campo, mas atua como um sistema vivo cujos padrões de uso moldam nossa relação com os elementos biológicos (Couto, 2007; Poole, 2022). Investigar como as Ciências Agrárias nomeiam, categorizam e articulam o elemento animal em seus textos torna-se fundamental para compreender as ideologias ecológicas que subjazem à pesquisa científica de alto impacto (Bocorny et al., 2022).

Diante disso, este estudo, dentro da perspectiva da Linguística de Corpus, investiga as circunstâncias de uso do nódulo "*animal*" em um corpus especializado de artigos científicos das Ciências Agrárias (CorAgrarian). A partir de análises estatísticas de similitude e de classificação hierárquica descendente (CHD) viabilizadas pelo software Iramuteq, associadas à investigação das linhas de concordância de colocações dominantes como *population*, *breed* e *animal welfare*, objetiva-se mapear as polaridades e as prosódias semânticas que delimitam o discurso agrário. Busca-se, assim, revelar se a linguagem acadêmica caminha em direção à coexistência harmônica e sistêmica exigida pela sociedade ou se permanece ancorada em narrativas históricas de fragmentação e dominação do ser vivo (Hötzel; Machado Filho, 2004).

2. Ecolinguística

Segundo a Associação Internacional de Ecolinguística (2019), a Ecolinguística analisa as relações entre a linguagem e os sistemas que sustentam a vida dentro do meio ambiente. Nessa perspectiva, a "ecologia da linguagem em um corpus" é entendida como as ocorrências das interações entre qualquer palavra ou expressão e seu ambiente de uso. Os conceitos são vistos como construções da ecologia biológica a partir das bases epistemológicas apresentadas nos dados deste estudo, demonstrando que as interações do nódulo "*animal*" dentro do ambiente do corpus, portanto, utilizam o ecossistema, as inter-relações entre as diversas questões e com diferentes polaridades. Nesse sentido, "os fatos da linguagem em sua dinâmica e em suas inter-relações são de interesse tanto para a ecologia quanto para o conceito de linguagem", que permeia as várias formas de interação entre palavras em linhas de concordância, textos ou matrizes de coexistência e permanência da linguagem (Couto, 2007, p. 119). Desta forma, o corpus é compreendido como uma rede de inter-relações da ecologia fundamental da língua, cujo verdadeiro ambiente é a sociedade que a utiliza como código (Haugen, 1972 apud Couto, 2007). No contexto desta investigação, esse ambiente corresponde à comunidade científica das Ciências Agrárias, cuja organização social molda diretamente os processos discursivos e os padrões de interação da linguagem acadêmica.

3. Metodologia

Este artigo analisa o nódulo "animal" no CorAgrarian (Teixeira et al., 2022; Marques, 2024), um corpus especializado composto por 447 artigos científicos originais, publicados em inglês entre 2014 e 2023. Os critérios de seleção incluíram periódicos indexados de alto impacto (*Qualis* A quadriênio 2017-2020) nas áreas de Agronomia, Engenharia Agrícola, Engenharia de Alimentos, Engenharia Florestal e Zootecnia. O tamanho total do corpus é de 2.502.032 palavras, distribuídas uniformemente por seções (Tabelas 1 e 2).

Tabela 1. Composição do CorAgrarian por área

Subcorpus	Nº de artigos	Nº de palavras	Média
Agronomia	87	526639	6053,32
Engenharia Agrícola	90	523245	5813,83
Engenharia de Alimentos	90	459535	5105,94
Engenharia Florestal	90	505454	5616,16
Zootecnia	90	487159	5412,88
TOTAL	**447**	**2502032**	**5597,39**

Fonte: Marques (2024)

Tabela 2. Composição do CorAgrarian por seção

Seção	Nº de textos	Nº de palavras	Média
Resumo	447	346477	775,12
Introdução	447	368222	823,76
Métodos	447	440752	986,02
Resultados	444	501947	1130,51
Discussão	441	479004	1086,18
Conclusão	444	365630	823,49
TOTAL	**2670**	**2502032**	**937,09**

Fonte: Marques (2024)

As análises foram feitas com o apoio do Iramuteq (*Interface de R pour l'Analyse Multidimensionnelle de Textes et de Questionnaires*) que é um *software* gratuito, de código aberto que permite a análise estatística e multidimensional do corpus, revelando suas estruturas e conexões (Mahema, 2022; Ratinaud; Marchand, 2012). Essa ferramenta possibilita a análise prototípica, categorização, identificação de associações e propicia o estudo de termos lexicais convergentes e similares em um corpus textual, pois ele estabelece uma interface com o R. Para exemplificação, as colocações dominantes do nódulo "animal" foram extraídas com o auxílio do AntConc e identificadas por meio de critérios estritos, adotando uma janela de colocação de [3L a 3R].

4. Resultados

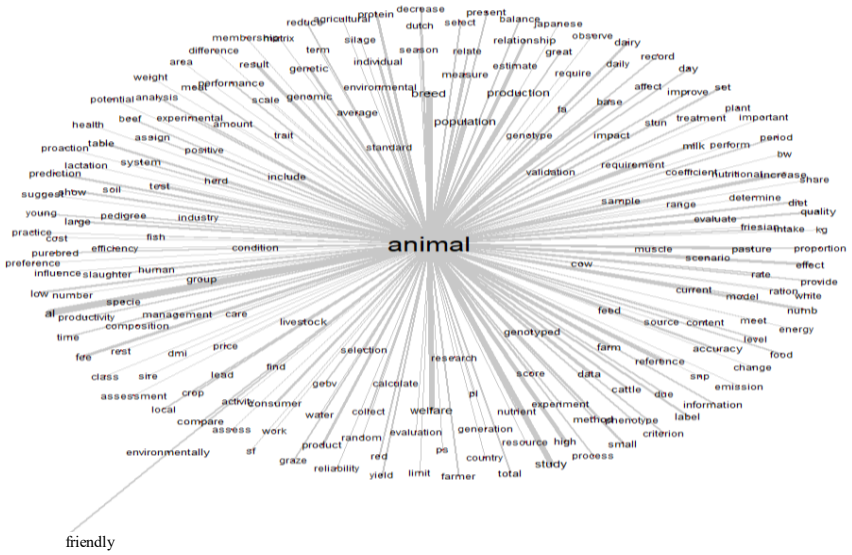
4.1 Da estrutura do subcorpus que apresenta o nódulo animal

A análise do nódulo animal feita automaticamente pelo Iramuteq confirma as escolhas lexicais comuns do grupo de escritores que compõem o subcorpus (ocorrências dentro do subcorpus em que a palavra aparece) e as semelhanças entre os textos que apresentam a palavra. Após a importação, o Iramuteq exibiu uma aba com a descrição completa do subcorpus, apresentando um resumo das estatísticas que oferecem um panorama geral dos dados. O nódulo "animal" apareceu em 244 textos, os quais foram subdivididos automaticamente pelo IRAMUTEQ em 996 segmentos de texto (ST), totalizando 36.656 ocorrências (palavras). Registrou-se uma média de 36,80 palavras por

segmento e 150,22 palavras por texto, com 4.329 formas ativas (palavras de conteúdo) e 1.928 hapax (5,26% das ocorrências).

4.2 Análise da similitude do nódulo "*animal*"

A Análise de Similitude desenvolvida pelo Iramuteq possibilita mapear a ocorrência simultânea entre palavras, indicando a conexão entre elas e auxiliando na identificação da estrutura representada (Camargo; Justo, 2013). O grafo abaixo mostra a representação do nódulo animal no corpus de estudo e revela as redes de significado, a centralidade das ideias e a frequência de association entre os léxicos por meio do tamanho das palavras e da espessura das linhas.



Grafo 1. Nódulo animal no CorAgrarian. Fonte: dados do estudo gerado pelo Iramuteq

A análise da nuvem das formas ativas relacionadas com o nódulo "*animal*" foi feita para revelar prosódias semânticas contrastantes, permitindo compreender como o nódulo "*animal*" se relaciona com as temáticas do corpus (Ratinaud, 2009). Com base na espessura das linhas e na proximidade com o núcleo, a rede se ramifica em três eixos temáticos propostos pelos autores: animal e sua gestão institucional, o mercado dos subprodutos animais e a desconstrução do animal em dados genéticos (Esses eixos serão melhor abordados na análise do dendograma do corpus). É importante ressaltar que o Iramuteq registra todos os colocados semânticos relacionados aos animais e suas respectivas espécies, nomes e tipos. Os termos de maior frequência e de linhas mais espessas são "*population*" (706), "*breed*" (560), e "*welfare*" (157). A seguir, as prosódias semânticas desses termos serão melhor exploradas por meio de suas linhas de concordâncias e exemplos extraídos com o AntConc.

4.2.1 *Population*

O termo "*population*" tem 706 ocorrências, e a análise sintática das linhas de concordância à esquerda são em maioria acompanhados de modificadores nominalizados (*self-sustainable population, reference population, test population, validation*

population) e essas nominalizações funcionam como termos técnicos da área. O padrão sintático à direita se configura como sintagmas preposicionados para especificar espécies de animais (*population of animals, population of wolves, populations of black salmon*). Além disso, os colocados estão relacionados à modelagem genética (*GEBV, SNPs, genotypes, breeds, sires*) e ao impacto espacial (*nearby population, livestock Farms*). Furthermore, the concordance lines below show other common patterns:

cause water supply problems to the nearby	population	and the agricultural areas that are irrigated
regulated by the size of the soil microbial	population	and the amount of fine roots and litter
used to assess sizes of the required reference	populations	and the associated genomic prediction
itself (sectoral), as well as for the nearby	population	and the environment. The methodology is
population is valuable, but unrealistic for small	populations,	and the gain in accuracy may not offset
for predicting accuracy of GEBV in the small	population,	and the prediction accuracy of GEBV in
indicated that both the size of the reference	population	and the relationship between reference
structure software, the current reference	population	and the selected SNPs was successful.
N cycle related genes, bacterial and fungal	populations,	and the soil microbiome, with a focus on

Figura 2. trecho das linhas de concordância de "population". Fonte: dados do estudo.

Nas linhas de concordâncias, "*nearby population*" refere-se a pessoas (comunidades humanas que sofrem com a falta d'água, por exemplo). Em seguida, "population" são fungos "*fungal population*" e micróbios "microbial population" revelando prosódias que variam entre as questões de manejo genético às demandas ecológicas. Além disso, em "*reference população*", o termo animal é uma unidade estatística de produção, como um dado na metodologia científica para o melhoramento, controle bio-estatístico e produtividade, conforme mostra o exemplo (1) abaixo.

> Exemplo (1)
>
> Each cattle breed was divided into **a reference population** and **a test population**. **The test population** was generated by randomly sampling 10 of **the animals** within each breed with a maximum of n 20. **The test population** included 4 Deep Red Cattle, 4 Dutch Belted, 5 Dutch Red and White Friesian, 20 Dutch Friesian, 12 Groningen White Headed and 20 Meuse-Rhine-Yssel (Table 1). **The remaining animals** formed **the reference population** (CorAgrarian.AnS.anbreeding.12_methods_mcl).

Como discutido anteriormente, "population" foca no papel do ser humano ou do ambiente como reguladores e limitadores do crescimento animal trazendo uma visão de controle sobre a densidade e o comportamento ecológico dos animais, como mostra o exemplo (2).

> Exemplo (2):
>
> These different estimates of additive genetic variance reflected variances in different conceptual base **populations**. Therefore, in ideal situations, the data sets used for predicting breeding values should be consistent with the data used for estimating variance components. It indicates that in practical genetic evaluation, variance components used for genetic evaluation should be estimated from the data set that is consistent with the data set for predicting breeding values with regard to conceptual base population. In **polytocous animal population** such as **pig population**, phenotypic data including recent three to five generations and pedigree data adding one to three ancestral generations will lead to

> asymptotic maximum reliability of EBV for **animals** in current generation, dependent on **population** structure (difference in population size and proportions of sires, dams and offspring in this study) and heritability of trait. In general, three generations of records plus two ancestral generations in pedigree are enough for predicting breeding values with regard to accuracy of EBV in **pig population**. (CorAgrarian.AnS.anbreeding.18_conclusion_mcl)

Além disso, os dados trazem o impacto ecológico sofrido por espécies silvestres e polinizadores devido à ação humana ou climática e provocam uma demanda urgente: a preocupação ambiental e a necessidade de diminuir os danos.

> Exemplo (3):
>
> **Salmon populations** have been dwindling since the inclusion of dams on the Klamath River (see Norgaard, 2005). Even adults that are 30-40 years old today have observed noticeable **decline of salmon populations** in the Klamath River. Thus, despite commercialization of the fishery by one tribe the sheer quantity required to provide salmon to all FDPIR clients nationwide is not feasible, yet local tribes would welcome greater access to this culturally significant food. (CorAgrarian.AnS.anhusbandry.04_results_mcl)

4.2.2 *Breed*

Há 560 ocorrências de "*breed*" no corpus de estudo tanto no singular "*breed*", quanto no plural "*breeds*" e para estabelecer análises comparativas intergrupais "*across-breed, multibreed*". Além disso, ele é relacionado ao tipo de técnica usada: "*genomic breeding"* (melhoramento genético; "*cattle / livestock breeding"* (pecuária / criação de gado) associado ao contexto de produção animal; "*traditional / conventional breeding"* (melhoramento convencional) usado para contrastar com métodos genéticos modernos; "*maize / plant / wheat breeding"* (melhoramento vegetal/de milho/de trigo) que foca na agricultura ou na alimentação animal. A figura 3 apresenta algumas das linhas de concordância de "*breed"*:

by randomly sampling 10 of the animals within each	breed	with a maximum of n 20. The test population included
official cattle breed. Both are now registered as one	breed	with an additional notification for colour. Similarly, the
size (Ne), because Ne controls the LD pattern within	breed	However, although we know that the current Ne of
unregistered animals that may belong to the same	breed	However, due to the missing pedigree it is not
is to increase productivity of food animals through	breed	improvement and a scientific approach to balanced
and a scientific approach to balanced feeding. The	breed	improvement programme needs to be supported by
with other SNP reduction techniques to identify	breed	informative SNPs and that PCA is the best technique
dimensionality of large SNP data sets and to identify	breed	informative SNPs (Lewis et al., 2011; Wilkinson et al.,
for which an allocation of an unknown individual to a	breed	is accepted was set to 0.775 (proportion of

Figura 3. Trecho das linhas de concordância de *breed*. Fonte: dados do estudo

Sintaticamente, adjetivos de tamanho e origem geográfica são recorrentes à esquerda do nódulo "*small, local, synthetic, traditional*". Há uma alta incidência de preposições e adjuntos voltados a testes e registros "*purity, registry, composition, purity test*". O termo breed também é associado a "*animal*" com foco estatístico com as palavras "*genomic, prediction, accuracy, selection e reference population"*. Do ponto de vista prosódico, a raça animal é tratada sob a ótica de controle de registros, herança legal e manutenção de

linhagens puras. O animal é validado pelo seu enquadramento institucional, conforme mostra o exemplo (4)

Exemplo (4):

Breed registries have been established for livestock species **to maintain the purity of breeds** and **to document the ancestry of animals**. However, a significant number of **animals** are unregistered with no or incomplete pedigree data and uncertain ancestral **breed origin**. Although many local livestock breeds are "at risk" on the basis of the number of purebred breeding females in a breed registry, there is often also a reservoir of **unregistered animals** that may belong to the same **breed**. However, due to the missing pedigree it is not possible for **breed societies** or herd books to include those **animals** in their **breeding program** for **purebred animals**. (CorAgrarian.AnS.anbreeding.12_abstract_mcl)

Algumas linhas de concordância enfatizam as consequências negativas da substituição de raças por opções mais produtivas, isolando ou colocando as raças nativas em risco de sumirem. Os textos parecem alertar contra a perda de diversidade biológica tradicional. Outras linhas já mostram os benefícios que certas linhagens animais específicas desempenham na manutenção e equilíbrio de biomas tradicionais associando a raça como patrimônio ecológico e cultural indispensável para a sustentabilidade regional.

Exemplo (5):

For implementing the methodology in practice, a rapid and reliable method for **genetic purity testing of animals is needed**, distinguishing **crossbred animals** from **purebred animals** and to determine the **breed composition**. Furthermore, there is genetic variation within **breeds** and consequently a **breed purity test** will depend on how well this genetic variation will be reflected in the **reference populations** dataset. Finally, some **introgression of genes of other breeds** is generally accepted, e.g. **animals** registered with 87.5 pedigree purity are generally considered **purebred**, so the challenge is to determine a threshold value for purity that is generally accepted. (CorAgrarian.AnS.anbreeding.12_introduction_mcl)

4.2.3 *Animal welfare*

Animal welfare aparece no corpus 157 vezes como colocado direto. Diferente de population ou breed, que operam em dimensões estatísticas, demográficas ou de melhoramento genético, animal welfare carrega uma prosódia fortemente regulatória, mercadológica e ética. O animal aqui deixa de ser apenas uma variável biológica e passa a ser o centro de um debate de valores de consumo, normas e políticas públicas.

As linhas de concordância de animal welfare demonstram uma preocupação com a gestão dos sistemas atrelada ao impacto ambiental e ao bem-estar animal, dentro do cenário das demandas por sustentabilidade, certificação e ética.

consumer segments for beef with information on animal	welfare	and environmentally friendly production. Using a choice
contribute to an understanding of whether the animal	welfare	and environmental impact certification labels are
animal production processes with regards to animal	welfare	and environmental protection will increase in the near
In fact, labeling certifications for animal	welfare	and reduced environmental impacts in beef production
strict scrutiny for what relates to sustainability, animal	welfare	and consumer perception of meat quality and safety.
of utmost importance to slaughter facilities both for	welfare	and regulatory reasons. The penetrating captive bolt is the

to produce high quality products (also related to animal	welfare)	and the possible social relationships and interaction with
friendly practices. Attribute information on animal	welfare	and the reduction of environmental impact in the animal
of these government departments for both animal	welfare	and agricultural productivity (see also Productivity
ammonia emissions which cause animal and human	welfare	and air pollution problems, as well as nitrous oxide
motivations: traceability, high quality products, animal	welfare	and fairness for farmers are motivations stressed by
models were used to evaluate the associations between	welfare	and productivity and profitability indicators. The
of beef production methods that promote animal	welfare	and reduce environmental impacts and the heterogeneity
had a positive preference for improvements in animal	welfare,	in agreement with our results concerning beef. Tait,
Organization for Animal Health (OIE) regarding animal	welfare)	in animal production (Tokyo 2020 Games, 2017).
department responsible for workplace health, safety, and	welfare	in England, Scotland, and Wales. The HSE produces
set up diagnostic tests with the aim to identify animal	welfare.	In this way, it will be possible to address
either met or did not meet the criteria. The animal	welfare	standards and on farm assessments that have been recently
or not meeting several criteria from an on farm animal	welfare	assessment and the farms' productivity and

Figura 4: Trecho das linhas de concordância de animal welfare. Fonte: dados do estudo

No padrão sintático de animal welfare à direita ocorre uma repetição de estruturas coordenadas, conectadas pela conjunção and, vinculando o bem-estar animal a critérios ambientais e de produtividade (animal welfare and environmentally friendly; animal welfare and regulatory, animal welfare and agricultural productivity). À esquerda, há uma alta frequência de termos que transformam o nódulo em um objeto de verificação institucional ou preferência (voluntary animal welfare standards, on farm animal welfare assessment, animal welfare preferring consumer). Além disso, animal welfare atua como um modificador adjetival composto para dar qualidade a um substantivo posterior (labels, practices, certification, standards).

O bem-estar animal no exemplo abaixo é visto como um valor humano e um diferencial de mercado que impulsiona o consumo consciente, associado a rótulos de sustentabilidade. Desta forma, o bem-estar como um ativo intangível de qualidade, ética de consumo e responsabilidade social, como mostra o exemplo (6)

> Exemplo (6)
>
> Consumers may view the **animal welfare and environmentally friendly labels** as indicators of food safety and quality (Harper Henson, 1999) because food safety is often linked to **food production methods** (Yee, Yeung, Morris, 2005). Grebitus et al. (2015) reported that human values contributed to a better understanding of consumer choices of **environmentally sustainable foods** relative to traditional sociodemographic explanatory variables. However, the present study does not imply that human values can function as the sole and direct predictors of consumer behaviors. Establishing a direct link between human values and behavior may result in significant but weak relationships because there is often a missing link between values and attitudes. (CorAgrarian.FEg.meat.12_discussion_mcl)

O nódulo "animal welfare" é inserido em um contexto de conformidade legal, auditorias agrícolas e critérios técnicos estritos que precisam ser seguidos e fiscalizados. Os animais passam a ser vistos como protocolos, auditorias, certificados, registros em conformidade jurídica, como no exemplo (7).

> In this respect, **labeling certifications promoting animal welfare** and reduced environmental impacts in beef production systems may be considered additional values by Japanese consumers. If Japanese consumers' concerns for these production methods increase, future perceptions of beef producing farmers on the production of such labeled beef and, thus, beef production methods may change rapidly. (CorAgrarian.AnS.anhusbandry.05_methods_mcl)

Algumas linhas de concordância focam no equilíbrio necessário entre os custos de implantar medidas humanitárias e éticas e o retorno financeiro do produtor. Esta visão é utilitarista e realista, pois pondera sobre a viabilidade econômica do bem-estar animal dentro da cadeia de suprimentos industrial. Outras linhas de concordância são voltadas para contextos em que o bem-estar é negligenciado, gerando problemas graves, tais como: falhas no abate, falhas técnicas no manejo, denúncias, ou falhas políticas de regulação.

> The **animal welfare standards** and on farm assessments that have been recently introduced in North America and Europe are primarily intended to **reassure consumers and food retailers that farm animals are being treated humanely.** Their primary purpose is **not to improve farm productivity and profitability.** Demonstrating the existence of positive associations between animal well being and farm productivity and profitability would provide a great incentive to stimulate dairy producers to invest toward improvements of the comfort and welfare of their animals. (CorAgrarian.AnS.anbreeding.09_discussion_mcl)

4.3 Dendograma do nódulo animal

A análise deste dendrograma (Classificação Hierárquica Descendente - CHD) gerado a partir do nódulo "animal" revela a figura do animal de forma abstrata e instrumentalizada no discurso técnico e científico. O dendrograma revela que, no corpus analisado, o nódulo "animal" não evoca uma narrativa de preservação, agência ecológica ou coexistência harmônica (Hötzel; Machado Filho, 2004). Pelo contrário, reflete uma história de dominação e utilitarismo, dividida em três classes temáticas conforme demonstrado na análise da nuvem de palavras: animal e sua gestão institucional, o mercado dos subprodutos animais e a desconstrução do animal em dados genéticos.

O desdobramento do dendrograma separa a Classe 1 (Vermelha) do bloco formado pelas Classes 2 e 3. Isso demonstra uma divisão conceitual clara na estrutura do corpus entre o espaço e a gestão humana em oposição ao corpo e a produção do animal. No segundo nível de ramificação, o discurso fragmenta o animal em duas dimensões complementares de exploração técnica: a redução genética (Classe 3) e a otimização industrial (Classe 2).

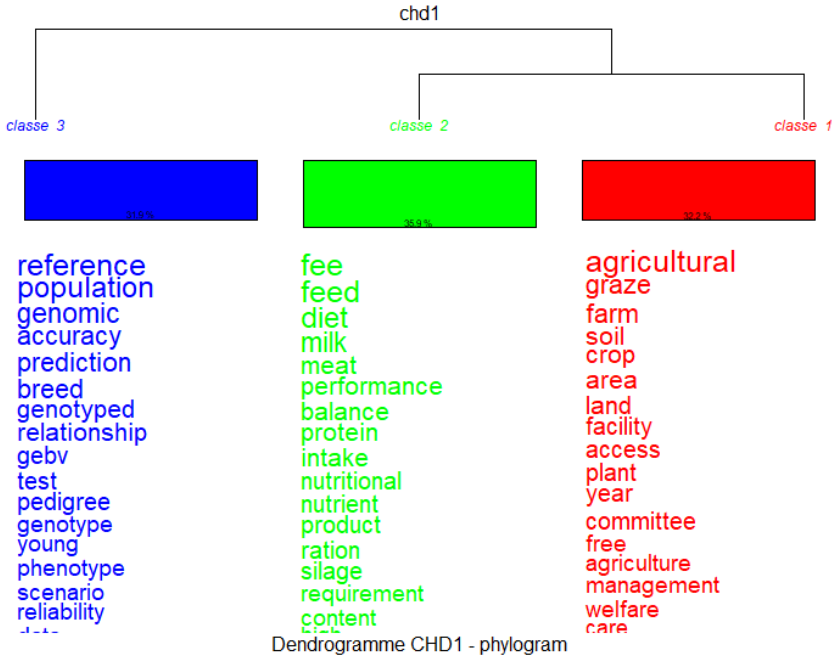


Figura 5: Dendograma do nódulo "animal". Fonte: dados do estudo

Classe 1 (Vermelha): O animal e sua gestão Institucional

As formas ativas mais frequentes são: *agricultural, graze, farm, soil, crop, land, facility, access, plant, committee, management, welfare, care*. Esta classe representa o enquadramento físico e institucional onde o animal é inserido. Termos como land, "soil" e "crop" não aparecem em um contexto de ecossistema livre, mas sim subordinados a "farm," "facility" e "management". A presença de welfare (bem-estar) e care (cuidado) ao lado de committee (comitê) e management (gestão) indica que o cuidado com o animal é mediado por uma lógica burocrática e regulatória humana, tratando a natureza como um espaço gerenciável e antropocêntrico.

Classe 2 (Verde): A otimização industrial e o mercado dos subprodutos animais

As formas ativas mais frequentes são: feed, diet, milk, meat, performance, balance, protein, intake, nutritional, nutrient, product, ration, silage. Esta classe constrói o que a ecolinguística crítica identifica como um modelo industrial ou puramente mercantil do ser vivo. O animal é tratado linguisticamente como uma máquina de conversão econômica na qual entram insumos (feed, diet, protein, nutrient, ration) e saem produtos comerciais (milk, meat, product) medidos pelo seu desempenho (performance). Desta forma, há uma coisificação explícita do animal, cujo valor de existência está estritamente atrelado à produtividade econômica.

Classe 3 (Azul): A desconstrução do animal em dados genéticos

Esta classe ilustra um processo de apagamento e abstração do ser animal. O animal real desaparece para dar lugar a dados estatísticos, códigos de seleção e valores preditivos (gebv, genotype,). A linguagem aqui serve para descolar o indivíduo biológico de sua existência física, transformando-o em um modelo matemático voltado à maximização da eficiência reprodutiva e de linhagem. De modo geral, os resultados indicam uma ênfase

predominante em abordagens éticas e voltadas para o bem-estar no tratamento dos animais, o que orienta a elaboração de materiais didáticos para o ensino de inglês para fins acadêmicos.

5. Conclusão

Este estudo demonstrou que, embora a sociedade pressione por sistemas agrícolas integrados e sustentáveis, a linguagem científica presente neste corpus advindo de revistas de alto impacto nas Ciências Agrárias ainda reflete uma narrativa de dominação e utilitarismo. As análises realizadas com a ferramenta Iramuteq evidenciaram um descompasso discursivo. Enquanto os nódulos "*population*" e "*breed*" reduzem o animal a variáveis técnicas e dados genômicos quantificáveis, a colocação "*animal welfare*" emerge condicionada por certificações, regulamentações e exigências econômicas de mercado. Desta forma, o discurso acadêmico e o desenvolvimento científico caminham mais lentamente do que as demandas ecológicas mundiais.

Referências

BOCORNY, A. E. P.; DUTRA, D. P.; TEIXEIRA, G. L. Mapping the language needs of the Brazilian researchers in the Agrarian Sciences: a corpus-based study of abstracts. In: *Anais... AACL*, 2022.

COUTO, Hildo Honório do. *Ecolingüística*. Estudo das relações entre língua e meio ambiente. Brasília: Thesaurus Editora, 2007.

FARIAS, G. D. *Animal inclusion in purely agricultural systems and anticipation of fertilization in productivity, resource use efficiency and plant nutrient status*. Thesis presented to the Animal Science Research Program as a requirement for obtaining the degree of Ph.D. in Animal Science, at the Faculty of Agronomy, Federal University of Rio Grande do Sul. 2021.

HÖTZEL, M. J.; MACHADO FILHO, L. C. P. Bem-estar Animal na Agricultura do Século XXI. *Revista de Etologia*, 2004, Vol.6, N°1, 03-15.

MAHEMA, K.M. *Análise de dados textuais em pesquisas de mobilidade urbana e transporte com Iramuteq*. Brasília: UnB, 2022.

MARQUES, C. G. de F. CorAgrarian: um corpus de artigos científicos de ciências agrárias. In: *SEvFALE*, Universidade Federal de Minas Gerais, Belo Horizonte, 2024.

POOLE, R. (2022). *Corpus-assisted ecolinguistics*. Bloomsbury Academic.

RATINAUD, P. (2009). *IRaMuTeQ: Interface de R pour les analyses multidimensionnelles de textes et de questionnaires* [Computer software]. Université de Toulouse. http://www.iramuteq.org

RATINAUD, P. MARCHAND, P. Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec Iramuteq. *Actes des 11ème Journées internationales d'Analyse statistique des Données Textuelles*, v. 3, p. 835-844, 2012.

SANTOS, V. S. (n.d.). Biodiversidade. *Brasil Escola*. Retrieved January 30, 2026, from https://brasilescola.uol.com.br/geografia/biodiversidade.htm

The International Ecolinguistics Association. Disponível em: https://www.ecolinguistics-association.org/

ARTIGOS CURTOS
ELC 2026

A inteligência artificial como parceira da escrita acadêmica em inglês: um assistente de feedback baseado em linguística de corpus

Vivian Meurer Lameira, Ana Eliza Pereira Bocorny, Walcir Cardoso

Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil

vilameira@gmail.com, ana.bocorny@gmail.com,
walcir.Cardoso@concordia.ca

Abstract. *Writing research articles in English has become a requirement for Brazilian researchers to take part in the international scientific community, yet less experienced authors are not always familiar with the rhetorical conventions and lexico-grammatical patterns valued by their disciplinary communities. This paper presents INPACT AI, a feedback assistant for academic writing in English that combines Corpus Linguistics with retrieval-augmented generation (RAG) and large language models (LLMs). The tool does not rewrite the user's text: it offers feedback grounded in patterns attested in specialized corpora, organized by field of knowledge and article section. We describe the project's motivation, its theoretical foundations, the architecture of the tool, and its pedagogical rationale.*

Keywords: Corpus Linguistics; academic writing in English; data-driven learning.

Resumo. *A escrita de artigos em inglês tornou-se exigência para a inserção de pesquisadores brasileiros no cenário científico internacional, mas autores menos experientes nem sempre dominam as convenções retóricas e os padrões léxico-gramaticais próprios de cada área. Este trabalho apresenta o INPACT AI, assistente de feedback para a escrita acadêmica em inglês que combina Linguística de Corpus, geração aumentada por recuperação (RAG) e modelos de linguagem de grande porte (LLMs). A ferramenta não reescreve o texto do usuário: oferece feedback ancorado em padrões atestados em corpora especializados, organizados por área do conhecimento e por seção do artigo. Descrevemos a motivação do projeto, seus fundamentos teóricos, a arquitetura da ferramenta e sua proposta pedagógica.*

Palavras-chave: Linguística de Corpus; escrita acadêmica em inglês; aprendizagem direcionada por dados.

1. Introdução

O inglês consolidou-se como a língua franca do mundo acadêmico (Mauranen et al., 2010), desempenhando papel central na disseminação do conhecimento científico (Jenkins, 2014). Pesquisadores que almejam publicar em periódicos internacionais deparam-se, portanto, com a necessidade de redigir seus textos em inglês. Como observa Dayrell (2014), a linguagem produzida por autores menos experientes na escrita

acadêmica em inglês difere daquela encontrada em artigos publicados por pesquisadores experientes, sobretudo porque os autores iniciantes nem sempre estão cientes das convenções do discurso acadêmico, em termos tanto de estrutura retórica quanto de padrões léxico-gramaticais.

Diante desse cenário, os recursos digitais mais difundidos para apoiar a escrita, como os sistemas automatizados de correção textual e os modelos de linguagem de uso geral, tendem a operar pela reescrita do texto do usuário. Embora úteis para ajustes pontuais, tais soluções deslocam a autoria e pouco contribuem para que o pesquisador internalize as convenções da escrita acadêmica em inglês. Na esteira de Swales (1981, 1990), cuja motivação ao propor a análise de movimentos retóricos foi assumidamente pedagógica, entendemos que o apoio à escrita deve favorecer o desenvolvimento da competência do autor, e não substituí-la.

Neste trabalho, apresentamos o INPACT AI, um assistente de feedback para a escrita acadêmica que combina princípios da Linguística de Corpus com técnicas de geração aumentada por recuperação (*Retrieval-Augmented Generation*, RAG) e modelos de linguagem de grande porte (*Large Language Models*, LLMs). Diferentemente de sistemas voltados à correção ou à reescrita automática, a ferramenta recupera evidências linguísticas extraídas de corpora especializados para oferecer feedback sensível às convenções retóricas e fraseológicas de diferentes áreas do conhecimento. O objetivo deste trabalho é descrever a motivação do projeto, seus fundamentos teóricos, a arquitetura da ferramenta e a proposta pedagógica que orienta seu funcionamento.

2. Fundamentação teórica

2.1. A escrita do artigo de pesquisa em inglês

Escrever um artigo de pesquisa não se reduz a transpor ideias de uma língua para outra. O artigo científico é um gênero altamente convencionalizado, cujas seções (introdução, métodos, resultados e discussão) organizam-se segundo expectativas retóricas relativamente estáveis, que variam conforme a área do conhecimento (Hyland, 2008). O domínio desses padrões é particularmente desafiador para falantes de inglês como língua adicional[37], razão pela qual diversos materiais e cursos têm sido desenvolvidos com esse público em vista (Swales e Feak, 1994; Morley, 2015). Oferecer orientação específica para cada seção do artigo, e não apenas para o texto em geral, é especialmente relevante para autores menos experientes na escrita acadêmica em inglês.

2.2. Movimentos retóricos e linguagem formulaica

Swales (1990) descreveu a organização retórica dos gêneros acadêmicos por meio da noção de movimentos, sistematizada no modelo *Create a Research Space* (CARS) para introduções de artigos, em que o autor estabelece um território, indica uma lacuna e ocupa esse espaço ao anunciar seus objetivos. A proposta foi ampliada para outras seções (Swales, 2004) e revisada do ponto de vista metodológico, com vistas a aproximar função e forma linguística (Moreno e Swales, 2018).

[37]O foco deste projeto é a produção de artigos acadêmicos em língua inglesa. Não há previsão de implementação da ferramenta em qualquer outra língua.

Reconhece-se, há muito, que a fluência na escrita depende, em grande medida, do uso de sequências pré-fabricadas de palavras (Wray, 2002). A Linguística de Corpus oferece instrumentos para descrever empiricamente esses padrões (Sinclair, 1991) e, no contexto brasileiro, Berber Sardinha (2004) consolidou os fundamentos da área. Entre as unidades mais estudadas estão os blocos formulaicos, ou *lexical bundles*, sequências recorrentes que cumprem funções discursivas específicas (Biber et al., 2004) e que variam conforme a disciplina (Hyland, 2008). Cortes (2013) relacionou essas unidades a movimentos retóricos, ao associar, por exemplo, a sequência *the purpose of this study is to* à função de anunciar objetivos. Investigações mais recentes voltaram-se às estruturas lexicais descontínuas, ou *lexical frames*, que organizam o discurso em torno de lacunas variáveis (Gray e Biber, 2013).

2.3. Aprendizagem direcionada por dados

O reconhecimento desses padrões abriu caminho para a aprendizagem direcionada por dados (*data-driven learning*), na qual o aprendiz consulta diretamente exemplos autênticos extraídos de corpora (Charles, 2007; Crosthwaite, 2019). Estudos que avaliaram tecnologias dessa natureza relatam efeitos positivos sobre a aprendizagem de gêneros (Cotos et al., 2017). No contexto de língua portuguesa, iniciativas voltadas a falantes de português já se apoiaram no modelo de Swales para orientar a produção de resumos e de introduções, como o sistema SciPo (Feltrim et al., 2006) e o portal Escrita Científica, da USP de São Carlos, aos quais se somam, mais recentemente, plataformas gratuitas de apoio à escrita assistidas por inteligência artificial[38]. Nessa perspectiva, a evidência de corpus funciona como andaime para a produção do texto (Wood, Bruner e Ross, 1976): o aprendiz observa o padrão em uso, decide sobre seu emprego e assume progressivamente o controle da escrita.

3. Metodologia

O INPACT AI converte as evidências discutidas na seção anterior em feedback acionável durante a escrita, apoiando-se em dois pilares: corpora especializados de artigos publicados e uma arquitetura computacional que recupera, nesses corpora, os padrões pertinentes ao texto em produção.

3.1. Arquitetura

A ferramenta combina LLMs com a técnica de RAG. Os LLMs, baseados na arquitetura Transformer (Vaswani et al., 2017), tornaram-se a base dos sistemas atuais de geração de texto; empregados isoladamente, contudo, tendem a produzir respostas plausíveis sem garantia de fidelidade às fontes. A RAG, proposta por Lewis et al. (2020), atenua essa limitação ao acoplar ao modelo uma etapa de recuperação de documentos pertinentes, que passam a fundamentar a resposta gerada. No INPACT AI, os documentos recuperados provêm do corpus CORHUM (Corpus of Humanities)[39]. Nesta versão, trata-se de uma

[38] Ver https://escritacientifica.sc.usp.br e https://rethorical-project.vercel.app. Acesso em: set. 2026.
[39] Corpus of Humanities (CORHUM): 24.453 textos de seções de artigos (25.499.733 palavras) publicados em inglês entre 2013 e 2023 no periódico de acesso aberto PLOS One, estratificados em quatro seções

base fixa, composta por artigos acadêmicos publicados e organizada por área e por seção, não cabendo ao usuário indicar artigos de referência de seu próprio interesse. Dessa forma, as recomendações apresentadas ao usuário não derivam de texto gerado livremente, mas de padrões efetivamente atestados na literatura da área-alvo, em consonância com a abordagem dirigida por corpus (Biber, 2009).

3.2. Funcionamento

O uso da ferramenta organiza-se em três etapas. Primeiro, o usuário seleciona a seção do artigo que pretende produzir (introdução, métodos, resultados/discussão, conclusão) e a área de conhecimento correspondente[40]. Em seguida, redige o próprio texto em inglês ou cola um trecho já produzido. Por fim, recebe feedback contextualizado, que aponta os movimentos retóricos esperados para aquela seção (Swales, 1990), sugere blocos formulaicos convencionais (Biber et al., 2004) e indica eventuais lacunas na estrutura do texto.

O INPACT AI integra-se a dois recursos do mesmo ecossistema: um curso on-line aberto (INPACT MOOC), voltado ao ensino da escrita acadêmica com base em corpora, e uma ferramenta de consulta autônoma a padrões linguísticos (INPACT TOOL). Em conjunto, os três recursos compõem um ambiente adaptativo de aprendizagem baseado em corpus.

4. Considerações finais e trabalhos futuros

O INPACT AI parte do princípio de que aprender a escrever artigos científicos em inglês envolve mais do que corrigir erros linguísticos: exige compreender como diferentes comunidades disciplinares organizam seus argumentos e utilizam padrões recorrentes de linguagem para construir conhecimento. Ao integrar inteligência artificial e Linguística de Corpus, a ferramenta procura tornar visíveis essas convenções, oferecendo ao usuário feedback fundamentado em evidências extraídas de artigos publicados. Essa decisão alinha-se aos princípios da aprendizagem direcionada por dados (Charles, 2007; Crosthwaite, 2019), segundo os quais o contato com exemplos autênticos favorece o desenvolvimento progressivo da autonomia do aprendiz.

Por se tratar de pesquisa em andamento, a ferramenta ainda não foi avaliada junto a seus usuários. Como continuidade deste trabalho, pretendemos investigar em que medida o feedback oferecido favorece o desenvolvimento da escrita acadêmica e a apropriação dos padrões retóricos e fraseológicos característicos de diferentes áreas do conhecimento[41]. Esperamos, assim, contribuir para a formação de autores mais autônomos e para a internacionalização da produção científica brasileira.

(resumo, introdução, resultados e discussão, conclusão) e em oito áreas: Antropologia; Arqueologia; Linguística, Letras e Artes; Psicologia; Educação; Ciência Política; Direito e Ciências Jurídicas; Economia.

[40] Apenas as seções identificadas nos artigos do corpus CORHUM (introdução, métodos, resultados/discussão e conclusão) serão utilizadas neste estudo.

[41] O sistema será avaliado empiricamente por meio de um desenho experimental composto por oito grupos, que utilizarão, de forma isolada ou combinada, o INPACT MOOC, a INPACT TOOL e o INPACT AI. Essa configuração permitirá avaliar comparativamente a eficácia de cada recurso e de suas integrações no aprimoramento da escrita acadêmica, mediante a coleta e a análise de dados quantitativos bem como de dados qualitativos.

Questões éticas

Este trabalho é de natureza teórico-descritiva e não envolveu a coleta de dados junto a seres humanos; os corpora que servem de base à ferramenta são compostos por artigos científicos publicados. Ferramentas de inteligência artificial foram utilizadas como apoio à redação e à revisão linguística do texto, sob supervisão e responsabilidade dos autores. A descrição da contribuição de cada autor será incluída na versão não anônima do trabalho.

Referências

Berber Sardinha, T. (2004) *Linguística de Corpus*, Manole, Barueri.

Biber, D. (2009) A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing, *International Journal of Corpus Linguistics*, v. 14, n. 3, p. 275-311.

Biber, D., Conrad, S. e Cortes, V. (2004) If you look at…: lexical bundles in university teaching and textbooks, *Applied Linguistics*, v. 25, n. 3, p. 371-405.

Charles, M. (2007) Reconciling top-down and bottom-up approaches to graduate writing: using a corpus to teach rhetorical functions, *Journal of English for Academic Purposes*, v. 6, n. 4, p. 289-302.

Cortes, V. (2013) The purpose of this study is to: connecting lexical bundles and moves in research article introductions, *Journal of English for Academic Purposes*, v. 12, n. 1, p. 33-43.

Cotos, E., Link, S. e Huffman, S. (2017) Effects of DDL technology on genre learning, *Language Learning & Technology*, v. 21, n. 3, p. 104-130.

Crosthwaite, P. (Ed.) (2019) *Data-driven learning for the next generation*: corpora and DDL for pre-tertiary learners, Routledge, London.

Dayrell, C. (2014) Instructional practices using corpus linguistics, In: Schuster, E., Levkowitz, H. e Oliveira Jr., O. N. (Eds.) *Writing scientific papers in English successfully*: your complete roadmap, hyprtek.com, inc.

Feltrim, V. D., Teufel, S., Nunes, M. G. V. e Aluísio, S. M. (2006) Argumentative zoning applied to critiquing novices' scientific abstracts, In: Shanahan, J. G., Qu, Y. e Wiebe, J. (Eds.) *Computing attitude and affect in text: theory and applications*, Springer, Dordrecht, p. 233-246.

Gray, B. e Biber, D. (2013) Lexical frames in academic prose and conversation, *International Journal of Corpus Linguistics*, v. 18, n. 1, p. 109-136.

Hyland, K. (2008) As can be seen: lexical bundles and disciplinary variation, *English for Specific Purposes*, v. 27, n. 1, p. 4-21.

Jenkins, J. (2014) *English as a lingua franca in the international university*: the politics of academic English language policy, Routledge, London.

Lewis, P. et al. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems*, v. 33, p. 9459-9474.

Mauranen, A., Hynninen, N. e Ranta, E. (2010) English as an academic lingua franca: the ELFA project, *English for Specific Purposes*, v. 29, n. 3, p. 183-190.

Moreno, A. I. e Swales, J. M. (2018) Strengthening move analysis methodology towards bridging the function-form gap, *English for Specific Purposes*, v. 50, p. 40-63.

Morley, J. (2015) *Academic Phrasebank: a compendium of commonly used phrasal elements in academic English*, 2015 enhanced edition, University of Manchester, Manchester, http://www.phrasebank.manchester.ac.uk, acesso em set. 2026.

Sinclair, J. M. (1991) *Corpus, concordance, collocation*, Oxford University Press, Oxford.

Swales, J. M. (1981) Aspects of article introductions, Aston ESP Research Reports, n. 1, *Language Studies Unit*, University of Aston in Birmingham, Birmingham.

Swales, J. M. (1990) *Genre analysis*: English in academic and research settings, Cambridge University Press, Cambridge.

Swales, J. M. (2004) *Research genres*: explorations and applications, Cambridge University Press, Cambridge.

Swales, J. M. e Feak, C. B. (1994) *Academic writing for graduate students: essential tasks and skills*: a course for nonnative speakers of English, University of Michigan Press, Ann Arbor.

Vaswani, A. et al. (2017) Attention is all you need, *Advances in Neural Information Processing Systems*, v. 30, p. 5998-6008.

Wood, D., Bruner, J. S. e Ross, G. (1976) The role of tutoring in problem solving, *Journal of Child Psychology and Psychiatry*, v. 17, n. 2, p. 89-100.

Wray, A. (2002) *Formulaic language and the lexicon*, Cambridge University Press, Cambridge.

Um MOOC informado por corpus para a Escrita Acadêmica em inglês na área das Humanidades: Design, Implementação e Avaliação Preliminar

Ana Eliza Pereira Bocorny, Juliana Monteiro dos Santos Almeida

Universidade Federal do Rio Grande do Sul (UFRGS)

juliana.teacher78@gmail.com, ana.bocorny@gmail.com

Abstract. *This study presents the design, implementation and evaluation of a corpus-informed Massive Open Online Course (MOOC) for teaching academic writing in English in the area of Humanities. Developed from empirical data within the CORHUM corpus (InPACT), the course integrates rhetorical moves with recurrent phraseological patterns typical of the research article genre. We describe the pedagogical design, a pilot implementation in an undergraduate Language and Literature course at a Brazilian federal university, and initial results. The data indicate that the MOOC enhanced students' awareness of form-function relationships in academic writing, utilizing lexical bundles as scaffolding. However, challenges regarding the virtual environment's instructional design emerged, highlighting the need to align corpus pedagogy with information architecture in digital contexts.*

Keywords: Corpus Linguistics; Academic Writing; MOOC.

Resumo. *Este trabalho apresenta o desenvolvimento, a implementação e a avaliação de um Massive Open Online Course (MOOC) informado por corpus para o ensino de escrita acadêmica em inglês nas Humanidades. Estruturado em quatro módulos, o curso integra movimentos retóricos e padrões fraseológicos do gênero artigo científico extraídos do Corpus de Humanidades (CORHUM-INPACT). Descreve-se o design pedagógico, a implementação piloto na graduação em Letras de uma universidade federal brasileira e os resultados iniciais desse teste. Os dados indicam que o MOOC favoreceu a conscientização sobre as relações forma-função na escrita, utilizando lexical bundles como andaimento. Contudo, surgiram desafios no design instrucional do ambiente virtual, evidenciando a necessidade de alinhar pedagogia e arquitetura da informação.*

Palavras-chave: Linguística de Corpus; Escrita Acadêmica em inglês; MOOC.

1. Introdução

A internacionalização da produção científica brasileira tem ampliado a demanda por recursos pedagógicos voltados ao ensino da escrita acadêmica em inglês. Entretanto, de acordo com Hyland (2014), os manuais de escrita acadêmica costumam adotar os modelos de Ciências Exatas e Biológicas como se fossem o padrão para toda a academia, deixando as Humanidades relativamente sub-representadas. Tal cenário evidencia a necessidade de propostas que considerem especificidades retóricas e fraseológicas próprias de diferentes disciplinas acadêmicas.

Nesse contexto, a Linguística de Corpus (LC) oferece contribuições relevantes para o ensino da escrita acadêmica ao possibilitar a identificação de padrões linguísticos convencionalmente usados em gêneros acadêmicos. Entre esses padrões, destacam-se os *lexical bundles*, sequências recorrentes de palavras que desempenham funções discursivas e organizacionais importantes em textos acadêmicos (Biber et al., 2021). Quando articulados à análise de movimentos retóricos proposta por Swales (1990; 2004), esses padrões permitem compreender como escolhas linguísticas contribuem para a realização de funções comunicativas específicas.

Inserido nesse cenário, o projeto InPACT (Internacionalização da Produção Acadêmica com Corpus e Tecnologia) desenvolveu recursos pedagógicos informados por corpus para apoiar pesquisadores brasileiros na produção de textos acadêmicos em inglês. Como desdobramento dessa iniciativa, o presente estudo apresenta a criação e a avaliação preliminar de um *Massive Open Online Course* (MOOC) destinado ao ensino de escrita acadêmica nas Humanidades.

O objetivo deste trabalho é descrever o design do MOOC, discutir sua implementação piloto e analisar resultados iniciais de um teste piloto realizado com 25 estudantes de graduação em Letras. O trabalho foca especialmente no uso de estratégias de andaimento, que são definidas por Boblett (2012), e de padrões fraseológicos como instrumentos de apoio à produção escrita em inglês (Chen e Baker, 2010).

2. Fundamentação teórica

A LC tem sido amplamente utilizada no ensino de línguas para fins acadêmicos devido à sua capacidade de fornecer evidências empíricas sobre o funcionamento da linguagem em contextos especializados. No campo do Inglês para Fins Acadêmicos (IFA), estudos apontam que a observação de padrões recorrentes favorece o desenvolvimento da consciência linguística e retórica dos aprendizes (Dutra; Berber Sardinha, 2024).

Entre os recursos mais relevantes identificados em corpora acadêmicos estão os *lexical bundles*, visto que esses agrupamentos fraseológicos desempenham funções como organização textual, expressão de posicionamento e construção argumentativa (BIBER et al., 2004).

Além da dimensão fraseológica, a escrita acadêmica também pode ser compreendida a partir da análise de gêneros e movimentos retóricos. Swales (1990; 2004) propõe que artigos científicos apresentam estruturas organizadas em movimentos e passos retóricos que respondem a objetivos comunicativos específicos dentro de comunidades discursivas. Em pesquisas posteriores, Yang e Allison (2003) demonstraram como diferentes seções de artigos científicos mobilizam funções retóricas particulares, especialmente nas seções de resultados e discussão.

A integração entre a Linguística de Corpus (LC) e a análise de gêneros tem possibilitado o desenvolvimento de abordagens pedagógicas fundamentadas em dados linguísticos autênticos. Entre essas abordagens, destaca-se a Aprendizagem Direcionada por Dados (Data-Driven Learning — DDL), proposta originalmente por Johns (1991), segundo a qual os aprendizes assumem um papel investigativo, observando exemplos de uso da língua para identificar regularidades léxico-gramaticais e formular generalizações. A DDL compreende diferentes graus de mediação pedagógica e de participação dos estudantes na consulta aos corpora (GILQUIN; GRANGER, 2010; RÖMER, 2011). Em

sua modalidade direta, os próprios aprendizes utilizam ferramentas de consulta para explorar os dados; na modalidade indireta, o professor seleciona e organiza evidências extraídas de corpora, incorporando-as a materiais e atividades pedagógicas, sem que os estudantes necessariamente operem programas de concordância ou acessem diretamente o corpus (BOULTON, 2010). No presente estudo, adotou-se essa segunda modalidade: os 25 estudantes interagiram com materiais pedagógicos elaborados a partir de análises de corpus, por meio dos quais foram orientados a observar e explorar padrões linguísticos recorrentes.

3. Metodologia

A pesquisa caracteriza-se como um estudo exploratório de natureza qualitativa, centrado na criação e implementação piloto de um MOOC para escrita acadêmica em inglês nas Humanidades.

O conteúdo do curso foi desenvolvido a partir do Corpus de Humanidades (CORHUM), compilado no âmbito do projeto InPACT. O corpus reúne aproximadamente 25 milhões de palavras distribuídas em diferentes disciplinas das Humanidades e organizadas conforme seções do gênero artigo científico. A extração dos dados linguísticos foi realizada com auxílio do software Sketch Engine, permitindo identificar *lexical bundles* recorrentes associados a funções retóricas específicas. O quadro a seguir exemplifica alguns *lexical bundles* de uma atividade do MOOC para a escrita da seção de Resultados e Discussão.

Quadro 1. Exemplos de *lexical bundles* (Resultados e Discussão)

Lexical Bundle	Uso
This study attempts to investigate	Comunicar o propósito da pesquisa.
The total number of [...] compared to One-third of the	Relatar os resultados.
results suggest that findings suggest that this suggests that	Comentar sobre os resultados

A elaboração pedagógica do MOOC foi construída a partir de evidências linguísticas extraídas do CORHUM, articulando princípios da Linguística de Corpus, da análise de gêneros acadêmicos e do andaimento instrucional. O curso foi organizado em quatro módulos autoinstrucionais que contemplavam: (1) ativação de conhecimento prévio; (2) explicitação da estrutura retórica; (3) exploração de elementos linguísticos recorrentes; e (4) produção escrita orientada.

O teste foi realizado em uma disciplina de graduação em Letras, na qual 25 estudantes utilizaram o MOOC como suporte para a redação das seções de Metodologia, e Resultados e Discussão de artigos acadêmicos em inglês. Após a utilização do MOOC, os 25 estudantes responderam a dois formulários eletrônicos, via Google Forms,

destinados à avaliação da utilidade pedagógica do recurso e à identificação de possíveis dificuldades relacionadas ao ambiente virtual. Os gráficos a seguir ilustram a categorização das análises das respostas para duas perguntas, as quais tratam de aspectos positivos do MOOC e sugestões de melhorias a esse recurso.

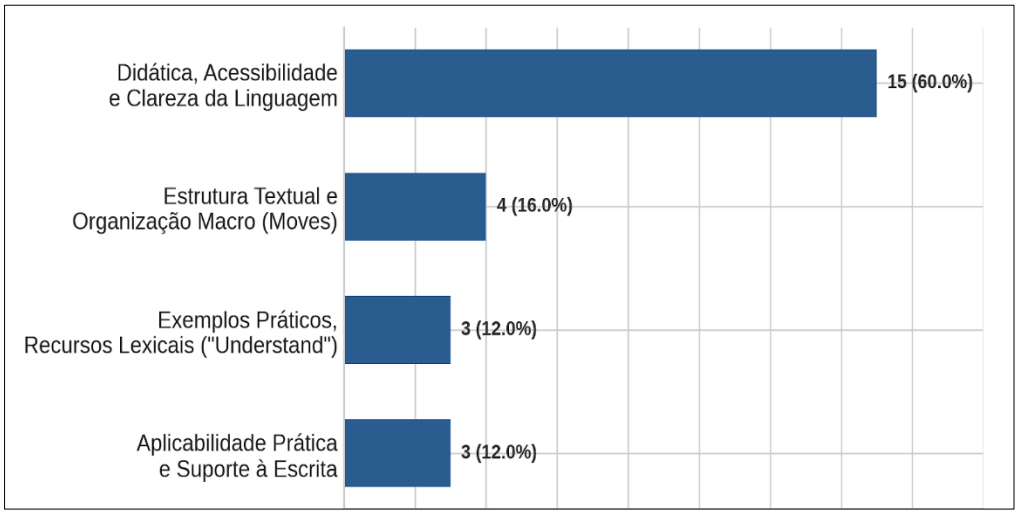


Figura 1. Aspectos Positivos do MOOC

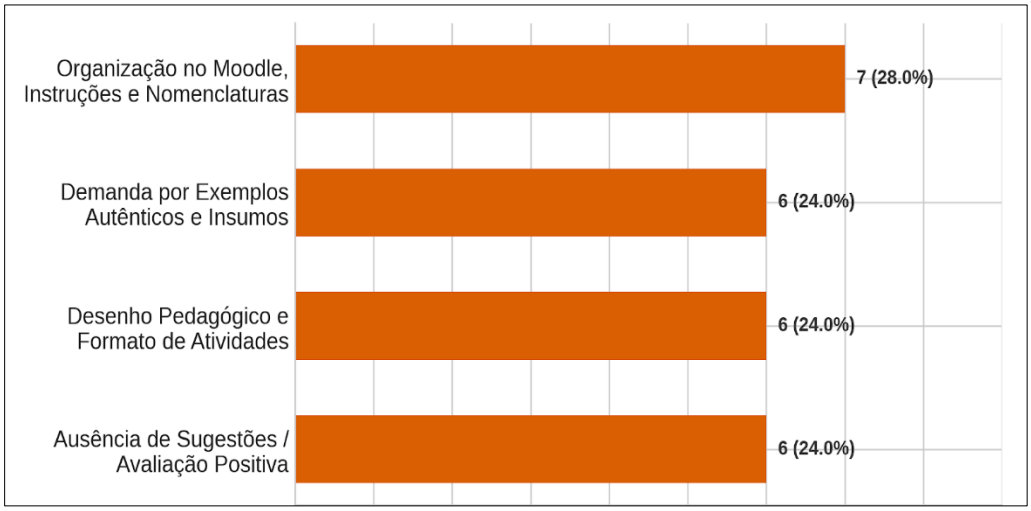


Figura 2. Sugestões de Melhorias ao MOOC

4. Resultados e discussão

Os resultados preliminares indicam que a maioria dos estudantes, 19 (76%), percebeu o MOOC como um recurso útil para apoiar a escrita acadêmica em inglês, enquanto 5 (20%) considerou o recurso parcialmente útil e apenas 1 (4%) não o considerou útil. Entre os aspectos mais valorizados por 15 estudantes (60%), destacaram-se a clareza das explicações, a apresentação de modelos textuais e o suporte fraseológico oferecido pelos *lexical bundles*. Essa avaliação converge com a caracterização dessas sequências como unidades recorrentes que desempenham funções discursivas e organizacionais relevantes nos textos acadêmicos (BIBER; CONRAD; CORTES, 2004). No contexto investigado, portanto, os *lexical bundles* parecem ter funcionado não apenas como repertório linguístico, mas também como uma forma de andaimento: ao disponibilizar padrões

recorrentes associados a funções discursivas específicas, o MOOC ofereceu apoio temporário à realização de uma tarefa de escrita ainda pouco familiar aos estudantes, em consonância com a concepção de andaimento discutida por Boblett (2012).

Os relatos de 7 (28%) dos participantes também sugerem que a articulação entre movimentos retóricos e padrões linguísticos contribuiu para a compreensão das funções comunicativas desempenhadas nas diferentes seções do artigo científico. Esse resultado pode ser interpretado à luz da concepção de gênero de Swales (1990, 2004), segundo a qual a organização dos textos acadêmicos está relacionada aos propósitos comunicativos reconhecidos pelas respectivas comunidades discursivas. Ao associar movimentos retóricos a escolhas fraseológicas recorrentes, o MOOC tornou mais explícita a relação entre o que os autores realizam discursivamente e os recursos linguísticos empregados para essa finalidade.

A contribuição dos exemplos foi particularmente percebida por 6 (24%) dos participantes na elaboração das seções de Resultados. Esse dado é relevante porque tais seções não se limitam à apresentação de informações: elas envolvem operações retóricas como relatar resultados, interpretá-los, relacioná-los a estudos anteriores e formular conclusões, conforme demonstram Yang e Allison (2003). Os exemplos extraídos do CORHUM parecem ter oferecido aos estudantes modelos concretos para reconhecer algumas dessas operações e observar como elas podem ser linguisticamente realizadas. Desse modo, a modelagem textual valorizada pelos participantes não se restringiu à reprodução de estruturas formais, mas envolveu a associação entre padrões fraseológicos e funções retóricas próprias das seções trabalhadas.

Essa percepção das relações entre forma e função aproxima-se dos princípios da Aprendizagem Direcionada por Dados, na qual a observação de ocorrências autênticas permite aos estudantes identificar regularidades e construir generalizações sobre o uso da língua (JOHNS, 1991; CHAMBERS, 2010). Embora os participantes não tenham consultado diretamente o corpus, os materiais elaborados a partir do CORHUM preservaram o princípio de exposição sistemática a evidências empíricas de uso, caracterizando uma modalidade indireta de DDL (BOULTON, 2010). Os resultados sugerem, assim, que a mediação dos dados pelo professor não eliminou seu potencial investigativo: a seleção e a organização pedagógica das ocorrências possibilitaram que estudantes iniciantes observassem padrões linguísticos sem precisar dominar previamente ferramentas de consulta a corpora.

Esses achados também dialogam com Dutra e Berber Sardinha (2024), para quem o trabalho com padrões recorrentes pode favorecer o desenvolvimento da consciência linguística e retórica no ensino de Inglês para Fins Acadêmicos. No caso deste estudo, a associação entre *lexical bundles* e movimentos retóricos parece ter ajudado os estudantes a compreender que as escolhas linguísticas não ocorrem de forma isolada, mas estão vinculadas à organização argumentativa e aos propósitos comunicativos do gênero. Além disso, o emprego de um corpus específico das Humanidades responde, ainda que em escala preliminar, à sub-representação dessas áreas em materiais de escrita acadêmica apontada por Hyland (2014), uma vez que os modelos apresentados foram derivados de práticas discursivas da própria área-alvo.

Os resultados, contudo, também evidenciaram limites na implementação tecnológica e pedagógica da proposta. Dos 25 participantes, 7 (28%) relataram dificuldades de navegação no Moodle, atribuídas ao excesso de informações, à

fragmentação da organização do curso e à dificuldade para distinguir os conteúdos explicativos das atividades práticas. Esses dados não contradizem o potencial pedagógico da abordagem informada por corpus, mas mostram que o acesso aos dados linguísticos e às atividades de observação depende da forma como o percurso de aprendizagem é organizado. Em outras palavras, a qualidade e a autenticidade dos exemplos não garantem, por si sós, que os estudantes consigam utilizá-los produtivamente: é necessário que a arquitetura do ambiente torne claras a sequência das atividades e a função de cada recurso.

A solicitação de simplificação de determinadas terminologias e de ampliação do número de exemplos contextualizados, por 6 estudantes (24%) reforça essa interpretação. A terminologia empregada na análise de gêneros e na Linguística de Corpus pode constituir uma carga adicional para estudantes que estão simultaneamente aprendendo a escrever em inglês e se familiarizando com as convenções do artigo científico. Nesse sentido, a necessidade de mais exemplos não deve ser entendida apenas como uma demanda quantitativa, mas como indicação de que o andaimento precisa ser gradual, articulando explicações conceituais, exemplificação contextualizada e oportunidades de aplicação. O resultado também corrobora a compreensão de que propostas de DDL podem envolver diferentes graus de mediação, definidos de acordo com a experiência e as necessidades dos aprendizes (CHAMBERS, 2010; BOULTON; TYNE, 2013).

De maneira geral, os resultados preliminares apontam para duas dimensões interdependentes do MOOC. Na dimensão linguístico-retórica, os participantes valorizaram a articulação entre movimentos retóricos, exemplos autênticos e *lexical bundles*, o que sugere o potencial do modelo para tornar mais visíveis as relações entre forma e função na escrita acadêmica. Na dimensão instrucional, entretanto, as dificuldades de navegação e compreensão da terminologia indicam que esse potencial depende de uma mediação cuidadosamente graduada e de uma organização mais clara do ambiente virtual. Em vez de confirmar de maneira definitiva a eficácia do modelo, os dados oferecem, portanto, indícios favoráveis à abordagem e identificam condições que precisam ser consideradas em seu refinamento e em avaliações posteriores.

5. Considerações finais

Este trabalho apresentou o desenvolvimento e a avaliação preliminar de um MOOC informado por corpus voltado ao ensino de escrita acadêmica em inglês para estudantes das Humanidades. A proposta buscou integrar evidências de corpus, análise de movimentos retóricos e estratégias de andaimento pedagógico em um ambiente virtual de aprendizagem.

Os resultados iniciais indicam que o curso contribuiu para a conscientização sobre padrões linguísticos e funções retóricas da escrita acadêmica, especialmente por meio da utilização de *lexical bundles* como apoio à produção textual. Ao mesmo tempo, os dados evidenciaram desafios relacionados à organização do ambiente virtual e ao sequenciamento das atividades pedagógicas.

Esta aplicação teve caráter preliminar e formativo, destinando-se ao aperfeiçoamento pedagógico e tecnológico do MOOC. Uma avaliação sistemática de seus efeitos sobre a escrita dos estudantes será realizada em etapa posterior, mediante projeto específico e após a correspondente apreciação ética.

Referências

BIBER, Douglas; CONRAD, Susan; CORTES, Viviana. Lexical bundles in university teaching and textbooks. *Applied Linguistics, Oxford*, v. 25, n. 3, p. 371-405, 2004. DOI: https://doi.org/10.1093/applin/25.3.371.

BIBER, Douglas; JOHANSSON, Stig; LEECH, Geoffrey; CONRAD, Susan; FINEGAN, Edward. *Grammar of spoken and written English*. Amsterdam: John Benjamins, 2021.

BOBLETT, Nancy. Scaffolding: Defining the metaphor. *Studies in Applied Linguistics and TESOL*, New York, v. 12, n. 2, p. 1-12, 2012.

BOULTON, Alex. Data-driven learning: taking the computer out of the equation. *Language Learning*, Ann Arbor, v. 60, n. 3, p. 534-572, 2010. DOI: https://doi.org/10.1111/j.1467-9922.2010.00566.x.

BOULTON, Alex; TYNE, Henry. Corpus linguistics and data-driven learning: A critical overview. *Bulletin suisse de linguistique appliquée*, Neuchâtel, v. 97, p. 97-118, 2013.

CHAMBERS, Angela. What is data-driven learning. In: O'KEEFFE, Anne; McCARTHY, Michael (org.). *The Routledge handbook of corpus linguistics*. London: Routledge, 2010. p. 345-358.

CHEN, Yu-Hua; BAKER, Paul. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, Honolulu, v. 14, n. 2, p. 30-49, 2010. DOI: https://doi.org/10.64152/10125/44213.

DUTRA, Deise Prina; BERBER SARDINHA, Tony. The role of Corpus Linguistics in EAP. In: SARMENTO, Simone et al. (org.). *English for Academic Purposes*. Porto Alegre: Zouk, 2024. p. 15-38.

GILQUIN, Gaëtanelle; GRANGER, Sylviane. How can data-driven learning be used in language teaching? In: O'KEEFFE, Anne; McCARTHY, Michael (org.). *The Routledge handbook of corpus linguistics*. London: Routledge, 2010. p. 359-370.

HYLAND, Ken. English for academic purposes. In: HYLAND, Ken; WONG, Lillian L. C. (org.). *The Routledge companion to English studies*. London: Routledge, 2014. p. 392-404.

JOHNS, Tim. Should you be persuaded: two samples of data-driven learning materials. *English Language Research Journal*, Birmingham, v. 4, p. 1-16, 1991.

RÖMER, Ute. Corpus Research Applications in Second Language Teaching. *Annual Review of Applied Linguistics, Cambridge*, v. 31, p. 205-225, 2011. DOI: https://doi.org/10.1017/S0267190511000055.

SWALES, John M. *Genre analysis*: English in academic and research settings. Cambridge: Cambridge University Press, 1990.

SWALES, John M. *Research genres*: Explorations and applications. Cambridge: Cambridge University Press, 2004.

YANG, Ruiying; ALLISON, Desmond. Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, Amsterdam, v. 22, n. 4, p. 365-385, 2003. DOI: https://doi.org/10.1016/S0889- 4906(02)00026-1.

A produção oral acadêmica em inglês: análise de dados extraídos de *corpora* para subsidiar a elaboração de materiais didáticos para o *webinar*

Rafael Leandro Götz, Ana Eliza Pereira Bocorny

Universidade Federal do Rio Grande do Sul (UFRGS)

rafaelleandrogotz@gmail.com, ana.bocorny@gmail.com

Abstract. *This study investigates academic oral production in English, focusing on the webinar genre from a Corpus Linguistics perspective within Applied Linguistics. The increasing internationalization of higher education has intensified the demand for effective oral communication in digital academic contexts, where webinars play a central role. Despite their relevance, academic oral genres remain underexplored when compared to written discourse, particularly in English for Academic Purposes. Grounded in genre analysis and Data-Driven Learning, this research aims to describe the rhetorical organization of webinars. The methodology adopts a mixed-method approach, including corpus compilation, transcription, annotation, and statistical analysis of linguistic features. Preliminary results are available. To date, ten webinars out of a corpus of 53 have been analyzed, totaling 68,941 tokens and 5,450 types, which enabled the identification of six recurrent discourse patterns representative of the genre (M1, institutional contextualization; M2, audience welcome and management; M3, webinar presentation; M4, introduction to the academic presentation; M5, development of the scientific content; and M6, conclusion of the presentation). The rhetorical analysis, based on Swales (2004), revealed a relatively stable rhetorical structure composed of six rhetorical moves.*
Keywords: corpus linguistics; webinar; academic discourse genre.

Resumo. *Este estudo investiga a produção oral acadêmica em língua inglesa, com foco no gênero webinar, sob a perspectiva da Linguística de Corpus, no campo da Linguística Aplicada. A internacionalização do ensino superior tem ampliado a demanda por comunicação oral eficaz em contextos acadêmicos digitais. Apesar de sua relevância, os gêneros orais acadêmicos permanecem menos explorados que os escritos, especialmente no Inglês para Fins Acadêmicos. Fundamentada na análise de gêneros e na Aprendizagem Orientada por Dados, a pesquisa busca descrever a organização retórica dos webinars. A metodologia adota abordagem mista, envolvendo compilação, transcrição e análise de corpus. Resultados preliminares já estão disponíveis. Até o momento, dez webinars, de um total de 53, foram analisados, totalizando 68.941 tokens e 5.450 types. A análise revelou possíveis seis movimentos retóricos recorrentes (M1, Contextualização institucional; M2, acolhimento da audiência; M3, apresentação do webinar; M4, introdução da exposição; M5, desenvolvimento científico; e M6, encerramento da apresentação)., indicando uma organização discursiva relativamente estável e fornecendo subsídios para o desenvolvimento de materiais didáticos baseados em dados autênticos.*

Palavras-chave: linguística de corpus; webinar; gênero discursivo acadêmico.

1. Introdução

A crescente internacionalização do ensino superior tem intensificado a demanda por competências linguísticas voltadas à comunicação acadêmica em língua inglesa, especialmente no que se refere à produção oral *online*. Nesse cenário, o domínio de gêneros orais acadêmicos, como apresentações, seminários e exposições breves, se torna fundamental para a participação efetiva em práticas científicas e educacionais. No entanto, segundo Clennell (1999), Souza (2023) e Castilho (2025), a literatura aponta que os gêneros acadêmicos orais ainda ocupam um espaço secundário em comparação aos gêneros acadêmicos escritos, tanto no ensino quanto na pesquisa em línguas para fins acadêmicos.

Além disso, segundo Morell (2015) e Silva, Brito, Franco, Almeida e Coelho (2020), a produção oral para fins acadêmicos (POFA) envolve não apenas aspectos linguísticos, mas também competências discursivas, pragmáticas e multimodais, o que reforça a complexidade desse objeto de ensino. Nesse sentido, a necessidade de desenvolver abordagens pedagógicas fundamentadas empiricamente se torna evidente, especialmente aquelas que considerem dados autênticos da língua em uso.

A Linguística de *Corpus* (LC), de acordo com Biber e Reppen (2015), surge como um campo particularmente promissor nesse contexto, ao possibilitar a análise sistemática de grandes volumes de dados linguísticos reais. A partir dessa perspectiva, é possível identificar padrões léxico-gramaticais e discursivos recorrentes, contribuindo para uma descrição mais precisa da linguagem acadêmica. Essa abordagem favorece a construção de materiais didáticos baseados em evidências, superando práticas intuitivas e promovendo maior alinhamento entre ensino e uso real da língua (McEnery e Hardie, 2012).

O presente estudo se justifica, portanto, pela necessidade de investigar empiricamente o gênero acadêmico oral *webinar*, a partir de dados autênticos, com o intuito de identificar padrões linguísticos recorrentes e compreender sua estrutura retórica. Ademais, ao articular pressupostos da LC, busca-se, considerando a análise de gêneros discursivos (Swales, 2004; Hyland, 2007; Flowerdew, 2012) e a Aprendizagem Orientada por Dados (AOD) (Boulton e Cobb, 2017), elaborar material didático para o ensino do gênero acadêmico em questão. Assim, o estudo busca não apenas descrever o funcionamento linguístico do gênero por meio do alinhamento entre forma linguística e função retórica (Moreno e Swales, 2018), mas também promover sua transposição didática para o ensino de Inglês para Fins Acadêmicos (IFA) (Hyland, 2006).

Tendo em vista o tema apresentado, o problema de pesquisa deste estudo é estruturado por meio das seguintes perguntas norteadoras: quais padrões léxico-gramaticais e retóricos caracterizam o gênero acadêmico oral *webinar* em língua inglesa e de que maneira esses padrões léxico-gramaticais e retóricos podem ser sistematizados e transpostos didaticamente, com base nos princípios da AOD, para a elaboração de materiais pedagógicos voltados ao ensino de Inglês para Fins Acadêmicos? Ademais, o objetivo geral é analisar os padrões léxico-gramaticais e sua relação com a organização retórica do gênero acadêmico oral *webinar* em língua inglesa, no domínio da Linguística Aplicada, com vistas à sua descrição linguística e à elaboração de materiais didáticos para o ensino de inglês para fins acadêmicos.

Ademais, este estudo é fruto de uma tese em fase inicial de desenvolvimento, e os resultados aqui apresentados são preliminares. Portanto, não contemplam, em sua totalidade, os objetivos propostos. Nesta etapa da pesquisa, a estrutura retórica ainda está sendo analisada de forma qualitativa e manual. A análise léxico-gramatical ainda não se iniciou.

2. Referencial Teórico

Os pressupostos teóricos se organizam em dois eixos: o primeiro se fundamenta na LC, voltado à investigação de padrões léxico-gramaticais (*key features*) (Biber, 1988; Biber; Conrad; Reppen, 1998) e retóricos (Flowerdew, 2012), a partir do *corpus* de estudo em comparação com *corpora* de referência; o segundo se baseia na AOD (Boulton; Cobb, 2017), visando à elaboração de materiais didáticos (Bocorny, Freitas e Rebechi, 2024) sobre o gênero *webinar* (Ruiz-Madrid; Fortanet-Gómez, 2017), disponibilizados no InPACT[42].

No eixo da LC, a análise envolve três etapas: (i) identificação de padrões retóricos por meio de movimentos e passos (Gray, 2019; Swales, 2004; Charles, 2007), organizando o *webinar* em introdução, desenvolvimento e encerramento; (ii) identificação de padrões léxico-gramaticais recorrentes por n-gramas, com base em etiquetagem morfossintática (Biber, 1988; Biber; Egbert, 2018) e sua relação com os movimentos retóricos (Biber; Conrad; Reppen, 1998); e (iii) análise de características-chave por meio da chavicidade (Biber; Egbert, 2018; McEnery; Hardie, 2012), comparando com *webinars* do *LAEL Webinars*[43] organizados pelo Programa de Pós-Graduação em Linguística Aplicada (LAEL) da Pontifícia Universidade Católica de São Paulo e pelo conjunto de 30 *webinars* organizado pela *University of Queensland*[44] (estudo) a *corpora* como da *American Association for Applied Linguistics* (AAL)[45], do *American English* (*AM*)[46] e da *Oxford Languages* (referência).

Para essas análises, utiliza-se o *Sketch Engine* (Kilgarriff *et al.*, 2004). Na transposição didática, emprega-se o *AntConc* (Anthony, 2005), possibilitando a criação de atividades baseadas em dados, alinhadas à AOD (Boulton e Cobb, 2017; Bocorny, Freitas e Rebechi, 2024).

2.1. A Linguística de corpus e os padrões linguísticos e retórico para descrever o gênero acadêmico oral *webinar*

Apesar de sua relevância, a POFA ainda é menos explorada no ensino de IFA, sobretudo quando comparada possivelmente à escrita, possivelmente em função de desafios metodológicos relacionados à coleta e análise de dados orais (Clennell, 1999). Ainda assim, trata-se de uma prática central nas comunidades acadêmicas, envolvendo gêneros com propósitos comunicativos específicos e organização retórica relativamente estável

[42] Disponível em: https://www.inpact.net.br/ Acesso em set. de 2026.

[43] Disponível em: https://www.youtube.com/channel/UC9-sGRAJFwaPj0I4U3aR0jA/videos Acesso em: 03 set. 2026.

[44] Disponível em: https://www.youtube.com/@petercrosthwaite8224/videos Acesso em: 03 set. 2026.

[45] Disponível em: https://www.aaal.org/webinar-recording Acesso em: 03 set. 2026.

[46] Disponível em: https://americanenglish.state.gov/ae-webinars Acesso em: 03 set. 2026.

(Swales, 2004; Hyland, 2004). Nesse contexto, a POFA demanda não apenas fluência linguística, mas também a capacidade de estruturar discursos de forma estratégica, articulando linguagem e função comunicativa (Flowerdew, 2012; Moreno e Swales, 2018).

A LC (Biber, 1988) oferece um aporte teórico-metodológico relevante ao possibilitar a análise de dados autênticos e a identificação de padrões léxico-gramaticais recorrentes (Biber, Conrad e Reppen, 1998; Biber e Barbieri, 2007). Esses padrões, quando estatisticamente salientes em um corpus de estudo comparado a um corpus de referência (Biber; Egbert, 1988; Biber, Conrad e Reppen, 1998), são denominados padrões láxico-gramaticais chave. Dentre esses padrões, destacam-se os *key lexical bundles* (*KLBs*), sequências lexicais contínuas e recorrentes, e os *key lexical frames* (*KLFs*), estruturas parcialmente fixas com variação lexical, ambos responsáveis por funções discursivas como organização temática e progressão do discurso (Biber e Conrad, 1999; Gray e Biber, 2013).

Esses recursos fraseológicos não apenas caracterizam o registro, mas também se relacionam diretamente à organização retórica dos gêneros, entendida como a distribuição de movimentos e etapas que estruturam o discurso (Flowerdew, 2012). No caso dos *webinars*, tais padrões podem sinalizar funções como abertura, desenvolvimento e encerramento. Paralelamente, a etiquetagem linguística, incluindo o *Part-of-Speech tagging* (*POS tagging*) (Biber, 1988; Berber Sardinha, 2004; Gray, 2015) e a identificação de movimentos retóricos (Swales, 2004), permite relacionar padrões formais às suas funções comunicativas (Biber, Conrad e Reppen, 1998).

2.2. A AOD, os gêneros discursivos e os materiais didáticos aplicados ao ensino do gênero acadêmico oral *webinar*

Nesse contexto, a AOD (Boulton e Cobb, 2017) orienta a elaboração de materiais para o ensino de IFA, possibilitando identificar padrões léxico-gramaticais, como *KLBs* e *KLFs*, e sua relação com movimentos retóricos (Swales, 2004; Charles, 2007).

Do ponto de vista pedagógico, a AOD promove autonomia e consciência linguística por meio da análise de dados autênticos (Boulton e Cobb, 2017; Römer, 2011). Tais dados podem ser extraídos por meio de ferramentas como o *AntConc* (Anthony, 2005), analisados e categorizados em padrões que poderão ser usados em atividades didáticas. Essas atividades podem incluir preenchimento de lacunas, expansão de frames levando à prática oral guiada (Chambers, 2007; Boulton e Cobb, 2017).

O estudo adota princípios de Bocorny, Freitas e Rebechi (2024), como uso de gêneros relevantes, textos autênticos e seleção baseada em frequência. Fundamenta-se na teoria de gêneros (Swales, 2004; Flowerdew, 2012), na perspectiva sociocultural (Bakhtin, 2010; Vygotsky, 1998; Gibbons, 2015) e no ensino por tarefas (Jacobs e Renandya, 2016), com base em dados de *corpus* (Chambers, 2007).

Como referência didática, utiliza-se o material didático *4 Point Speaking for Academic Purposes: Introduction to EAP* (Lockwood e Folse, 2017), articulando abordagem comunicativa (Tomlinson e Masuhara, 2018). Os materiais serão disponibilizados na plataforma *InPACT*, ampliando o ensino de IFA para a oralidade e contribuindo para práticas pedagógicas baseadas em dados.

3. Procedimentos Metodológicos

O estudo se caracteriza como documental e aplicado (Gil, 2008), com abordagem mista (Creswell e Creswell, 2021), fundamentada na LC (Biber e Reppen, 2015) e na AOD (Boulton; Cobb, 2017), organizando-se desde a construção dos *corpora* até a transposição didática. O *corpus* de estudo é composto pelo conjunto de 53 *webinars* do *LAEL Webinars* e pelo conjunto de 30 *webinars* organizado pela *University of Queensland*, comparados com os *corpora* de referência organizados pelo conjunto de 24 *webinars* da AAL, pelo conjunto de 52 *webinars* do *AM* e pelo conjunto de 11 *webinars* da *Oxford Languages* Sua comparabilidade com o corpus de estudo se justifica por três critérios: (i) pertencem ao domínio da Linguística Aplicada; (ii) compartilham o mesmo gênero discursivo, o *webinar*; e (iii) apresentam o mesmo meio de circulação, caracterizado pela comunicação acadêmica oral em ambiente digital. Esses critérios oportunizam uma comparação mais controlada dos padrões léxico-gramaticais entre os corpora. A análise é realizada via *AntConc* (Anthony, 2005) e *Sketch Engine* (Kilgarriff *et al*., 2004).

Como critérios de seleção dos *webinars*, foram considerados aqueles inseridos na grande área das Humanidades, com recorte nos campos da LC e da Linguística Aplicada, áreas nas quais os pesquisadores possuem maior domínio teórico e maior familiaridade para a compreensão dos conteúdos especializados. Quanto à duração, os *webinars* selecionados variam de aproximadamente 35 minutos a 1 hora e 30 minutos.

Para a etapa preliminar apresentada neste estudo, foram selecionados os dez *webinars* mais recentes disponíveis na página do LAEL no momento da constituição da amostra. Desse modo, os dez *webinars* analisados não constituem uma amostra aleatória nem estratificada, mas uma amostra preliminar não probabilística, definida por um critério cronológico de seleção, correspondente às dez publicações mais recentes disponibilizadas pelo LAEL.

Na identificação e anotação dos movimentos e etapas retóricas, os *webinars* são segmentados manualmente em unidades discursivas menores. A delimitação considera a função comunicativa de cada segmento e sua recorrência no corpus: movimentos correspondem a funções mais amplas e etapas a objetivos mais localizados (Swales, 2004). A consistência da classificação é buscada pela observação sistemática dessas recorrências e pelo confronto com descrições do *webinar* na literatura especializada.

A análise inclui extração de n-gramas e identificação de *KLBs* e *KLFs*, seguida de análise de palavras-chave por log-likelihood (Rayson e Garside, 2000; Brezina, 2018) e de Análise de Características-Chave por *Cohen's d* (Biber e Egbert, 2018). Os resultados são interpretados à luz da teoria de gêneros (Swales, 2004; Hyland, 2004; Flowerdew, 2012) e da LC (Biber, Conrad e Reppen, 1998; Biber e Egbert, 2018).

A transposição didática utiliza *AntConc* (Anthony, 2005), gerando atividades baseadas em dados (Bocorny, Freitas e Rebechi, 2024), inspiradas nas estruturas didáticas do material *4 Point Speaking for Academic Purposes: Introduction to EAP* (Lockwood e Folse, 2017), e disponibilizadas na plataforma InPACT. Quanto aos aspectos éticos, o estudo segue a Resolução CNS nº 510/2016 (Brasil, 2016). Predominantemente documental, utiliza dados públicos, sem coleta de informações sensíveis. A etapa com especialistas envolve participação voluntária mediante Termo de Consentimento Livre e Esclarecido, garantindo anonimato, confidencialidade e uso exclusivo para fins acadêmicos. A seleção considera formação mínima na área, e os participantes podem

desistir a qualquer momento, assegurando os princípios de autonomia, não maleficência e uso responsável dos dados.

4. Resultados Preliminares

Até o momento, foram analisados dez *webinars* dos 53 do LAEL, totalizando 68.941 *tokens* e 5.450 *types*. Nesta etapa, foi realizada exclusivamente a análise da estrutura retórica, conduzida manualmente pelos pesquisadores do estudo, especialistas no campo da Linguística Aplicada e da LC. Ademais, não foi feita nenhuma etapa com outros especialistas em relação à validação dos resultados da estrutura retórica encontrados. Informa-se que os achados obtidos nesta etapa devem ser compreendidos como preliminares e restritos à amostra analisada, não sendo ainda considerados representativos do gênero *webinar* em sua totalidade. A análise retórica (Swales, 2004) sugere, nessa amostra inicial, a ocorrência de seis movimentos retóricos, quais sejam os seguintes: M1, estabelecimento do contexto institucional; M2, acolhimento do contexto institucional; M3, apresentação do *webinar*; M4, introdução da exposição acadêmica; M5, desenvolvimento do conteúdo científico e M6, encerramento da apresentação. Esses resultados apontam para indícios de uma possível regularidade na organização discursiva dos *webinars* analisados. Contudo, sua confirmação depende da ampliação da análise para uma parcela maior do *corpus* analisado.

Referências

Anthony, L. (2005). *AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom*. In *2005 IEEE International Professional Communication Conference Proceedings* (pp. 729–737). IEEE. https://doi.org/10.1109/IPCC.2005.1494244

Bakhtin, M. M. (2010). *Estética da criação verbal* (5ª ed.; P. Bezerra, Trad.). Martins Fontes.

Berber Sardinha, T. (2004). *Linguística de corpus*. Manole.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181–190). Rodopi.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., & Reppen, R. (Eds.). (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139764377

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge University Press.

Bocorny, A. E. P., Freitas, A. L., & Rebechi, R. (2024). From specialized corpus to EAP classroom: Integrating authentic data into materials design. In S. Sarmento, R. Rebechi, & M. Matte (Eds.), *English for academic purposes: Reflections, description and pedagogy* (pp. 55–96). Zouk.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348–393. https://doi.org/10.1111/lang.12224

Brasil. Ministério da Saúde. Conselho Nacional de Saúde. (2016). *Resolução nº 510, de 7 de abril de 2016*. https://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Castilho, A. (2025). Língua falada e teorização linguística. *Cadernos de Estudos Linguísticos, 67*, e025047. https://doi.org/10.20396/cel.v67i00.8680378

Chambers, A. (2007). Integrating corpora in language learning and teaching. *ReCALL, 19*(3), 249–251. https://doi.org/10.1017/S0958344007000134

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing. *Journal of English for Academic Purposes, 6*(4), 289–302. https://doi.org/10.1016/j.jeap.2007.09.009

Clennell, C. (1999). Promoting pragmatic awareness and spoken discourse skills with EAP classes. *ELT Journal, 53*(2), 83–91.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*, 75–95.

Creswell, J. W., & Creswell, J. D. (2021). *Projeto de pesquisa: Métodos qualitativo, quantitativo e misto* (5ª ed.). Penso.

Flowerdew, L. (2012). *Corpora and language education*. Palgrave Macmillan.

Gibbons, P. (2015). *Scaffolding language, scaffolding learning: Teaching English language learners in the mainstream classroom* (2nd ed.). Heinemann.

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social* (6ª ed.). Atlas.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics, 18*(1), 109–136. https://doi.org/10.1075/ijcl.18.1.08gra

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story*. John Benjamins.

Gray, B. (2019). Tagging and counting linguistic features for multi-dimensional analysis. In T. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis* (pp. 11–27). Bloomsbury.

Hyland, K. (2004). *Genre and second language writing*. University of Michigan Press.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing, 16*(3), 148–164. https://doi.org/10.1016/j.jslw.2007.07.005

Hunston, S., & Francis, G. (2000). *Pattern grammar*. John Benjamins.

Jacobs, G. M., & Renandya, W. A. (2016). Student-centred learning in ELT. In W. A. Renandya & H. P. Widodo (Eds.), *English language teaching today: Linking theory and practice* (pp. 13–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-38834-2_2.

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–116).

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Lockwood, R. B., & Folse, K. S. (2017). *4 Point Speaking for Academic Purposes: Introduction to EAP*. University of Michigan Press. https://doi.org/10.3998/mpub.9216736

Max Planck Institute for Psycholinguistics. (2024). *ELAN (EUDICO Linguistic Annotator)*. https://archive.mpi.nl/tla/elan

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes, 50*, 40–63. https://doi.org/10.1016/j.esp.2017.11.006

Morell, T. (2015). International conference paper presentations: A multimodal analysis to determine effectiveness. *English for Specific Purposes, 37*, 137–150. https://doi.org/10.1016/j.esp.2014.10.002

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *The Workshop on Comparing Corpora* (pp. 1–6). Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117730

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519–549.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205–225. https://doi.org/10.1017/S0267190511000055

Ruiz-Madrid, N., & Fortanet-Gómez, I. (2017). An analysis of multimodal interaction in a webinar. In C. Vargas-Sierra (Ed.), *Professional and academic discourse* (pp. 274–282). EasyChair.

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Silva, B. A., Brito, D. M., Franco, A. P. C. L., Almeida, M. E. G., & Coelho, S. F. (2020). Produção textual acadêmica: receptividade de graduandos com os gêneros multimodais "análise de dados" e "apresentação em vídeo". *Texto Livre, 13*(2), 1–16. https://doi.org/10.35699/1983-3652.2020.24393

Souza, J. W. C. (2023). Criação do CODiC - Corpus Oral de Divulgação Científica: considerações linguísticas e metodológicas. *Porto das Letras, 9*(2), 243–267. https://sistemas.uft.edu.br/periodicos/index.php/portodasletras/article/view/16223

Swales, J. (2004). *Research genres*. Cambridge University Press.

Tomlinson, B., & Masuhara, H. (2018). *The complete guide to materials development*. Wiley Blackwell.

Vygotsky, L. S. (1998). *A formação social da mente: O desenvolvimento dos processos psicológicos superiores* (6ª ed.). Martins Fontes.

Redução do espaço acústico em pessoas com esquizofrenia

Vitor Marçal de Oliveira, Bruno Neves Rati de Melo Rocha

Universidade Federal de Minas Gerais (UFMG), Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL)

timarcal@hotmail.com, bbruno791@gmail.com

Abstract. *This study investigates acoustic vowel space reduction in people with schizophrenia compared to a healthy control group. The vowels /a/, /i/, and /u/ were analyzed from 16 male speakers with schizophrenia (C-ORAL-ESQ corpus) and 14 controls (C-ORAL-MED corpus), based on F1, F2 measurements, the Formant Centralization Ratio (FCR), and the Phonation-Vowel Index (PHI). Results indicate that the schizophrenia group exhibited significantly higher FCR and lower intercategory dispersion (inter_ms) relative to controls. Additionally, intracategory dispersion (intra_ms) correlated negatively with extrapyramidal symptoms (EPS). These findings support the hypothesis of vowel centralization in schizophrenia and suggest an association with the extrapyramidal side effects of antipsychotic medication.*

Keywords: vowel centralization; schizophrenia; drug-induced parkinsonism; C-ORAL-ESQ; spontaneous speech.

Resumo. *Este trabalho investiga a redução do espaço acústico em pessoas com esquizofrenia em comparação a um grupo controle sadio. Foram analisadas as vogais /a/, /i/ e /u/ de 16 falantes do sexo masculino com esquizofrenia (C-ORAL-ESQ) e 14 controles (C-ORAL-MED), a partir de medidas de F1, F2, do índice de centralização formântica (FCR) e do índice de organização vocálica (PHI). Os resultados indicam que o grupo com esquizofrenia apresentou FCR significativamente maior e menor dispersão intercategorial (inter_ms) em relação ao controle. Além disso, o índice de dispersão intracategorial (intra_ms) correlacionou-se negativamente com os sintomas extrapiramidais (EPS). Os achados sustentam a hipótese de centralização vocálica na esquizofrenia e sugerem associação com efeitos extrapiramidais dos antipsicóticos.*

Palavras-chave: centralização vocálica; esquizofrenia; parkinsonismo medicamentoso; C-ORAL-ESQ; fala espontânea.

1. Introdução

O presente trabalho busca verificar se indivíduos falantes de português brasileiro com esquizofrenia medicados com antipsicóticos apresentam redução do espaço vocálico com relação a um grupo controle sadio. A hipótese de redução do espaço acústico baseia-se no fato de que certos tipos de antipsicóticos (medicamentos usados para o tratamento dos sintomas positivos da esquizofrenia, como delírios, alucinações e comportamento desorganizado) podem provocar os chamados sintomas extrapiramidais (transtornos de movimento involuntários decorrentes do bloqueio dos receptores D2 de dopamina no cérebro, dentre os quais se destaca o parkinsonismo induzido por medicação). O parkinsonismo é caracterizado por rigidez muscular, tremores em repouso e lentidão nos

movimentos, de forma análoga ao observado na Doença de Parkinson. Em indivíduos com Parkinson, especialmente aqueles afetados por disartria, é documentada uma redução do trapézio vocálico – representação acústica bidimensional delimitada pelos formantes F1 e F2 – com relação a controles sadios (Roland *et al*., 2023; Sapir *et al*., 2010). Com base nessa observação, estudos iniciais conseguiram caracterizar a redução do espaço vocálico em indivíduos com esquizofrenia a partir de medidas acústicas correlacioná-la à presença de sintomas extrapiramidais e dosagem de antipsicóticos (Sinha *et al*., 2014; Hogoboom *et al*., 2023).

A centralização do espaço vocálico é investigada neste trabalho a partir de: a. medidas de F1 e F2, b. índice de centralização formântica (FCR) e c. índice do nível de organização do espaço vocálico (PHI), calculado como uma razão entre o índice de separação intercategorial (*inter_ms*, que mede a dispersão dos centroides das categorias vocálicas) e o índice de separação intracategorial (*intra_ms*, que quantifica a variabilidade nas repetições da vogal).

2. Metodologia

2.1. Participantes

O grupo alvo (ESQ) é composto por 16 indivíduos do sexo masculino provenientes do C-ORAL-ESQ (Rocha *et al*., 2026), corpus de fala de pessoas com esquizofrenia coletada durante consultas psiquiátricas no Instituto Raul Soares (FHIEMG) e Hospital das Clínicas da UFMG (EBSERH), em Belo Horizonte. A média de idade foi de 37,2 anos (DP 13,2), com escolaridade média de 8,5 anos (DP 3,2) e renda média de 1,8 salários-mínimos (DP 1,9). As médias de PANSS (Positive and Negative Syndrome Scale) positivo e negativo (Kay *et al*., 1987) foram, respectivamente, 12,5 e 17,5. Todos os participantes estavam em tratamento com antipsicóticos: 8 recebiam apenas fármacos de segunda geração, 7 recebiam combinação de primeira e segunda geração (2 deles com anticolinérgico associado), e 1 tinha apenas o valor de equivalente de clorpromazina registrado sem especificação de prescrição. A média de equivalente de clorpromazina foi de 799 mg/dia (DP 351,1); no subgrupo de 11 participantes com prescrição de antipsicóticos de maior risco extrapiramidal (primeira geração, risperidona ou olanzapina), a média foi de 716,6 mg/dia (DP 398,2). A média de equivalente de benzatropina foi de 23,9 mg/dia (DP 32,1).

O grupo controle é composto por 14 indivíduos do sexo masculino provenientes do C-ORAL-MED (Rocha *et al*., em preparação), corpus de consultas de clínica médica e cardiologia com pacientes sem transtornos mentais, coletadas no Ambulatório da Igreja Nossa Senhora do Carmo, em Belo Horizonte. A média de idade foi de 54 anos (DP 13), com escolaridade média de 8,5 anos (DP 4,5) e renda média de 2,5 salários-mínimos (DP 1,2). Dois participantes previamente selecionados foram excluídos por uso de medicação e comorbidades que poderiam afetar o a produção de vogais e não fazem parte do número total de 14.

A decisão de se utilizar apenas participantes do sexo masculino se deveu à sua maior disponibilidade de dados no momento da realização da pesquisa. Foram selecionados todos os participantes que possuíam avaliação de sintomas extrapiramidais e que produziram número suficiente de realizações de vogais tônicas durante a gravação (cf. 2.2.)

2.2. Procedimentos de coleta e tratamento de dados

Para cada participante, foram identificadas manualmente no software Praat (Boersma 2001) 5 realizações das vogais /i/, /a/ e /u/ com duração igual ou superior a 50 ms, em posição tônica (450 ocorrências no total). A escolha das vogais periféricas justifica-se pela sua maior ocorrência em posição tônica e por permitirem, por si sós, a caracterização aproximada do trapézio vocálico. Os formantes F1 e F2 foram extraídos por meio de um script do Praat que calcula a média formântica da porção central (50%) de cada segmento vocálico anotado em um *interval tier*.

Os dados foram inspecionados manualmente para identificar erros de rastreamento de formantes. Nas ocorrências em que os valores de F1 e F2 apresentaram forte divergência visual com as faixas de energia do espectrograma (sobretudo em /u/ e algumas realizações de /i/) foi feita uma nova extração com ajuste no número máximo de formantes nas configurações de formante, com conferência no sintetizador de vogais do Praat. Em uma ocorrência de /u/ em que esse ajuste não foi suficiente, os valores foram extraídos manualmente, com conferência no sintetizador de vogais.

A partir dos valores de F1 e F2, foram calculados o índice de centralização formântica (FCR, *Formant Centralization Ratio*, Sapir *et a*l., 2010), dado por FCR = (F2_u + F2_a + F1_i + F1_u) / (F2_i + F1_a); e o índice PHI (Roland *et al*., 2023), razão entre o inter_MS (dispersão entre categorias vocálicas) e o intra_MS (variabilidade interna a cada categoria), análoga à estatística F de Fisher em uma ANOVA.

Devido à heterogeneidade de tratamentos farmacológicos e às diferentes dosagens entre os pacientes do grupo SCZ, optou-se em utilizar a dosagem equivalente de clorpromazina (Gardner *et al*., 2010). Essa é uma métrica que permite converter a dosagem de diferentes antipsicóticos em uma única unidade de medida. Por estarem associados a maior risco de efeitos extrapiramidais, foram considerados apenas os antipsicóticos de primeira geração, além de risperidona e olanzapina.

Com base nos valores individuais de F1, F2, FCR, PHI, inter_ms, intra_ms e idade, foram realizadas comparações entre grupos e associações com variáveis clínicas. Para o grupo SCZ, as variáveis clínicas consideradas foram a pontuação na Escala Simpson-Angus (Simpson & Angus 1970), o equivalente de clorpromazina e o tempo de doença.

3. Resultados

3.1. Comparação entre grupos

A Figura 1 apresenta o espaço acústico vocálico por grupo, com ESQ em vermelho e controle em azul, bem como o espaço acústico vocálico de falantes brasileiros do sexo masculino conforme estudo de Barbosa e Madureira (2015), em cinza, para comparação. Na imagem, as realizações individuais das vogais são sinalizadas pelo código do falante seguido da vogal (/i/, /a/ e /u/), com os centróides em destaque. A inspeção visual da imagem sugere uma tendência à centralização do grupo SCZ com relação ao controle, mais evidente no eixo de F2, bem como uma centralização dos grupos SCZ e controle com relação à descrição de Barbosa e Madureira (2015), sobretudo em F1 para /i/ e em F2 para /a/ e /u/.

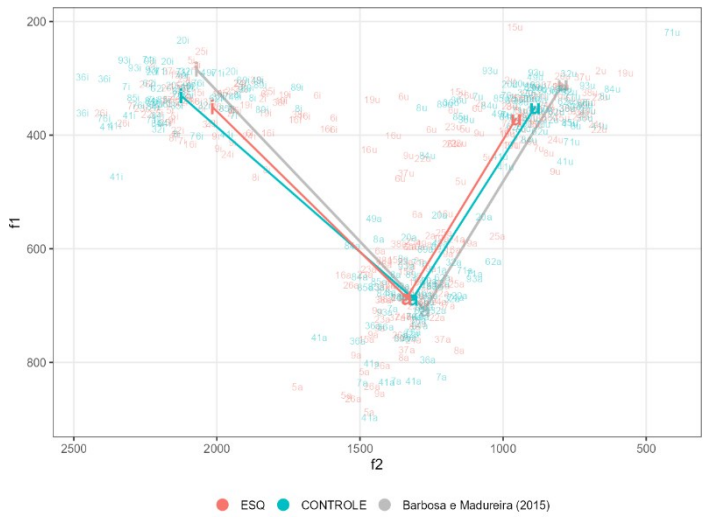


Figura 1. Espaço acústico vocálico por grupo (SCZ e controle) e referência (Barbosa; Madureira 2015). Médias de F1 e F2 (Hz) para as vogais /a/, /i/ e /u/ nos grupos SCZ (vermelho) e controle (azul), com referência de Barbosa e Madureira (2015) em cinza.

A Figura 2 a seguir apresenta boxplots com valores de FCR (em cima à esquerda), PHI (em cima, à direita), inter_ms (embaixo, à esquerda) e intra_ms (embaixo, à direita) por grupo (SCZ em vermelho e controle em azul). Os boxplots de FCR e inter_ms (ambos à esquerda) possuem maior diferença entre as áreas ocupadas por cada grupo. Os boxplots de PHI e intra_ms exibem maior sobreposição entre os grupos. Os participantes de código 6 e 16 do grupo ESQ e 89 do grupo controle foram identificados como *outliers* para algumas medidas. No grupo experimental, essas discrepâncias acústicas coincidem com perfis clínicos extremos. Já no grupo controle, o participante 89 destacou-se por uma acentuada centralização restrita à vogal /i/. Não havendo motivos clínicos para descartar os *outliers*, foram mantidos em todas as comparações entre grupos apresentadas a seguir.

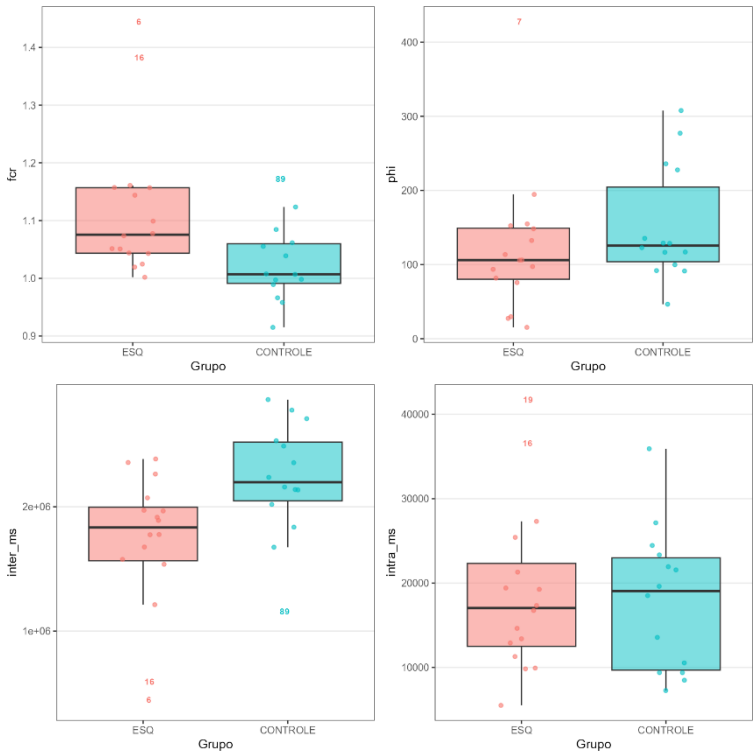


Figura 2. Distribuição do FCR, PHI, inter_ms e intra_ms por grupo. Pontos representam valores individuais. Números rotulados identificam participantes com valores extremos.

O grupo SCZ apresentou distribuições significativamente diferentes do grupo controle em duas medidas acústicas, o FCR e o inter_ms. O índice de centralização formântica (FCR) foi maior no grupo SCZ (Mdn = 1,078) do que no grupo controle (Mdn = 1,007; U = 52,0; p = 0,024; r = −0,50 [IC 95%: −0,83; −0,09]), indicando maior centralização vocálica. O índice de separação intercategorial (inter_ms) foi menor no grupo SCZ (Mdn = 1.833.147) do que no grupo controle (Mdn = 2.197.530; U = 157,0; p = 0,021; r = 0,51 [IC 95%: 0,11; 0,84]), indicando menor distinção entre categorias vocálicas. Ambos os efeitos se mantiveram após controle por idade (FCR: Z = 2,14; p = 0,049; inter_ms: Z = 2,27; p = 0,035), mas não após a correção da Taxa de Falsas Descobertas pelo método de Benjamini-Hochberg (FDR-BH; q = 0,097 para ambos).

As demais variáveis do falante testadas, a saber o índice de variabilidade interna das vogais (*intra_ms*, que mede a variabilidade nas repetições do fonema), e o índice de organização do sistema vocálico (PHI, fruto da razão entre inter_ms e intra_ms) não diferiram significativamente entre os grupos em nenhuma das análises.

Apesar das diferenças visuais entre grupos observadas na Figura 1, as comparações por vogal (F1 e F2 de /a/, /i/ e /u/) não revelaram diferenças significativas após correção para múltiplas comparações. A única exceção é F2 de /i/, parâmetro que mais difere SCZ do grupo controle no gráfico da Figura 1, que é significativamente maior para o grupo controle (t(28) = 2,63; p = 0,014; d = -0,96 [IC 95%: -1,72; -0,20]), mas não sobreviveu à correção para múltiplas comparações.

3.2. Associações com variáveis clínicas

Foram testadas associações bivariadas entre as variáveis linguísticas FCR, PHI, inter_ms, intra_ms e valores de F1 e F2 de cada vogal e as variáveis clínicas equivalente de

clorpromazina e EPS total. A única associação que sobreviveu à correção para múltiplas comparações foi a correlação entre intra_ms e EPS total (ρ = -0,658; p = 0,006; q = 0,011; IC 95% [-0,910, -0,211]). Esse achado motivou a realização de análise *post-hoc* testando correlações de intra_ms e os subitens da escala EPS, em que foram encontradas três correlações negativas que sobrevivem à correção FDR: EPS3, chacoalhar o ombro (p = 0,006; q = 0,019; ρ = -0,658 [IC 95%: -0,850; -0,284]); EPS4, rigidez de cotovelo (p = 0,005; q = 0,019; ρ = -0,662 [IC 95%: -0,817; -0,294]); EPS6, pêndulo das pernas (p = 0,006; q = 0,019; ρ = -0,658 [IC 95%: -0,850; -0,284]).

4. Discussão

Esse trabalho investigou a existência de redução no espaço acústico em pessoas com esquizofrenia com relação a um grupo controle e buscou correlacionar medidas das vogais com características clínicas dos pacientes. O grupo alvo foi formado por 16 indivíduos do C-ORAL-ESQ, corpus de consultas psiquiátricas de pacientes com esquizofrenia. O grupo controle foi formado por 14 indivíduos do C-ORAL-MED, corpus de consultas de clínica geral e cardiologia de pessoas sem transtornos mentais. Na comparação entre grupos, ESQ diferiu significativamente do controle com relação a FCR, inter_ms – ambos sobrevivendo à correção para múltiplas comparações – e F2 de /i/, sem sobreviver à correção de múltiplas comparações. PHI e intra_ms não foram significativamente diferentes entre os grupos. Esses resultados dão suporte à hipótese de que o grupo ESQ apresenta centralização do trapézio vocálico com relação ao controle.

Com relação às associações bivariadas entre os parâmetros linguísticos e variáveis clínicas de ESQ, a única correlação significativa foi entre intra_ms e o EPS total, que avalia sintomas extrapiramidais. O equivalente de clorpromazina não correlacionou com nenhuma variável linguística. A correlação com EPS total motivou uma análise *post-hoc* com os subitens da escala, em que revelou correlações negativas entre intra_ms e EPS3 (chacoalhar o combro), EPS4 (rigidez de cotovelo) e EPS6 (pêndulo das pernas). O intra_ms expressa a variabilidade interna a cada categoria vocálica: valores superiores indicam maior instabilidade articulatória. Chama atenção o fato de que indivíduos com valores superiores de EPS3, EPS4 e EPS6 apresentem menor instabilidade articulatória, sobretudo considerando que Roland *et al.* (2023) demonstraram intra_ms elevado em pacientes parkinsonianos em relação ao grupo controle. Esse achado deve ser observado com cautela e confirmado em estudos futuros com amostras maiores. Destaca-se que o C-ORAL-ESQ não apresenta histórico completo das prescrições medicamentosas, contendo apenas os medicamentos prescritos na data exata da consulta. Dessa forma, os dados de EPS podem atuar como indicador indireto de efeitos medicamentosos a longo prazo.

Questões éticas

O C-ORAL-ESQ e C-ORAL-MED foram aprovados pelos Comitê de ética da UFMG (COEP: CAAE 0209.0.203.000-07).

Referências

Barbosa, P. A. and Madureira, S. (2015) *Manual de fonética acústica experimental: aplicações a dados do português*, Cortez Editora, São Paulo.

Boersma, P. (2001) "Praat, a system for doing phonetics by computer", *Glot*

International, v. 5, n. 9–10, p. 341–345.

Gardner, D. M. et al. (2010) "International consensus study of antipsychotic dosing", *American Journal of Psychiatry*, v. 167, n. 6, p. 686–693.

Hogoboom, L., van Dijk, M. W. M., Castelein, S., Bruggeman, R. and Wieling, M. (2023) "Initial Evidence of Vowel Space Reduction in a Subset of Individuals with Schizophrenia", *Journal of Speech, Language, and Hearing Research*, v. 66, n. 12, p. 4949–4963\. DOI: 10.1044/2023_JSLHR-23-00062.

Kay, S. R., Fiszbein, A. and Opler, L. A. (1987) "The Positive and Negative Syndrome Scale (PANSS) for schizophrenia", *Schizophrenia Bulletin*, v. 13, n. 2, p. 261–276.

Rocha, B. et al. (2026) "C-ORAL-ESQ: Corpus Orale di Schizofrenici", http://www.c-oral-brasil.org.

Rocha, B., Toledo, G. and Raso, T. (Em preparação) "C-ORAL-MED: Corpus de fala espontânea de interação entre médico e paciente".

Roland, N. J., Antoniou, M. and Meyer, B. T. (2023) "Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson's disease", *Journal of Parkinson's Disease*, v. 13, n. 3, p. 399–413\. DOI: 10.3233/JPD-225050.

Sapir, S., Ramig, L. O., Spielman, J. L. and Fox, C. (2010) "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech", *Journal of Speech, Language, and Hearing Research*, v. 53, n. 1, p. 114–125\. DOI: 10.1044/1092-4388(2009/08-0184).

Simpson, G. M. and Angus, J. W. S. (1970) "A rating scale for extrapyramidal side effects", *Acta Psychiatrica Scandinavica Supplementum*, v. 45, supl. 212, p. 11–19.

Sinha, V. K., Praharaj, S. K., Chakraborty, V. and Singh, A. (2014) "Predictors of effect of atypical antipsychotics on speech", *Psychiatry Research*, v. 220, n. 3, p. 1003–1008\. DOI: 10.1016/j.psychres.2014.10.008.

Detecção e diagnóstico de erros segmentais de pronúncia em corpus oral de aprendizes de Inglês com modelos auto-supervisionados sem ajuste de pesos

Marcos Antonio R. Pinto Junior[1,2], Michaël Grégoire[1], Jackson Wilke da C.Souza[2]

[1]Departamento de Estudos Hispânicos e Ibero-Americanos Université Clermont Auvergne (UCA) – Clermont-Ferrand – França, [2]Programa de Pós-Graduação em Língua e Cultura (PPGLinC) Universidade Federal da Bahia (UFBA) – Salvador, BA – Brasil

marcos_antonio.ramos_pinto_junior@doctorant.uca.fr

Abstract. *This paper investigates segmental mispronunciation detection and diagnosis (MDD) in the manually annotated 16 kHz L2-ARCTIC. We evaluate a retrieval-based method that keeps SSL speech models frozen and labels test frames by cosine nearest-neighbour search over a labelled embedding pool. Five retrieval paths are compared on F1 and PER, and the gold errors are analysed by native language and type. The validation-selected ensemble reaches 74.25% test F1 against 73.36% for the HuBERT baseline, a small, untested gain with many more insertions — automatic feedback remains limited when insertions are frequent.*

Resumo. *Este trabalho investiga a detecção e o diagnóstico de erros segmentais de pronúncia (MDD) na versão manualmente anotada a 16 kHz do L2-ARCTIC. Avaliamos um método que mantém modelos SSL congelados e rotula quadros de teste por vizinhos mais próximos (cosseno) em um conjunto rotulado de embeddings. Cinco caminhos são comparados por F1 e PER, e os desvios do teste são analisados por língua materna e tipo. O ensemble selecionado na validação alcança 74,25% de F1 contra 73,36% da base HuBERT, ganho pequeno, sem significância testada e com muito mais inserções: o feedback automático segue limitado quando as inserções são frequentes.*

1. Introdução

Corpora orais anotados de aprendizes permitem observar, em escala, como falantes de diferentes línguas maternas realizam sons de uma língua-alvo [Hérry-Bénit *et al.*, 2021; Han, Kin e Chung, 2024], e a Linguística de *corpus* fornece o enquadramento para tratá-los como base empírica de descrição e comparação em dados linguísticos [McEnery e Hardie 2012]. Quando o conjunto de dados registra tanto a sequência canônica esperada quanto a percebida por anotadores, é possível investigar quantitativamente padrões de erro segmental, tais como inserções, omissões e substituições. Esse tipo de operação é definido como *Mispronunciation Detection and Diagnosis* (MDD), que é o nível da detecção e diagnóstico de erros de pronúncia, que visa identificar onde a produção diverge do *corpus* padrão-ouro e de que tipo é a divergência.

A construção de sistemas de MDD baseados em unidades fonêmicas depende de dados de treinamento anotados nesse nível, além de modelos especializados, o que pode dificultar sua aplicação em cenários com poucos dados ou com novos perfis de aprendizes [Ye *et al.*, 2022; Shahin e Ahmed, 2024]. Uma alternativa recente mantém modelos auto-

supervisionados (SSL) estáticos e compara representações acústicas a um conjunto rotulado por vizinhos mais próximos [Tu *et al.*, 2026]. Neste trabalho, investigamos essa abordagem no L2-ARCTIC, anotado manualmente [Zhao *et al.*, 2018; Chikingsley s.d.] a partir das seguintes perguntas de pesquisa: (i) Representações estáticas detectam erros segmentais? (ii) Quais escolhas de recuperação melhoram o F1 sobre a base HuBERT, e o ganho é robusto? (iii) Quais limites esse procedimento impõe ao *feedback* automático quando o ganho vem com mais inserções? Assim, espera-se contribuir com a comparação de cinco caminhos de recuperação e a caracterização dos desvios anotados por língua materna e tipo.

2. Trabalhos relacionados

Avaliar pronúncia não se confunde com buscar acento nativo: a inteligibilidade prioriza o que afeta a compreensão, em vez de penalizar toda marca de não natividade [Levis 2005]. Nesse caso, a saída gerada automaticamente é tratada como apoio à análise, não como julgamento do falante. Tecnicamente, a MDD é formulada por alinhamentos entre as sequências canônica, percebida e reconhecida, medindo aceitação, rejeição e diagnóstico correto [Leung *et al.*, 2019]; modelos SSL como HuBERT, wav2vec2 e WavLM reduzem a dependência de treinamento específico [Baevski et al. 2020; Hsu *et al.*, 2021; Chen *et al*, 2022], o que importa quando a anotação fina é custosa.

Tu *et al.* (2026) propõem MDD por recuperação com o modelo estático. Nesse modelo, o áudio é codificado quadro a quadro, em que cada um recebe o rótulo dos vizinhos mais próximos em um conjunto rotulado, e a sequência prevista é alinhada à canônica. No L2-ARCTIC reportam F1 de 69,60% com HuBERT-Large ls960-ft, *pool* de 500 enunciados, k=10, limiar 0,7, quadro central e maioria simples, com PER de 104,08% pela quantidade de inserções; em *preprint*, o CROTTC-IF reporta 71,77% com alinhamento monótono por quadro [Geng *et al.*, 2026].

3. Metodologia

Com relação ao *corpus*, utilizamos o L2-ARCTIC [Zhao *et al.*, 2018] na versão anotada manualmente a 16 kHz [Chikingsley s.d.], com fala de aprendizes de seis línguas maternas e, por enunciado, a sequência canônica, a percebida e a anotação por segmento (fonema canônico, percebido e tipo de erro), já alinhada temporalmente. A divisão pública é *por locutor:* teste com 900 enunciados de seis locutores (NJS, TLV, TNI, TXHC, YKWK, ZHAA), validação com 899 de outros seis (MBMPS, NCC, SVBI, THV, YBAA, YDCK) e treino com os doze restantes, do qual se amostram 500 enunciados para o *pool*. Nenhuma voz do *pool* reaparece na avaliação, em que a semelhança acústica entre falantes infla o desempenho. A amostragem foi única e sua variação não foi medida.

Em "*I had faith in them*" (TXHC, L1 mandarim, *arctic_a0030*), a canônica AY HH AE D F EY TH IH N DH EH M é percebida como AY HH AE D **AH** F EY **S IY NG** DH EH M: inserção de /AH/ e substituições /TH/→/S/, /IH/→/IY/, /N/→/NG/. A inserção é o caso delicado: não corresponde a fonema canônico algum e escapa à comparação posição a posição.

A base usa HuBERT-Large [Hsu *et al.*, 2021] ls960-ft, congelado; no ensemble somam-se wav2vec2-Large [Baevski *et al.*, 2020] e WavLM-Large [Chen *et al.*, 2022],

também sem ajuste de pesos. Cada modelo produz um *embedding* por quadro (20 ms) da camada final, normalizado (norma L2), e no *pool* cada fonema é representado pelo quadro central. Na inferência, cada quadro é rotulado pelos k vizinhos mais próximos por cosseno, retidos acima de um limiar τ. Assim, onde o trabalho de referência usa maioria simples, adotamos votação por temperatura: cada vizinho pesa exp(similaridade / T), de modo que, com T baixo, o mais semelhante domina sem descartar os demais. Sem vizinho acima de τ, atribui-se *blank*; a sequência é decodificada colapsando repetições e removendo *blanks*.

Todos os caminhos partem de k = 5, τ = 0,6, T = 0,05, pool = 500, camada final e quadro central, diferindo apenas na regra de decisão. Os hiperparâmetros que passaram por *fine-tuning* proposto seriam: (1) *Base HuBERT*: votação por temperatura com τ e T globais, não sendo réplica exata do sistema de referência (k = 10, τ = 0,7, maioria simples), e sim nossa linha de base local; (2) *Limiar por fonema*: τ passa a depender do fonema vencedor, escolhido na validação e mantido global para os pouco frequentes; (3) *Temperatura por classe*: T depende do fonema canônico alinhado ao quadro, também selecionada na validação; (4) *Ensemble SSL*: os três modelos recuperam em índices independentes do mesmo *pool*; por quadro, os pesos de cada um viram distribuição sobre os rótulos, combinadas por média com pesos iguais, e o rótulo é o argmax. (5) *Suavização temporal*: voto majoritário em janela deslizante e descarte de segmentos muito curtos, reatribuídos ao vizinho mais longo.

A avaliação segue o protocolo hierárquico de MDD [Leung *et al.*, 2019], que combina os alinhamentos canônico×percebido, percebido×reconhecido e canônico×reconhecido para classificar cada fonema em aceitação ou rejeição, verdadeira ou falsa. Como medidas de avaliação, o sistema reporta Precisão, Revocação e F1 de detecção (classe positiva = erro) e a taxa de erro de fonema (PER) entre a sequência reconhecida e a percebida; a taxa de diagnóstico por tipo de erro não é reportada (Seção 4.1). Como a validação orienta a seleção da configuração, os valores são exploratórios: sem intervalos de confiança ou testes de significância e com execução única por caminho, diferenças de F1 abaixo de 1 ponto percentual são apenas indicativas.

4. Resultados e discussão

Avaliamos na validação (899 enunciados) e no teste (900), em GPU NVIDIA H100 NVL.

Tabela 1. Varredura exploratória no L2-ARCTIC, em %

Caminho	Val. F1	Teste F1	Teste PER
Base HuBERT	75,93	73,36	128,19
Limiar por fonema	75,93	73,37	128,19
Temperatura por classe	70,87	68,95	134,65
Ensemble SSL	77,56	74,25	178,42
Suavização temporal	72,31	69,28	90,48

O maior F1 por validação foi do ensemble SSL, que também supera a base no teste (74,25% contra 73,36%). O ganho vem da revocação (71,24%), mas é de 0,89 ponto percentual, em execução única e sem teste de significância, enquanto o PER sobe de

128,19% para 178,42%: o ensemble recupera mais desvios ao custo de muito mais inserções espúrias, deslocando o ponto de operação em vez de ser um sistema melhor. O limiar por fonema quase não altera o resultado, a temperatura por classe degrada o F1 e a suavização reduz o PER a 90,48% derrubando o F1 a 69,28%: somando F1 e PER, a base HuBERT segue a mais equilibrada.

Como referência externa, PER-MDD reporta 69,60% e CROTTC-IF, 71,77% [Tu *et al.*,2026; Geng *et al.*, 2026]. Esses resultados não são diretamente comparáveis, pois *pool*, amostragem, alinhamento e métricas diferem. Ainda assim, a ablação de Tu et al. associa o wav2vec2 a um PER de 173,28%, o mesmo padrão observado em HuBERT ao *ensemble*.

Com relação à distribuição de erros no corpus e os limites do diagnóstico, no teste, os 900 enunciados somam 29.787 fonemas canônicos, 4.322 deles (14,5%) com anotação de desvio, conforme demonstrado na Tabela 2.

Tabela 2. Desvios anotados no teste, por L1 e tipo, em % dos fonemas do locutor

L1 (locutor)	Subst.	Omissão	Inserção	Total
Árabe (ZHAA)	7,7	2,0	1,0	10,7
Mandarim (TXHC)	9,7	2,6	1,5	13,8
Hindi (TNI)	9,9	1,4	0,2	11,5
Coreano (YKWK)	9,4	2,6	0,6	12,6
Espanhol (NJS)	11,1	1,6	0,7	13,4
Vietnamita (TLV)	15,4	8,7	0,8	24,9
Total	10,6	3,2	0,8	14,5

A partir desses resultados, é possível interpretar destacar três aspectos da Tabela 1. Primeiro, há um *desbalanceamento* na tarefa, pois 85,5% dos fonemas estão corretos, de modo que rejeitar pouco reduz o F1 e rejeitar muito, como o *ensemble*, ganha revocação às custas de precisão. Substituições respondem por 72,7% dos desvios e omissões por 21,8%, contra 5,5% de inserções: as inserções que dominam o PER são majoritariamente espúrias, de quadros de transição, não desvios do aprendiz. E a taxa de erro varia de 10,7% a 24,9% conforme a L1, com o vietnamita concentrando omissões (8,7%, o triplo da média) — assimetria que a leitura agregada esconde. Os pares mais frequentes são /DH/→/D/ (14,3% das substituições), /Z/→/S/ (13,1%) e /IH/→/IY/ (5,2%), com perfis por L1: /DH/→/D/ lidera entre vietnamita, hindi e coreano; /Z/→/S/ entre espanhol, árabe e mandarim; /V/→/F/ só no árabe.

Cabe explicitar o que não medimos: a taxa de diagnóstico correto por tipo de erro, que exigiria contabilizar, entre as rejeições verdadeiras, quantas identificam o fonema produzido. Sem ela não se pode afirmar que o sistema *diagnostica* os três tipos com qualidade comparável, e a assimetria da Tabela 2 sugere o contrário.

5. Considerações finais

Sem ajuste de pesos para MDD, a base HuBERT alcança 73,36% de F1 e o ensemble selecionado pela validação, 74,25%; o ganho é pequeno, não testado estatisticamente e acompanhado de forte aumento do PER, enquanto a suavização reduz inserções e o F1. A

limitação central é o equilíbrio entre revocação e inserções espúrias, somada à única amostra do pool e à execução única por caminho. Como etapas futuras: reportar o diagnóstico por tipo de erro e por L1; repetir a varredura com múltiplas amostras; suprimir inserções sem perder revocação; e estender a *corpora* de aprendizes de português brasileiro. Para usos pedagógicos, os resultados são apoio à análise, não avaliação definitiva do falante.

Questões éticas e uso de IA

O estudo usa *corpus* público já anotado, sem novas gravações nem identificação de participantes; os resultados servem à análise agregada de padrões, não ao julgamento de competência individual, pois modelos acústicos podem refletir desequilíbrios de cobertura de sotaques e de gravação. Ferramentas de IA apoiaram a organização textual, a revisão bibliográfica e a conversão de formatos; a interpretação e a responsabilidade científica permanecem com os autores.

Referências

Baevski, A. *et al.* (2020) wav2vec 2.0: Self-Supervised Learning of Speech Representations, In: *NeurIPS*, 33, 12449-12460.

Chen, S. *et al.* (2022) WavLM: Large-scale self-supervised pre-training for full-stack speech processing, *IEEE JSTSP*, 16(6), 1505-1518.

Chikingsley (s.d.) L2-ARCTIC manual v5.0 16k, Hugging Face Datasets, huggingface.co/datasets/chikingsley/l2-arctic-manual-v5.0-16k, acesso em 28 jun. 2026.

Geng, H. *et al.* (2026) Beyond Acoustic Sparsity and Linguistic Bias: A Prompt-Free Paradigm for Mispronunciation Detection and Diagnosis, arXiv:2604.22133.

Han, S., Kim, S. and Chung, M. (2024). Constructing Korean learners' L2 speech corpus of seven languages for automatic pronunciation assessment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3772–3781, Torino, Italy. ELRA and ICCL.

Herry-Bénit, N., Lopez, S., Kamiyama, T. and Tennant, J. (2021). The interphonology of contemporary English corpus (IPCE-IPAC). *International Journal of Learner Corpus Research*, 7(2):275–289. DOI: 10.1075/ijlcr.20010.her.

Hsu, W.-N. *et al.* (2021) HuBERT: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM TASLP*, 29, 3451-3460.

Leung, W.-K., Liu, X. and Meng, H. (2019) CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis, In: *ICASSP 2019*, 8132-8136.

Levis, J. M. (2005) Changing contexts and shifting paradigms in pronunciation teaching, *TESOL Quarterly*, 39(3), 369-377.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics*: Method, Theory and Practice, Cambridge Univ. Press.

Shahin, M. and Ahmed, B. (2024). Phonological-level mispronunciation detection and diagnosis. In *Proceedings of Interspeech 2024*, pages 4099–4103, Kos, Greece. ISCA.

Tu, H. T. *et al.*(2026) Mispronunciation Detection and Diagnosis Without Model Training: A Retrieval-Based Approach, In: *ICASSP 2026*.

Ye, W., Mao, S., Soong, F., Wu, W., Xia, Y., Tien, J. and Wu, Z. (2022). An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (APL) embeddings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Singapore. IEEE. DOI: 10.1109/ICASSP43922.2022.9747654.

Zhao, G. *et al.* (2018) L2-ARCTIC: A non-native English speech corpus, *Interspeech* 2018, 2783-2787.

Prosódia e função discursiva em palavrões no português brasileiro: uma análise preliminar de "pô" na fala espontânea baseada em corpora

Victor Hugo Medina, Átila Soares Vital

Universidade Federal de Minas Gerais, Radboud Universiteit

victormedina@ufmg.br, atila.vital@ru.nl

Abstract. *This paper presents a preliminary corpus-based study of the pragmatic and prosodic uses of the lexical items* pô *and* porra *in Brazilian Portuguese, contrasting spontaneous speech from the C-ORAL-BRASIL I corpus with informal digital writing from a sample of the FDP Corpus, consisting of* Reddit *comments. The study is based on the hypothesis that the modality of circulation and prosodic realization may influence the distribution and pragmatic interpretation of these forms. The analysis combines relative frequencies, concordance lines and acoustic analysis in* Praat*, with the aid of the Parselmouth library. In the spoken corpus, 30 occurrences of* pô *and 13 of* porra *were identified, while the written sample contained 7 occurrences of* pô *and 224 of* porra*. Ten spoken occurrences of* pô *were selected for acoustic analysis. Nine of them displayed falling or level contours and relatively short duration, patterns compatible with less expressive uses related to discourse organization and turn management. In the written sample,* porra *occurred more frequently than* pô *and appeared in contexts involving intensification and attitude expression. The results suggest a possible functional specialization between the reduced and full forms, as well as differences in the expressive resources mobilized by spoken and digital written language.*

Resumo. *Este trabalho apresenta um estudo preliminar baseado em corpora sobre os usos pragmáticos e prosódicos dos itens lexicais* pô *e* porra *no português brasileiro, contrastando a fala espontânea do corpus C-ORAL-BRASIL I e a escrita digital informal, representada por uma amostra do Corpus FDP, constituída por comentários do Reddit. O estudo parte da hipótese de que a modalidade de circulação e a realização prosódica podem influenciar a distribuição e a interpretação pragmática dessas formas. A análise combina frequências relativas, linhas de concordância e análise acústica realizada no* software Praat, *com auxílio da biblioteca Parselmouth. No corpus oral, foram identificadas 30 ocorrências de* pô *e 13 de* porra*, enquanto a amostra escrita contém 7 ocorrências de* pô *e 224 de* porra*. Dez ocorrências orais de* pô *foram selecionadas para análise acústica. Em nove delas, observaram-se contornos de* f_0 *descendentes ou nivelados e duração relativamente curta, padrões compatíveis com usos menos expressivos relacionados à organização discursiva e ao gerenciamento do turno. Na amostra escrita,* porra *apresentou frequência superior à de* pô *e ocorreu em contextos de intensificação e expressão de atitude. Os resultados sugerem uma possível especialização funcional entre as formas reduzida e plena, bem como diferenças nos recursos de expressividade mobilizados pela fala e pela escrita digital.*

1. Introdução

Os palavrões são parte de um fenômeno linguístico de interesse para o estudo acadêmico porque circulam entre diferentes funções discursivas e pragmáticas. Ainda que sejam frequentemente associados apenas à expressão de emoção ou agressividade, seu uso efetivo na língua mostra uma diversidade muito maior de funções, incluindo avaliação, intensificação, desabafo, marcação de atitude e organização do discurso. Por conta disso, o estudo de palavrões oferece uma janela relevante para a análise da linguagem em uso, especialmente quando se considera a diferença entre registros falados e escritos.

Neste trabalho, os palavrões são entendidos como formas associadas a domínios socialmente estigmatizados, cuja interpretação pode variar conforme o contexto e a função discursiva [Jay; Janschewitz, 2008]. Dessa forma, uma mesma forma pode ocorrer em insultos, avaliações, intensificações, expressões emotivas e usos como marcadores discursivos. Essa variação funcional justifica a consideração de fatores lexicais, pragmáticos e prosódicos.

No caso da fala espontânea, a interpretação de um palavrão depende tanto de sua forma lexical quanto de fatores prosódicos, como entonação, intensidade e duração. Em estudos de fala espontânea sobre o português brasileiro (PB), a prosódia participa da segmentação do fluxo da fala e do reconhecimento de funções pragmáticas e ilocucionárias [Raso; Rocha, 2025]. Nesse sentido, a análise de *pô* deve, portanto, considerar a sua forma lexical e sua realização entoacional e temporal. Esses elementos podem reforçar seu valor expressivo ou, em contrapartida, enfraquecê-lo e aproximá-lo de uma função de cunho mais discursivo. Na escrita em meios digitais, por sua vez, os palavrões não contam diretamente com um componente entoacional, o que torna ainda mais importante observar os contextos sintáticos e pragmáticos em que ocorrem. A comparação entre esses dois registros pode, com isso, revelar diferenças significativas na distribuição funcional dessas expressões. Além disso, ainda que a escrita digital disponha de outros recursos expressivos, como pontuação, repetição, capitalização e grafias não convencionais, a comparação neste estudo se concentra na distribuição lexical e nos contextos sintáticos e pragmáticos das formas analisadas.

Este trabalho investiga esses problemas a partir da forma *pô*, observando sua ocorrência em fala espontânea no corpus C-ORAL-BRASIL I [Raso; Mello, 2012], composto por dados informais do PB, e sua relação com dados escritos do *Reddit* no Corpus FDP, em construção pelos autores. A hipótese inicial é que *pô* apresenta na fala um comportamento associado a marcadores discursivos e a unidades entoacionais específicas, enquanto o corpus escrito pode mostrar usos distintos, mais dependentes do contexto textual. Com isso, espera-se contribuir para uma descrição mais precisa do funcionamento dos palavrões no PB, articulando contribuições da linguística de corpus, da pragmática e da prosódia. A comparação proposta busca ainda verificar se as restrições que são próprias de cada modalidade influenciam a seleção e a função das formas concorrentes *pô* e *porra*.

2. Análise de corpus e distribuição lexical na fala e na escrita digital

Com o objetivo de contrastar o comportamento dos palavrões na fala espontânea e na escrita digital, foram comparados dados do corpus informal C-ORAL-BRASIL I [Raso & Mello, 2012] e uma amostra do Corpus FDP, extraída do *Reddit* e processada por meio

da plataforma *Sketch Engine*. A amostra escrita contém 78.378 tokens, correspondentes a 63.569 palavras, distribuídos em 1.894 sentenças. Devido às diferenças de tamanho, modalidade, composição e situação comunicativa entre os corpora, todas as frequências absolutas foram normalizadas para frequências relativas por milhão de tokens (i.p.m.). A comparação aqui deve ser compreendida como exploratória, uma vez que os *corpora* diferem em modalidade, composição temática e situação comunicativa.

Tabela 1. Distribuição de frequências absolutas e relativas (por milhão de tokens) entre as formas *pô* e *porra*

Forma lexical	Ocorrências na fala	Frequência relativa na fala	Ocorrências escritas	Frequência relativa na escrita
pô	30	Aprox. 150	7	89,31
porra	13	Aprox. 65	224	2.857,94

Nota: As frequências relativas foram calculadas por milhão de tokens. A amostra escrita do Corpus FDP contém 78.378 tokens e 63.569 palavras. Os dados da fala foram adaptados de Raso e Mello (2012) e os dados da escrita foram gerados por meio da plataforma *Sketch Engine*.

Os resultados sugerem que, no português falado, a variante reduzida *pô* é sensivelmente mais produtiva que sua contraparte canônica (150,00 i.p.m. vs. 65,00 i.p.m.). Conforme discutido na literatura [Santos, 2024], a oralidade tende a favorecer processos de desgaste fonético e *bleaching* semântico em itens de alta frequência, promovendo a gramaticalização de palavrões em marcadores discursivos esvaziados de teor ofensivo. Neste estudo, essa possibilidade é tratada como uma hipótese interpretativa em vez de um processo de gramaticalização consolidado.

Em contrapartida, os dados de escrita digital no *Reddit* parecem ir de encontro a essa lógica. A forma canônica *porra* apresenta, na amostra analisada, uma frequência relativa muito superior à de *pô* (2.857,94 i.p.m. contra 89,31 i.p.m., respectivamente). A análise das linhas de concordância (KWIC) no *Sketch Engine* contribui para essa interpretação: na ausência de parâmetros acústicos de relevância informacional, os dados sugerem que a variante tabu canônica pode atuar em construções de intensificação sintática e pragmática (ex.: "essa *porra* é um bang bang do caralho", "que *porra* é essa na cabeça dela?").

Adicionalmente, as linhas de concordância sugerem co-ocorrências de *porra* com outras formas tabu, como *puta que pariu*, *do caralho* e *tomar no cu*. Como essas ocorrências ainda serão quantificadas, elas são apresentadas como uma observação de caráter exploratório. Nesta etapa, não há evidência quantitativa suficiente para estabelecer uma diferenciação entre as modalidades. Ainda assim, os dados permitem hipotetizar que enquanto a fala espontânea pode gerenciar a atenuação pragmática por meio de recursos prosódicos, como contornos menos proeminentes, e fonéticos, como a apócope, a escrita na internet pode recorrer a formas lexicais plenas para produzir expressividade e saliência pragmática.

3. Análise Prosódica de *pô* na Fala Espontânea

A principal vantagem de se mobilizar um corpus de fala em contraposição a dados

puramente escritos reside no acesso a categorias entoacionais determinantes para a interpretação de fenômenos pragmáticos. Para além da estrutura morfossintática, os palavrões e suas formas reduzidas podem ser avaliados em relação ao seu papel informacional dentro dos enunciados por meio de parâmetros acústicos como a intensidade, a duração e a frequência fundamental (f_0). Para esta análise, as ocorrências de *pô* realizadas em unidades entoacionais dedicadas no C-ORAL-BRASIL I foram filtradas, resultando em 10 ocorrências passíveis de modelação acústica isolada após o descarte de sobreposições de fala ou ruídos de captação. Foram consideradas passíveis de análise as ocorrências que permitiam delimitação acústica do item e apresentavam sinal suficiente para a observação da duração e da trajetória de f_0. A seleção portanto não corresponde à totalidade das ocorrências de *pô* no corpus, já que se tem aqui um subconjunto para a análise acústica. Essa seleção segue a ideia de que a segmentação prosódica constitui uma etapa importante para a interpretação da pragmática da fala, já que as fronteiras prosódicas ajudam a delimitar os domínios nos quais as relações sintáticas, semânticas e ilocucionárias são reconstruídas [Raso; Rocha, 2025].

As análises acústicas obtidas com a biblioteca Parselmouth [Jadoul; Thompson; De Boer, 2018], que adapta o *Praat* [Boersma; Weenink, 2026] em Python, revelaram uma tendência preliminar de diferenciação prosódica e funcional. Em 9 dos 10 casos investigados, tanto a curva melódica quanto a duração das unidades exibiram comportamentos comparáveis: uma entonação descendente ou nivelada (*flat*) e uma duração temporal entre 120 e 200 ms. A descrição aqui deve ser entendida em termos relativos, já que diferentes contornos podem desempenhar funções distintas dependendo de sua posição no enunciado e de seu alinhamento com unidades prosódicas. Estudos sobre a entoação do PB também mostram que movimentos ascendentes, descendentes e ascendente-descendentes podem se associar a diferentes graus de proeminência e organização discursiva [Lucente, 2025]. Esse padrão é compatível com usos que são associados a marcadores discursivos e partículas de preenchimento ou gerenciamento de turno na fala espontânea [Santos, 2024]. Considerando o número de ocorrências analisadas, essa associação se entende como indicação preliminar. Sob essa configuração prosódica, a interpretação semântica tradicional associada ao palavrão de base, como indexador direto de hostilidade ou quebra severa de tabu, é atenuada, operando estritamente na gestão conversacional.

Por outro lado, um único caso manifestou maior duração temporal e um padrão melódico ascendente-descendente, descrito preliminarmente como uma possível realização de maior ilocução autônoma [Austin, 1962]. Essa interpretação inicial se justifica porque a função depende também dos parâmetros pragmáticos e contextuais do enunciado [Raso; Rocha, 2025]. Tal padrão merece atenção porque combina mudança de direção da f_0 e maior duração, características que podem aumentar a proeminência perceptual da unidade. Esse padrão é compatível com uma realização mais expressiva, embora não seja suficiente, isoladamente, para determinar sua função discursiva [Lucente, 2025]. Embora o item lexical seja idêntico (*pô*), a configuração acústica torna menos provável o seu enquadramento como mero marcador de preenchimento. Trata-se de uma realização fonética saliente, que pode veicular de forma mais direta uma atitude afetiva ou uma possível indignação do locutor, podendo aproximar-se da carga expressiva original da variante canônica. Na figura 1, essa ocorrência é representada pelo último espectrograma, com subtítulo em azul.

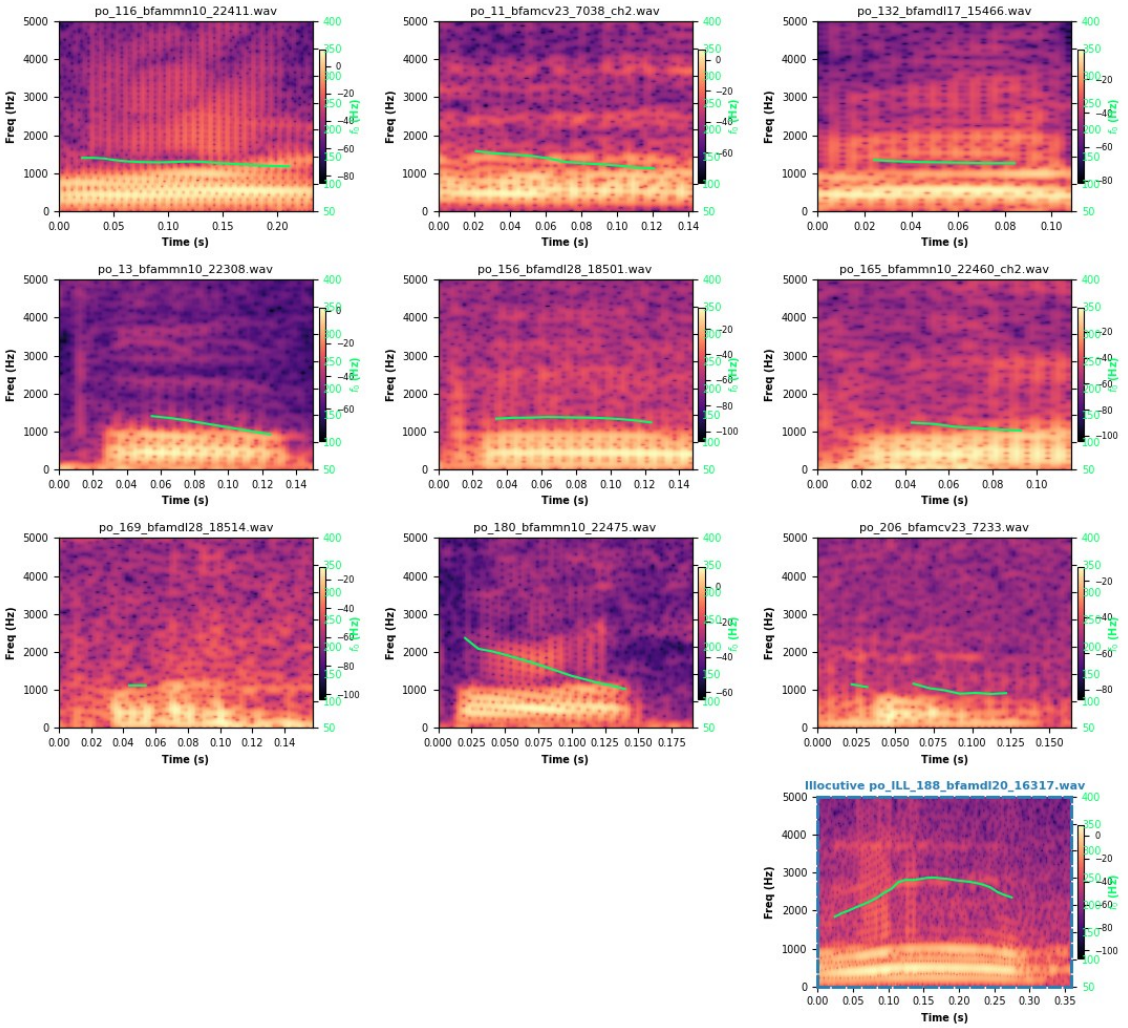


Figura 1. Espectrogramas e f_0 das ocorrências de *pô*.

A figura 1 apresenta as ocorrências individualmente e permite comparar, de modo descritivo, a trajetória de f_0 e a duração das realizações. Observa-se uma diferença entre as ocorrências analisadas preliminarmente, a diferença entre as ocorrências com contorno descendente ou nivelado e o caso destacado em azul. Nas primeiras, a duração curta e a ausência de um movimento melódico amplo são compatíveis com uma função menos expressiva e mais ligada ao gerenciamento conversacional. No caso destacado, a duração maior e o contorno ascendente-descendente dão maior proeminência à forma, permitindo sustentar uma interpretação mais expressiva. É importante ressaltar que essa interpretação se baseia na curva de f_0 e, conforme discutem Raso e Rocha (2025), a análise de funções ilocucionárias deve considerar conjuntamente a forma prosódica e os parâmetros pragmático-cognitivos do contexto.

4. Intermodalidade e estratégias de compensação

A análise intermodal entre o C-ORAL-BRASIL I e o Corpus FDP (*Reddit*) apresenta uma assimetria funcional que contribui para a hipótese central deste estudo. Conforme mapeado na Tabela 1, o registro falado privilegia a forma apocopada *pô* (~150,00 i.p.m.) em detrimento de *porra* (~65,00 i.p.m.), ao passo que o registro escrito digital inverte

essa proporção de forma acentuada, com a variante canônica atingindo 2.857,94 i.p.m. em comparação com as 89,31 i.p.m. da variante reduzida. Esta divergência pode estar relacionada às diferenças entre as modalidades, à composição temática e às situações comunicativas representadas nos *corpora*. Uma hipótese a se confirmar é a de que, na ausência de recursos prosódicos, mas em presença de outros recursos expressivos próprios da escrita digital, essa modalidade possa apresentar maior recorrência de formas lexicais plenas e de co-ocorrências com outras expressões tabu para produzir saliência pragmática. Na fala espontânea, a perda de carga ofensiva (*semantic bleaching*) e a consequente gramaticalização de um palavrão em marcador discursivo são mediadas pela prosódia. Nas ocorrências analisadas, a menor duração e os contornos nivelados ou descendentes são compatíveis com uma interpretação menos expressiva e mais ligada ao gerenciamento conversacional. Na escrita digital informal do *Reddit*, os usuários não têm esses recursos acústicos. A ausência de tais recursos pode favorecer, em determinados contextos, o uso de formas lexicais plenas e de co-ocorrências com outras expressões tabu. Essa possibilidade é compatível com a maior frequência relativa de *porra* observada na amostra, mas deverá ser estudada em análises futuras com controle dos temas e das situações comunicativas. Para compensar a ausência de entonação e garantir a expressividade do enunciado, o produtor textual parece recorrer a uma estratégia de saturação lexical. Em vez de utilizar a forma atenuada *pô*, mobiliza-se a força expressiva máxima da variante canônica *porra*, frequentemente integrada em estruturas de intensificação sintática ou em ambientes de clustering expressivo, onde co-ocorre com fórmulas como *puta que pariu* e *tomar no cu*. Quando a forma *pô* surge no *Reddit*, ela parece se concentrar em posições iniciais do comentário ou do turno (ex: "*Pô*, Lula..."), de modo semelhante ao que ocorre em certos usos da fala.

5. Considerações Finais

Este estudo preliminar sugere que a distribuição e a especialização funcional dos palavrões no PB podem ser condicionadas pelo meio de circulação. Na amostra oral analisada, a prosódia parece atuar como um mecanismo de atenuação do valor de tabu do palavrão e da aproximação de *pô* a funções discursivas. Na amostra escrita, por outro lado, observa-se uma maior frequência relativa de *porra*. Esses resultados iniciais indicam que a modalidade, a situação comunicativa e a disponibilidade de recursos prosódicos devem ser considerados na análise das formas em estudo. No entanto, as diferenças observadas não podem ser atribuídas exclusivamente à modalidade, uma vez que os *corpora* também diferem quanto ao tamanho, aos temas dos comentários e à situação comunicativa. As próximas etapas visam ampliar a amostra escrita e quantificar as co-ocorrências entre as formas.

Questões éticas

Este estudo declara, para os devidos, fins que os dados de fala analisados são anonimizados no C-ORAL-BRASIL I. As amostras de textos escritos do Corpus FDP consistem de comentários extraídos para fins de pesquisa acadêmica, com todos os identificadores de usuários dos dados analisados sendo deletados a fim de garantir total privacidade. O uso de IA generativa não foi feito neste estudo. Átila cuidou do processamento dos dados e análise acústica em Python. Victor Hugo foi responsável pela

extração e análise dos dados no *Sketch Engine*. Os dois autores estiveram envolvidos no recorte dos dados, na discussão dos resultados e aprovação da versão final.

Referências

Austin, J. L. *How to do things with words*. Oxford: Oxford University Press, 1962.

Boersma, P.; Weenink, D. *Praat: doing phonetics by computer* [Computer program]. Version 6.4.09. Amsterdam, 2026. Disponível em: http://www.praat.org/. Acesso em: 1 jun. 2026.

Jadoul, Y.; Thompsons, B.; De Boer, B. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics,* v. 71, p. 1-15, 2018.

Jay, Timothy; Janschewitz, Kristin. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, v. 4, n. 2, p. 267-288, jan. 2008. DOI 10.1515/JPLR.2008.013.

Lucente, Luciana. Funções Executiva e Reguladora do Oscilador Glotal. *In*: Arantes, Pablo; Rothe-Neves, Rui; Lucente, Luciana (Orgs.). *Entoação: Interface e Experimentação*. Campinas: Mercado de Letras, 2025. p. 187–208.

Raso, T.; Mello, H. *O Corpus C-ORAL-BRASIL I*: português falado informal. Belo Horizonte: Editora UFMG, 2012.

Raso, Tommaso; Rocha, Bruno. Prosódia: Ilocução e Atitude. *In*: Arantes, Pablo; Rothe-Neves, Rui; Lucente, Luciana (Orgs.). *Entoação: Interface e Experimentação.* Campinas: Mercado de Letras, 2025. p. 51–96.

Santos, S. M. *How to deal with Discourse Markers*: a prosodic, corpus-based, computational and experimental proposal. 2024. Tese (Doutorado em Linguística) – Université Paris-Saclay, Paris; Universidade Federal de Minas Gerais, Belo Horizonte, 2024.

A anotação morfossintática de um corpus pedagógico do kubeo em universal dependencies: desafios tipológicos e decisões linguísticas

Pedro Paulo Freire Oliveira, Thiago Blanch Pires

Instituto de Letras – Universidade de Brasília

pedro.freire@aluno.unb.edu.br, thiagocomaga@gmail.com

Resumo. *Este trabalho apresenta a construção e a anotação morfossintática de um corpus piloto e pedagógico da língua Kubeo segundo o modelo Universal Dependencies (UD), visando ampliar os recursos disponíveis para a pesquisa em Linguística de Corpus e a documentação de línguas indígenas. Compilado a partir da cartilha escolar Pamiene Mama Buei-Tukubo, o corpus foi anotado manualmente em formato CoNLL-U, contemplando categorias gramaticais, traços morfológicos (FEATS) e relações de dependência sintática. A análise mostra que fenômenos como animacidade, indexação pronominal verbal, marcação de caso e classificadores nominais exigem decisões de anotação que preservem as características tipológicas da língua, distinguindo mecanismos previstos pelo próprio UD de escolhas analíticas específicas para o Kubeo. O corpus piloto resultante constitui um recurso inicial para a descrição linguística, o ensino e futuras aplicações computacionais voltadas à língua.*

Palavras-chave: língua kubeo; anotação morfossintática; linguística de corpus.

Abstract. *This paper presents the construction and morphosyntactic annotation of a pilot, pedagogically oriented Kubeo corpus following the Universal Dependencies (UD) framework, aiming to expand linguistic resources for Corpus Linguistics and the documentation of Indigenous languages. Compiled from the school primer Pamiene Mama Buei-Tukubo, the corpus was manually annotated in CoNLL-U format, including part-of-speech tags, morphological features (FEATS), and syntactic dependency relations. The analysis shows that phenomena such as animacy, verbal pronominal indexing, case marking, and nominal classifiers require annotation decisions that preserve the language's typological properties, distinguishing mechanisms already provided by UD from analytical choices specific to Kubeo. The resulting pilot corpus provides an initial reusable resource for linguistic description, Kubeo language teaching, and future computational applications.*

Keywords: Kubeo language; morphosyntactic annotation; corpus linguistics.

1. Introdução

A Linguística de Corpus tem papel central na produção de recursos que subsidiam investigações descritivas e aplicações em Processamento de Linguagem Natural (PLN), permitindo observar regularidades gramaticais, documentar fenômenos linguísticos e disponibilizar bases de dados reutilizáveis para fins científicos e educacionais.

No contexto das línguas indígenas brasileiras, ainda são escassos os corpora morfossintaticamente anotados, o que limita estudos baseados em evidências e o

desenvolvimento de ferramentas computacionais. Iniciativas recentes, como o treebank UD para o Nheengatu (ALENCAR, 2021), demonstram a viabilidade de tais recursos mesmo a partir de corpora reduzidos, reforçando a relevância de esforços semelhantes para outras línguas amazônicas, como o Kubeo.

Entre os recursos mais utilizados para anotação sintática e morfossintática destaca-se o Universal Dependencies (UD), modelo que busca compatibilizar descrições linguísticas por meio de um conjunto comum de categorias morfológicas e relações de dependência. Sua aplicação a línguas tipologicamente distintas, porém, exige adaptações cuidadosamente fundamentadas, sobretudo em contextos de baixa disponibilidade de recursos.

A língua Kubeo, pertencente à família Tukano Oriental, apresenta características morfossintáticas que desafiam a aplicação direta de modelos universais de anotação. Fenômenos como classificadores nominais, marcação de caso, evidencialidade, animacidade e mecanismos de indexação verbal desempenham papel central na organização gramatical da língua e exigem decisões de anotação capazes de preservar informações relevantes tanto para sua descrição linguística quanto para sua representação computacional.

Este trabalho apresenta a construção e a anotação morfossintática de um corpus piloto e pedagógico da língua Kubeo, compilado a partir da cartilha Pamiene Mama Buei - Tukubo e estruturado segundo as diretrizes do UD, distinguindo, sempre que possível, o que decorre diretamente dessas diretrizes do que constitui uma decisão analítica dos autores. Além de contribuir para a documentação linguística do Kubeo, o corpus anotado constitui um recurso potencialmente reutilizável em pesquisas de Linguística de Corpus, em estudos tipológicos, no desenvolvimento de aplicações de PLN e em iniciativas de ensino e fortalecimento da língua em contexto escolar.

2. Fundamentação teórica

2.1. Universal Dependencies e a representação morfossintática

A construção de corpora morfossintaticamente anotados demanda um modelo de representação que permita descrever de forma consistente as propriedades gramaticais das línguas e, ao mesmo tempo, favoreça a comparabilidade entre diferentes recursos linguísticos. Nesse contexto, o UD constitui atualmente um dos principais padrões internacionais para anotação sintática e morfológica, adotando um conjunto compartilhado de categorias gramaticais, atributos morfológicos e relações de dependência sintática aplicáveis a diferentes famílias linguísticas (NIVRE et al., 2016; DE MARNEFFE et al., 2021).

Embora proponha um inventário universal de categorias, o UD não pressupõe uniformidade estrutural entre as línguas: suas diretrizes reconhecem que fenômenos específicos devem ser representados por decisões linguísticas fundamentadas, o que torna o modelo especialmente relevante para línguas indígenas, cuja diversidade tipológica frequentemente desafia categorias concebidas a partir de línguas amplamente documentadas.

Entre os mecanismos previstos pelo próprio esquema, destaca-se a camada de traços morfológicos (FEATS), que permite registrar propriedades linguísticas específicas

de cada língua sem alterar a arquitetura geral do modelo — mecanismo explorado de maneira central na anotação do Kubeo, como se discute na Seção 3.

2.2. O "Problema de Anchieta" e a preservação da informação morfológica

Um dos desafios recorrentes na descrição de línguas indígenas consiste em representar adequadamente estruturas morfológicas complexas sem descaracterizar sua organização gramatical. Sandalo e Galves (2017) denominam esse desafio de Problema de Anchieta, em referência à tendência histórica, inaugurada pelas gramáticas missionárias, como a de José de Anchieta para o tupi, de descrever línguas indígenas por meio de categorias gramaticais concebidas para línguas indo-europeias, o que pode obscurecer processos morfológicos próprios dessas línguas.

Embora formulada a partir da crítica às descrições coloniais, a noção permanece atual em iniciativas de anotação linguística computacional: modelos que privilegiam segmentações incompatíveis com a estrutura da língua podem comprometer a representação de relações morfológicas fundamentais, dificultando tanto a descrição linguística quanto a reutilização dos dados em pesquisas baseadas em corpus. Sob essa perspectiva, a construção de um corpus anotado requer o equilíbrio entre dois objetivos complementares: manter a compatibilidade com um padrão internacional de representação, como o UD, e preservar informações linguísticas essenciais para a descrição da língua documentada.

2.3. Implicações para a anotação do Kubeo

A língua Kubeo apresenta características morfossintáticas — como classificadores nominais, marcação de caso, animacidade, evidencialidade, indexação verbal e aglutinação — que tornam esse equilíbrio particularmente relevante e nem sempre encontram correspondência direta nas categorias dos modelos universais de anotação.

Entre esses fenômenos, dois têm impacto direto sobre as decisões de anotação discutidas neste trabalho. Primeiro, a animacidade constitui uma distinção morfológica obrigatória no Kubeo, afetando a categorização dos nomes e seu comportamento na estrutura argumental (CHACON, 2012). Segundo, a indexação de argumentos ocorre predominantemente por meio de marcas de pessoa, gênero e número incorporadas ao complexo verbal — com frequência a um auxiliar —, e não por pronomes independentes (CHACON; GENETTI, 2019). Essa incorporação pronominal exige que a anotação registre, na camada FEATS do elemento verbal ou auxiliar, informações que, em línguas de marcação predominantemente nominal, estariam associadas a um argumento lexicalmente realizado.

Dessa forma, a construção do corpus anotado exigiu decisões linguísticas fundamentadas na descrição gramatical do Kubeo, buscando preservar informações morfológicas relevantes sem comprometer a compatibilidade com as diretrizes do UD. Nesse processo, convém distinguir dois tipos de decisão: de um lado, o uso extensivo da camada FEATS para registrar animacidade, caso e indexação verbal segue diretamente o mecanismo de extensibilidade previsto pelas próprias diretrizes do UD; de outro, a escolha de manter certas formas aglutinadas como um único token — em vez de segmentá-las em unidades morfológicas independentes — constitui uma decisão analítica

dos autores, fundamentada na descrição gramatical do Kubeo e documentada em um manual interno de anotação (Seção 3.2).

Essas decisões evidenciam que a anotação morfossintática é também um processo de investigação linguística, e não apenas uma etapa de preparação de dados — perspectiva que orienta a análise de exemplos concretos apresentada a seguir.

3. Metodologia

3.1. Compilação do corpus

O corpus analisado foi compilado a partir da cartilha pedagógica Pamiene Mama Buei-Tukubo (CHACON, 2023), organizada para o ensino da língua Kubeo em contexto escolar, em parceria com a comunidade e a escola indígena responsável pelo material. A seleção das sentenças seguiu leitura integral do material didático e filtragem por critérios de representatividade e complexidade crescente, priorizando sentenças mono-oracionais e, em seguida, exemplos com classificadores nominais, marcação de caso, animacidade, indexação verbal e construções existenciais.

Esse processo resultou em um corpus piloto de 30 sentenças. A cartilha tem 77 páginas organizadas em seis unidades: as quatro primeiras dedicadas ao sistema de escrita (letras, sílabas, nasalidade, tons e acentos) e as Unidades 5 e 6 dedicadas, respectivamente, a canções e a narrativas tradicionais do povo Kubeo — seções das quais foram extraídas prioritariamente as sentenças do corpus piloto, por reunirem estruturas autênticas em contextos poéticos e narrativos.

As sentenças foram transcritas, organizadas em formato digital e estruturadas no padrão CoNLL-U. Paralelamente, utilizou-se a base lexical previamente desenvolvida no FieldWorks Language Explorer (FLEx), o que padronizou lemas, categorias gramaticais e informações morfológicas ao longo da anotação.

3.2. Anotação morfossintática

A anotação foi realizada manualmente, com cada sentença analisada à luz da descrição gramatical de referência do Kubeo (CHACON, 2012; CHACON; GENETTI, 2019). As decisões recorrentes foram sistematizadas em um manual interno de anotação, revisado em conjunto com o linguista responsável pela descrição de referência da língua, assegurando consistência interna e fidelidade à gramática do Kubeo.

Como discutido na Seção 2.3, essas decisões distinguem mecanismos previstos pelo UD (uso da camada FEATS) de escolhas analíticas específicas do Kubeo (segmentação de formas aglutinadas), detalhadas na análise a seguir. Os exemplos discutidos neste artigo foram selecionados por representarem fenômenos morfossintáticos cuja anotação exigiu esse tipo de adaptação fundamentada às diretrizes gerais do UD, evidenciando os desafios envolvidos na representação de uma língua indígena em um modelo universal de anotação.

4. Análise Linguística e Resultados Preliminares

A análise do corpus piloto revelou que reduzir o Kubeo às etiquetas simples do UD gera perda descritiva relevante, caso não sejam exploradas as camadas de traços morfológicos e de dependência. A Tabela 1 ilustra a anotação completa, em CoNLL-U, da sentença 1 do corpus.

Tabela 1. Anotação morfossintática da sentença 1 ("Yawi oko hārāwʉi āino wore kuibi.", 'A onça procura comida no dia de chuva') no formato CoNLL-U.

ID	Forma	Lema (Glosa)	UPOS	Feats	DepRel	Head
1	Yawi	yawi (onça)	NOUN	Animacy=Anim| Gender=Masc| Number=Sing	nsubj	5
2	oko	oko (água/chuva)	NOUN	Animacy=Inan| Number=Sing	nmod	3
3	hārāwʉi	hārāwʉ (dia) -i (LOC)	NOUN	Animacy=Inan| Case=Loc	obl	5
4	āino	āino (comida)	NOUN	Animacy=Inan| Case=Acc| Number=Sing	obj	5
5	wore	wo (procurar) -re	VERB	VerbForm=Vnoun	root	0
6	kuibi	kui (estar/AUX) -bi (3M)	AUX	Gender=Masc| Number=Sing| Person=3	aux	5

Na sentença analisada, observa-se a flexibilidade da ordem dos constituintes e a separação linear entre o sujeito nominal explícito (Yawi, "onça", ID 1) e o verbo núcleo (wore, ID 5), do qual dependem também o modificador locativo hārāwʉi (obl), o objeto āino (obj) e o auxiliar kuibi (aux). A anotação evidencia a distinção obrigatória de animacidade prevista pela gramática do Kubeo: o sujeito Yawi recebe o traço Animacy=Anim, ao passo que os demais nomes da sentença — oko ("chuva"), hārāwʉi ("dia", com marca locativa) e āino ("comida") — recebem Animacy=Inan, respeitando os eixos ontológicos que regem a gramática da língua. O pronome indexado ao final da sentença, incorporado ao auxiliar kuibi ("estar" + 3ª pessoa masculina singular), ilustra o mecanismo de indexação pronominal discutido na Seção 2.3: em vez de um pronome sujeito independente, a concordância com o argumento nominal é registrada nos traços Gender, Number e Person do próprio auxiliar (CHACON; GENETTI, 2019).

Do ponto de vista da anotação, a decisão de manter a palavra hārāwʉi ("no dia de chuva") como um único token, classificado como NOUN e estendido com o traço morfológico Case=Loc, evita a fragmentação artificial de aglutinações nativas em preposições isoladas (ADP) que não existem de forma independente na língua — trata-se, como discutido na Seção 3.2, de uma decisão analítica dos autores, e não de uma exigência das diretrizes gerais do UD. Já o uso do traço Animacy e a codificação da indexação verbal em FEATS seguem diretamente o mecanismo de extensibilidade previsto pelo próprio modelo.

Um levantamento preliminar da relação entre sujeito nominal explícito e verbo núcleo no conjunto do corpus piloto indica um padrão compatível com o exemplo discutido: sujeitos lexicalmente realizados ocorrem em cerca de 90% das sentenças, com distância linear entre 1 e 4 tokens em relação ao verbo (média aproximada de 2,5) e forte predominância da ordem Sujeito-Verbo. Embora esses números devam ser interpretados com cautela, dada a reduzida extensão do corpus piloto, eles sugerem que o protocolo de anotação adotado é capaz de representar de forma consistente tanto dependências locais quanto dependências de maior alcance entre sujeito e verbo.

5. Considerações Finais

A estruturação de um corpus da língua Kubeo orientado pelo framework Universal Dependencies demonstra-se uma estratégia viável, desde que conduzida com rigor e sensibilidade tipológica: a anotação morfossintática, ao lidar com fenômenos como animacidade, indexação pronominal e aglutinação, exige a exploração sistemática da camada FEATS para enfrentar o Problema de Anchieta (SANDALO; GALVES, 2017) sem descaracterizar a gramática nativa.

Deve-se ressaltar que os resultados aqui apresentados são preliminares e derivam de um corpus piloto ainda reduzido (30 sentenças); por isso, mais do que consolidar um modelo empírico definitivo, este trabalho estabelece um protocolo de anotação reprodutível e um ponto de partida metodológico para etapas futuras. Com consolidação prevista para o primeiro semestre de 2027, as próximas etapas incluem a expansão do corpus anotado e o aprofundamento na anotação de fenômenos ainda pouco explorados, como a serialização verbal e os marcadores de evidencialidade, fortalecendo a presença digital, educacional e científica da língua Kubeo.

Questões Éticas

O corpus deriva de uma cartilha pedagógica publicada para o ensino da língua Kubeo em contexto escolar, organizada por um pesquisador com atuação de longa data junto à comunidade e à escola indígena responsável pelo material (CHACON, 2023); nenhuma coleta de dados diretamente junto a falantes foi realizada nesta etapa. As decisões de anotação foram revisadas em conjunto com o linguista responsável pela descrição de referência do Kubeo, resguardando a fidelidade da análise à gramática da língua e às escolhas já validadas pela comunidade no material de origem. Ferramentas de inteligência artificial (assistente de linguagem) foram utilizadas como apoio na revisão textual deste artigo; a anotação linguística e as análises apresentadas são de autoria e responsabilidade exclusiva do autor.

Referências

ALENCAR, L. F. de. Uma gramática computacional de um fragmento do nheengatu. *Revista Estudos da Linguagem, Belo Horizonte*, v. 29, n. 3, p. 1717-1777, 2021.

CHACON, T. C. *The phonology and morphology of Kubeo: the documentation, theory, and description of an Amazonian language*. Tese (Doutorado em Linguística) – University of Hawai‘i at Manoa, Honolulu, 2012.

CHACON, T. C. (org.). *Pamiene Mama Buei-Tukubo: primeiro livro para o ensino da língua Kubeo.* Escola Estadual Indígena Huredanaçu Nossa Senhora da Imaculada Conceição; SEDUC-AM, 2023.

CHACON, T. C.; GENETTI, C. Grammatical relations in Kubeo. In: WITZLACK-MAKAREVICH, A.; BICKEL, B. (eds.). *Argument Selectors*: A new perspective on grammatical relations. Amsterdam: John Benjamins, 2019. p. 399-432.

DE MARNEFFE, M. C. et al. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255-308, 2021.

NIVRE, J. et al. Universal Dependencies v1: a multilingual treebank collection. In: *INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC)*, 10., 2016, Portorož. Proceedings… Portorož: European Language Resources Association, 2016. p. 1659-1666.

SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do Brasil: o problema de Anchieta. *Cadernos de Estudos Linguísticos*, Campinas, v. 59, n. 3, p. 631-648, 2017.

Preparação auditável de corpus histórico: um pipeline em camadas para análise linguística baseada em corpus

Luiz Henrique Santos Torres

Mestrando em Ciências da Linguagem na Universidade NOVA de Lisboa/NOVA FCSH
Avenida Berna 26 C, 1069-061 Lisboa

luizhstorres@gmail.com

Abstract. *This work-in-progress reports preliminary results from the auditable preparation of a historical corpus based on The Secret Doctrine, vol. I (1888). The corpus is mainly in English and includes multilingual material. The layered pipeline preserves RAW and intermediate versions and links human-reviewed decisions to versioned rules, logs, hashes, residual scans, and post-stage diagnostics, supporting traceability and revision. Initial diagnosis, dehyphenation, paratext treatment, and running-header/pagination treatment have been completed and validated; review of 681 footnote candidates remains in progress. The contribution is a controlled preparation procedure prior to corpus-based linguistic analysis.*

Resumo. *Este trabalho em andamento apresenta resultados preliminares da preparação auditável de um corpus histórico baseado em The Secret Doctrine, vol. I (1888). A língua principal é o inglês, com material multilíngue. O pipeline em camadas preserva RAW e versões intermediárias e articula decisões revistas por humano a regras versionadas, logs, hashes, varreduras residuais e diagnósticos pós-etapa, favorecendo rastreabilidade e revisão. Foram concluídos e validados o diagnóstico inicial, a hifenização e o tratamento de paratextos e cabeçalhos/paginação; a revisão de 681 candidatos a notas permanece em andamento. A contribuição é um procedimento controlado anterior à análise linguística baseada em corpus.*

1. Introdução

A Linguística de Corpus pressupõe que decisões de seleção, preparação, anotação e limpeza não são neutras, pois afetam medidas como frequência, dispersão, colocação, palavras-chave e padrões lexicais (Sinclair, 1991; Biber et al., 1998; McEnery & Hardie, 2012); em estudos assistidos por corpus, essas decisões também sustentam inferências qualitativas (Baker, 2006).

Esse problema é particularmente relevante em corpora históricos digitalizados. Edições antigas, reedições digitais e processos de OCR podem introduzir ruídos de paginação, cabeçalhos correntes, hifenização por quebra de linha, índices, paratextos editoriais, notas integradas ao fluxo principal e caracteres especiais, afetando medidas quantitativas de corpus (Gries, 2008; Brezina, 2018; Egbert & Biber, 2019). Estudos sobre OCR histórico reforçam a necessidade de curadoria e validação explícitas (Hill & Hengchen, 2019; Reul et al., 2018). No contexto brasileiro, Paixão de Sousa (2014), em estudo sobre o Corpus Tycho Brahe, discute a construção de corpora históricos eletrônicos em termos de decisões de representação e da combinação de procedimentos filológicos, linguísticos e computacionais.

Este artigo apresenta resultados preliminares de um estudo-piloto em andamento

dedicado à preparação auditável de corpus histórico, tomando como objeto o primeiro volume de *The Secret Doctrine*, de H. P. Blavatsky. A língua principal do corpus é o inglês, embora a obra inclua termos, citações e sequências em outras línguas e sistemas de escrita. Sua complexidade editorial, notas autorais, terminologia, material multilíngue, transliterações, fórmulas e paratextos fazem do volume um objeto-piloto metodologicamente exigente para testar uma preparação controlada em camadas. O foco recai sobre a preparação do corpus, anterior à análise lexical ou discursiva propriamente dita.

O objetivo é demonstrar como um texto histórico complexo pode ser transformado em dados mais adequados à análise linguística por meio de camadas sucessivas de processamento, revisão e validação, tornando a preparação textual rastreável e revisável em vez de um procedimento invisível.

2. Corpus piloto e pipeline

O estudo-piloto toma como objeto o primeiro volume de *The Secret Doctrine*, de H. P. Blavatsky, publicado originalmente em 1888. A escolha da obra justifica-se por sua relevância histórica no desenvolvimento do discurso teosófico moderno e pela complexidade textual já descrita: extensão, densidade terminológica, referências intertextuais, notas autorais, vocabulário especializado, termos em línguas antigas e modernas, transliterações e alternância entre exposição doutrinária, comentário, citação e discussão crítica.

O corpus foi derivado do ebook em PDF de *The Secret Doctrine*, Vol. I - *Cosmogenesis*, publicado pela Theosophical University Press em 2014 como *Unabridged Verbatim Edition*, reprodução da primeira edição de 1888. O PDF foi convertido localmente para TXT com PDF-XChange Editor, e o TXT resultante foi preservado como camada RAW antes de qualquer transformação. O diagnóstico inicial registrou hifenização por quebra de linha, material editorial e índice, cabeçalhos correntes, paginação e notas de rodapé integradas ao fluxo principal.

Esses elementos têm efeitos diferentes sobre a análise de corpus: a hifenização pode fragmentar itens lexicais; cabeçalhos e paginação podem introduzir repetições artificiais; índices e paratextos podem deslocar frequências; e notas podem misturar comentário autoral, referência bibliográfica e discussão paralela ao texto principal. Ao mesmo tempo, nem todo elemento aparentemente ruidoso deve ser removido, pois notas, números, termos estrangeiros e caracteres não ASCII podem carregar informação textual relevante.

Por essa razão, o pipeline foi organizado em etapas sucessivas, com revisão humana dos candidatos antes da aplicação determinística das decisões e diagnóstico posterior. Em termos operacionais, "auditável" significa preservar o estado anterior (RAW e versões intermediárias), registrar decisões e transformações em registros de revisão, regras versionadas e logs, e verificar cada etapa com hashes, varreduras residuais e diagnósticos pós-etapa, permitindo rastrear e revisitar decisões. O fluxo geral é: RAW → diagnóstico inicial → dehyphenation → paratexto/índice → cabeçalhos/paginação → notas de rodapé → reconstrução de parágrafos → CLEAN_MAIN → NOTES → chunks analíticos.

Na dehyphenation, os candidatos foram revisados para distinguir hifenização

produzida por quebra editorial de linha de casos em que o hífen devia permanecer como parte da forma lexical ou histórica; as decisões revisadas foram aplicadas deterministicamente e submetidas a diagnóstico pós-etapa. Um exemplo real é *Sacred Sci- / ence → Sacred Science*, em que a barra representa a quebra de linha. No tratamento de paratextos, material editorial moderno e índice foram retirados do CLEAN_MAIN e preservados separadamente, enquanto material autoral relevante, como dedicatória e prefácio, foi mantido. Para cabeçalhos/paginação, candidatos recorrentes foram revisados para distinguir material editorial de conteúdo legítimo e verificados por diagnóstico residual. A etapa de notas de rodapé continua sob revisão humana, sem transformação final reportada neste estágio.

3. Resultados preliminares

Os resultados apresentados dizem respeito à preparação do corpus, e não à análise lexical ou discursiva final. Eles registram o estado validado das etapas iniciais do pipeline e o estágio ainda em revisão das notas de rodapé.

Tabela 1. Resultados preliminares do pipeline

Etapa	Resultado preliminar
Diagnóstico inicial	identificação de hifenização, paratexto, cabeçalhos, notas e caracteres especiais
Hifenização por quebra de linha	1.562
Hífens removidos	1.437
Hífens preservados	125
Paratexto, metadados e índice	368 linhas removidas ou movidas
Cabeçalhos/paginação - 1ª rodada	664 linhas removidas
Cabeçalhos residuais - correção	30 linhas removidas
Versão intermediária validada	28.492 linhas
Tokens operacionais	318.011
Diagnóstico pós-cabeçalhos/paginação	PASS; 0 flags de risco
Candidatos a notas de rodapé	681
Recomendados para extração	658
Com revisão de limite	23

Os números evidenciam transformações controladas: a correção de hifenização evita fragmentação lexical; o tratamento de cabeçalhos/paginação reduz repetições editoriais; e a separação entre paratexto editorial e material autoral preserva elementos como dedicatória e prefácio sem manter no corpus principal índice, metadados ou material editorial moderno. Foram removidas 694 linhas de cabeçalhos/paginação nas duas rodadas registradas na Tabela 1. Na Tabela 1, PASS indica que as verificações configuradas para a etapa não produziram condições de não aprovação; no diagnóstico de cabeçalhos/paginação, isso resumiu ausência de flags de risco, verificações não aprovadas, candidatos residuais inesperados e falhas de preservação. A contagem de 318.011 é operacional, usada nos diagnósticos do pipeline, considera caracteres Unicode e trata formas com hífen ou apóstrofo interno como uma unidade; ela não pressupõe a

mesma definição de token em ferramentas analíticas posteriores.

Um ponto importante foi a identificação de falsos positivos. Na revisão residual de cabeçalhos/paginação, 31 candidatos foram examinados; 30 foram removidos como cabeçalhos com numeração romana e um foi preservado por corresponder a material legítimo de nota/fórmula, e não a paginação. O caso ilustra a necessidade de combinar detecção automática, revisão humana e diagnóstico posterior.

A etapa atualmente em curso envolve a revisão dos candidatos a notas de rodapé. Foram identificados 681 candidatos, dos quais 658 foram inicialmente recomendados para extração; outros 23 exigem revisão de limite. Esses valores descrevem uma revisão ainda em andamento: a extração final das notas não é apresentada como concluída.

4. Discussão e próximos passos

Os resultados preliminares indicam que a preparação de corpora históricos complexos deve ser tratada como etapa metodológica substantiva. No caso analisado, distinguir ruído editorial de material textual potencialmente relevante exigiu critérios diferentes para hifenização, paratextos, cabeçalhos, paginação e notas, combinando detecção automática, revisão humana e diagnóstico pós-etapa.

A organização em camadas reduz o risco de tornar invisíveis essas decisões. Ao preservar RAW, versões intermediárias, texto principal, registros de revisão, regras, logs e diagnósticos, o processo permite retornar a decisões anteriores e verificar como cada transformação afetou o corpus, em consonância com princípios de documentação, rastreabilidade e reutilização de dados de pesquisa (Wilkinson et al., 2016).

O trabalho apresenta limites. O pipeline foi aplicado, até o momento, a um volume de uma única obra, *The Secret Doctrine*, e sua generalização a outros corpora históricos não está estabelecida. A etapa de notas de rodapé ainda está em revisão, e os resultados reportados dizem respeito à preparação do corpus, não à análise lexical ou discursiva final. Os próximos passos incluem concluir a extração das notas para NOTES, realizar a reconstrução controlada de parágrafos, gerar uma versão CLEAN_MAIN estável e segmentar o texto em chunks analíticos.

Questões éticas

O estudo não envolve participantes humanos, dados pessoais ou dados sensíveis; o corpus é uma obra publicada em 1888. IA generativa foi usada apenas como apoio à revisão linguística, condensação e organização do manuscrito, com revisão e decisão final humanas. As transformações do corpus descritas no estudo permanecem documentadas por regras determinísticas, logs e diagnósticos. O autor foi responsável pela concepção do estudo, preparação e validação do corpus, análise dos dados e redação e revisão do manuscrito.

Referências

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Blavatsky, H. P. (1888). *The secret doctrine: The synthesis of science, religion, and philosophy* (Vol. 1: Cosmogenesis). The Theosophical Publishing Company.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77-104.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437.

Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities, 34*(4), 825-843.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Paixão de Sousa, M. C. (2014). O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa, 16*(n. esp.), 53-93. https://doi.org/10.11606/issn.2176-9419.v16ispep53-93

Reul, C., Springmann, U., Wick, C., & Puppe, F. (2018). Improving OCR accuracy on early printed books by combining pretraining, voting, and active learning. *Journal for Language Technology and Computational Linguistics, 33*(1), 3-24.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018.

Investigando marcas de autoria em contos atribuídos a Machado de Assis

Ursula Puello Sydio

Universidade de São Paulo (USP)

ursula.sydio@usp.br

Resumo. *Na década de 1950, o crítico Raimundo Magalhães Júnior atribuiu ao Machado de Assis a autoria de dezenove contos publicados sob pseudônimos no século XIX. Entretanto, a ainda não há consenso da crítica machadiana a respeito da autoria desses contos. Aproveitando essa lacuna, os recursos computacionais da Linguística de Corpus (LC) foram usados para encontrar marcas de autoria que corroborem ou não com a hipótese de Magalhães Júnior. Este trabalho visa apresentar os resultados encontrados, graças a LC, de marcas de autoria externa, comparando o corpus Pseudônimos sob Investigação (PI), compostos por esses dezenove contos e divididos em treze subcorpora, com outro corpus intitulado Obra Completa (OC), composto por duzentas e seis obras cuja autoria machadiana não é disputada.*

Abstract. *In the 1950s, the literary critic Raimundo Magalhães Júnior attributed to Brazilian writer Machado de Assis the authorship of nineteen short stories published under pseudonyms in the 19th century. However, there is no consensus among Machadian scholars on the authorship of these works. In order to fill this gap, the computational resources of Corpus Linguistics (CL) were used to map the authorship markers which may support or debunk Magalhães Júnior's hypothesis. This research aims to present the findings obtained for external authorship markers, due to the CL, comparing the corpus titled Pseudonyms under Investigation (PI), comprised by these nineteen short stories and divided into thirteen subcorpora, to the corpus Complete Works (CW), composed of 206 works whose Machadian authorship is undisputed.*

1. Introdução

Machado de Assis é um dos autores mais estudados da literatura brasileira e publicou romances, contos, crônicas, peças e poemas, ora com o próprio nome ou iniciais, ora sob pseudônimos. As obras publicadas sob pseudônimo hoje incorporadas ao cânone machadiano, com autoria consensual, baseiam-se nos achados de Galante de Sousa, que reuniu vinte e três pseudônimos na *Bibliografia de Machado de Assis* (1955).

Na década de 1950, Raimundo Magalhães Júnior organizou cinco volumes de contos machadianos publicados pela Civilização Brasileira. Entre textos já consagrados, incluiu 19 obras assinadas por pseudônimos ou em anonimato e publicadas em jornais e periódicos das décadas de 1860 e 1870. Nos prefácios, apresentou razões para atribuí-las a Machado de Assis. Contudo, esses pseudônimos não foram corroborados por Galante de Sousa e, apesar da vasta fortuna crítica sobre o escritor, a hipótese de Magalhães Júnior permanece em aberto (Guimarães, 2016; Suriani da Silva, 2019).

Assim, esta pesquisa busca identificar marcas de autoria que aproximem ou afastem a hipótese de autoria machadiana desses dezenove contos. Os avanços da

Linguística de Corpus e de áreas afins permitem aplicar métodos indisponíveis na época de Magalhães Júnior. A relevância duradoura de Machado de Assis na academia, na escola e no mercado editorial internacional justifica essa investigação.

2. Pressupostos teóricos

Esta seção apresenta a descoberta dos dezenove contos por Raimundo Magalhães Júnior e discute pesquisas recentes de atribuição de autoria na obra de Machado de Assis que fundamentam a metodologia aqui adotada.

Raimundo Magalhães Júnior, crítico literário, jornalista e biógrafo de Machado de Assis, dedicou-se ao estudo das obras do autor publicadas em periódicos do século XIX, especialmente nas coleções do *Jornal das Famílias* da Biblioteca Nacional, com o propósito de divulgá-las e preservá-las em livro (Santos, 2021, p. 80). Dessa pesquisa resultou uma coleção de cinco volumes de contos machadianos, publicada pela Civilização Brasileira em 1956. Nesses cinco volumes, entre obras de autoria não disputada, identifiquei dezenove contos. A Figura 1 reúne seus títulos, os pseudônimos que os assinam, os jornais ou periódicos de publicação original, os anos de publicação e as coletâneas em que foram posteriormente incluídos.

Título	Pseudônimo	Jornal	Ano	Coletânea
"Fernando e Fernanda"	Máximo	Jornal das Famílias	1866	Contos Recolhidos
"Francisca"	Máximo	Jornal das Famílias	1867	Contos Recolhidos
"Onda"	Máximo	Jornal das Famílias	1867	Contos avulsos
"Curiosidade"	M.	A Estação	1879	Contos sem data
"O Pai"	M.	Jornal das Famílias	1866	Contos Recolhidos
"O Teles e o Tobias (Quadro de Costumes Políticos)"	M. *** anônimo	Semana Ilustrada	1865	Contos avulsos
"Diana"	anônimo	Jornal das Famílias	1866	Contos avulsos
"A Vida Eterna"	Camilo da Anunciação	Jornal das Famílias	1870	Contos avulsos
"Possível e Impossível",	Marco Aurélio	Jornal das Famílias	1867	Contos avulsos
"Incorrigível"	***	A Estação	1884	Contos sem data
"Duas Juízas"	Próspero	A Estação	1883	Contos sem data
"Um Bilhete"	z. z. z.	A Estação	1885	Contos sem data
"História de uma lágrima"	J. B.	Jornal das Famílias	1867	Contos sem data
"A Menina dos Olhos Pardos"	Otto	Jornal das Famílias	1873	Contos Esparsos
"Quinhentos Contos"	Otto	Jornal das Famílias	1868	Contos Esparsos
"Canseiras em Vão"	o.o.	Jornal das Famílias	1872	Contos Esparsos
"Quem Boa Cama Faz"	o.o.	Jornal das Famílias	1875	Contos Esparsos
"A Felicidade"	x.	Jornal das Famílias	1871	Contos Esparsos
"Felicidade pelo Casamento"	F. S.	Jornal das Famílias	1866	Contos Esparsos

Figura 1. A figura apresenta uma lista dos contos investigados na pesquisa bem como informações relevantes à obra

Na última década, as principais contribuições para a investigação de autoria de textos machadianos "perdidos" no anonimato vieram de Sílvia Maria Azevedo (2019) e Fernando Borsato de Santos (2019). Azevedo investigou as crônicas da coluna "Badaladas", da *Semana Ilustrada*, assinadas pelo pseudônimo Dr. Semana, propondo uma metodologia baseada em marcas de autoria externas e internas ao texto. Suas categorias — temas recorrentes, referências históricas, culturais e literárias, citações e recursos estilísticos — inspiraram as categorias adotadas nesta pesquisa. Em estudo

derivado de seu doutorado, Borsato (2019) identificou um novo pseudônimo com base no conceito de autorreferência (Love, 2002), concluindo que o criptônimo "?" pertencia a Machado de Assis. Se a hipótese de Magalhães Júnior se confirmar, esse não seria o único criptônimo do autor.

3. Metodologia

Com base nas categorias de marca de autoria externa usadas por Azevedo na investigação do Dr. Semana (Assis; Azevedo, 2019) e no site *Referências na Ficção Machadiana* (Senna, 2008), as referências encontradas no PI e no OC foram organizadas nas categorias a seguir:

Tabela 1. Divisão das categorias de referências da obra machadiana

Categoria	Definição da categoria
Topônimos – Rio de Janeiro	Bairros, logradouros, estabelecimentos, instituições e outros locais dentro da cidade do Rio de Janeiro
Topônimos – Geral	Outras cidades, estados e países
Personalidades contemporâneas	Pessoas célebres contemporâneas a Machado de Assis
Personalidades históricas	Pessoas que fazem parte da História, como reis, imperadores e filósofos
Eventos políticos	Instituições e fatos históricos da política brasileira
Autores	Escritores, dramaturgos e músicos
Obras	Obras literárias, óperas, sonatas e outras obras citadas
Personagens literárias	Personagens de outros autores
Greco-romanas	Referências a mitologia, obras e filósofos da Antiguidade greco-romana
Bíblicas	Referências a passagens, livros e personagens da Bíblia
Jornais e revistas	Jornais, revistas e publicações da época de Machado

Para investigar a recorrência dessas referências, foi necessário compilar os corpora. O corpus de estudo, Pseudônimos sob Investigação (PI), reúne os dezenove contos de autoria questionada e foi dividido em treze subcorpora conforme o pseudônimo que assina cada obra, como mostra a tabela abaixo.

Tabela 2. Os treze subcorpora do PI e os contos que os compõe

	Pseudônimo	Contos assinados pelo Pseudônimo
1	Máximo	"Fernando e Fernanda"; "Francisca"; "Onda".
2	M.	"Curiosidade"; "O Pai"; "O Teles e o Tobias (Quadro de Costumes Políticos)"
3	***	"Incorrigível"
4	Anônimo	"Diana"
5	Camilo da Anunciação	"A Vida Eterna"
6	Marco Aurélio	"Possível e Impossível"
7	Próspero	"Duas Juízas"
8	z.z.z.	"Um Bilhete"
9	J. B.	"História de uma lágrima"
10	Otto	"A Menina dos Olhos Pardos"; "Quinhentos Contos"
11	O. O.	"Canseiras em Vão"; "Quem Boa Cama Faz"
12	x.	"A Felicidade"

13	F. e S.	"Felicidade pelo Casamento"

O corpus de referência, intitulado Obra Completa (OC), é composto exclusivamente por obras literárias de Machado de Assis com autoria consensual: 9 romances, 190 contos e outras 7 narrativas. Com os corpora compilados, utilizei o software AntConc para extrair listas de N-grams e, sobretudo, a ferramenta KWIC para pesquisar nos corpora as referências apresentadas na Tabela 1. O fluxo de trabalho está resumido na Figura 2.

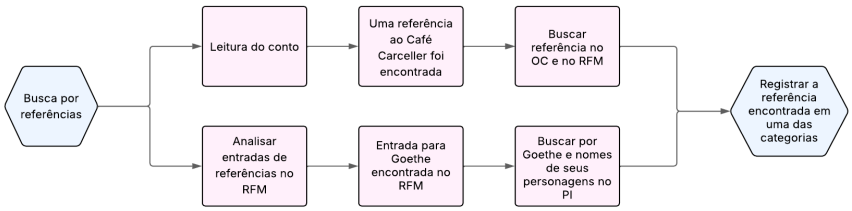


Figura 2. Fluxograma com etapas da pesquisa por referências

Os resultados da pesquisa por referências foram armazenados em bancos de dados e serão apresentados na próxima seção.

4. Resultados iniciais e discussão

Foram encontradas referências cruzadas entre os corpora PI e OC em todas as categorias apresentadas na Tabela 1. Entretanto, dada a natureza da apresentação em formato de pôster e a limitação de tempo para discutir todas as categorias, este trabalho discute apenas os resultados relativos à categoria dos Topônimos, por sua expressividade e relevância para a caracterização da escrita do autor. O cotejo entre PI e OC revelou diversas referências cruzadas a logradouros do Rio de Janeiro oitocentista. Ruas centrais e bairros tradicionais como Andaraí, Catumbi, Gamboa, Largo do Paço, Largo de São Francisco, Praia dos Mineiros, Rua da Ajuda, Rua da Assembleia, Rua da Vala e Tijuca aparecem tanto nas narrativas do PI — sobretudo em "Canseiras em Vão", assinado por O. O. — quanto em romances do OC, como *Quincas Borba* e *Esaú e Jacó*. A Rua do Ouvidor, por exemplo, ocorre 150 vezes no OC e aparece quatro vezes em três contos do PI, apesar do menor tamanho desse corpus. O Café Carceller, por sua vez, surge quatro vezes no PI, nos contos "Canseiras em Vão" e "A Felicidade", e 26 vezes no OC, em obras como "Miss Dollar" e *Quincas Borba*. Em contraste, topônimos fora da capital, como Paris, Madrid e Porto Alegre, surgem em ambos os corpora apenas como referências passageiras ou comparações culturais, sem descrição realista de seus espaços urbanos.

Para visualizar as demais marcas de autoria encontradas na categoria Topônimos – Rio de Janeiro, consulte o Anexo I após as Referências, que organiza o banco de dados de topônimos em três colunas: a referência, as obras do PI em que ocorre e as obras correspondentes do OC.

5. Próximos passos

O procedimento usado para comparar os topônimos cariocas foi repetido nas demais categorias, como menções a personalidades contemporâneas e históricas, eventos

políticos, autores, obras, personagens literárias, referências greco-romanas e bíblicas, além de jornais e revistas. Esses resultados serão discutidos em maior detalhe na tese. Paralelamente, os corpora seguem sendo analisados à luz de outras abordagens da Linguística Forense e da Estilometria.

Referências

Assis, M. de. (2015) *Obra completa*. 3ª ed. São Paulo: Editora Nova Aguilar.

Assis, M. de; Magalhães Júnior, R. (1985a) *Contos Esquecidos*. Rio de Janeiro: Ediouro.

Assis, M. de; Magalhães Júnior, R. (1985b) *Contos sem Data*. Rio de Janeiro: Ediouro.

Assis, M. de; Magalhães Júnior, R. (1986) *Contos Recolhidos*. Rio de Janeiro: Ediouro.

Assis, M. de; Magalhães Júnior, R. (1956) *Contos Avulsos*. Rio de Janeiro: Civilização Brasileira.

Assis, M. de; Magalhães Júnior, R. (1996). *Contos Esparsos*. Rio de Janeiro: Ediouro.

Azevedo, S. M. (org.). (2019) *Badaladas Dr. Semana*, por Machado de Assis: crônicas de Machado de Assis. São Paulo: Editora Unesp.

Borsato, F. R. (2019) "O pseudônimo "?" e a produção dispersa de Machado de Assis nos jornais do século XIX". Machado de Assis em Linha, v. 12, n. 28, p. 55–72.

Galante de Sousa, J. (1955) *Bibliografia de Machado de Assis*. Rio de Janeiro: Instituto Nacional do Livro.

Guimarães, H. S. (2016) *Machado de Assis, o escritor que nos lê*. São Paulo: Editora Unesp.

Lawrence, Anthony. (2004) AntConc. Tóquio, Japão: Waseda University. Disponível em: http://www.antlab.sci.waseda.ac.jp/

Love, H. (2002) *Attributing autorship*: an introduction. Cambridge: Cambridge University press.

Senna, M. (2008) *Referências na ficção machadiana*, 2008. Disponível em: https://machadodeassis.net/.

Suriani da Silva, A. (2019). "Os contos de Machado de Assis: Periodicidade e ficção no Brasil no século XIX". 2019.

Anexo I - Topônimos cariocas no PI e no OC

Lugar	Obra do PI	Obra do OC
Andaraí	Felicidade pelo Casamento Canseiras em Vão	O Caso da Viúva Dom Casmurro Dona Benedita Esaú e Jacó Helena Memorial de Aires A Mulher Pálida Qual Dos Dois? Só! Três Tesouros Perdidos Um Erradio Um Quarto de Século Uma Partida Uma Águia Sem Asas A Última Receita
Catumbi	História de uma Lágrima Curiosidade	Último Capítulo
Gamboa	Curiosidade Canseiras em Vão	O Cônego ou Metafísica do Estilo Memórias Póstumas de Brás Cubas Noite de Almirante O Programa Quincas Borba Um Erradio Anedota Pecuniária Conto de Escola Dívida Extinta Silvestre
Largo do Paço	Quem Boa Cama Faz	Quincas Borba Vidros Quebrados Uns Braços
Largo de São Francisco	Canseiras em Vão	A Carteira Dona Jocunda Esaú e Jacó João Fernandes Maria Cora Mariana (1871) Memorial de Aires Memórias Póstumas de Brás Cubas A Mão e a Luva O Programa Quincas Borba A Senhora do Galvão Um Capitão de Voluntários Um Quarto de Século Vinte Anos! Vinte Anos! Entre Santos
Praça do Comércio	Quinhentos Contos	Orai por Ele O Anel de Polícrates Casa Velha Esaú e Jacó A Inglesinha de Barcelos Maria Cora Memorial de Aires Quincas Borba Sales O Segredo de Augusta Tempo de Crise Um Almoço
Praia dos Mineiros	Diana	Tempo de Crise
Praia Vermelha	A Felicidade A Vida Eterna	Quincas Borba O Segredo de Augusta
Rua da Ajuda	Canseiras em Vão	Antes Que Cases... O Astrólogo A Melhor das Noivas Pai Contra Mãe Pobre Finoca! Quem Não Quer Ser Lobo... Quincas Borba A Senhora do Galvão Vinte Anos! Vinte Anos!
Rua da Assembleia	A Felicidade Canseiras em Vão	Antes Que Cases... A Carteira Casa Velha Esaú e Jacó Mariana (1896)
Rua/Praça da Constituição	Quem Boa Cama Faz	A Causa Secreta Decadência de Dois Grandes Homens Não É O Mel Para A Boca do Asno Quincas Borba O Sainete
Rua da Vala	Quem Boa Cama Faz	O Caso da Vara
Rua de D. Manoel	Curiosidade	A Causa Secreta Esaú e Jacó A Mulher Pálida
Rua Direita	Felicidade pelo Casamento Canseiras em Vão	Antes Que Cases... O Astrólogo João Fernandes Memorial de Aires Quincas Borba Ressurreição
Rua do Conde	Canseiras em Vão	Casa Não Casa... O Caso Barreto A Causa Secreta Dona Mônica Ernesto de Tal Miss Dollar A Mulher de Preto Último Capítulo

Lugar	Obra do PI	Obra do OC
Rua do Ouvidor	Diana Felicidade pelo Casamento Canseiras em Vão	O Anel de Polícrates O Anjo Rafael O Caminho de Damasco O Carro Número 13 Casa Velha O Caso da Viúva O Caso do Romualdo Dom Casmurro Esaú e Jacó O Espelho Helena Iaiá Garcia O Lapso Maria Cora Mariana (1871) O Melhor Remédio Memorial de Aires Miss Dollar A Mão e a Luva Nem Uma Nem Outra Papéis Velhos O Programa Pilades e Orestes Quincas Borba Ressurreição O Sainete Jogo do Bicho
Rua dos Ciganos	Onda Quem Boa Cama Faz	Confissões de Uma Viúva Moça Conversão de Um Ávaro A Mulher de Preto A Mão e a Luva Um Erradio
Tijuca	Felicidade pelo Casamento	Dona Benedita Dona Paula Memórias Póstumas de Brás Cubas Esaú e Jacó Iaiá Garcia O Anjo Rafael Antes Que Cases... As Bodas de Luís Duarte Casada e Viúva O Caso da Vara Decadência de Dois Grandes Homens Dom Casmurro Dívida Extinta O Empréstimo Evolução Helena A Inglesinha de Barcelos Jogo do Bicho Maria Cora Memorial de Aires Miloca Qual Dos Dois? Quem Não Quer Ser Lobo... Quincas Borba Ressurreição Sales Só! Trio em Lá Menor Troca de Datas Um Para o Outro Uma Partida Uma Águia Sem Asas Vidros Quebrados Vênus Divina Vênus!
Café Carceller	Canseiras em Vão A Felicidade	Antes Que Cases... Decadência de Dois Grandes Homens Esaú e Jacó Iaiá Garcia Mariana (1871) Memorial de Aires Miss Dollar A Mão e a Luva Onze Anos Depois Quincas Borba O Rei dos Caiporas Ressurreição To Be Or Not To Be Um Homem Superior Uma Por Outra Valério
Livraria Garnier	Canseiras em Vão A Vida Eterna	Valério O Caminho de Damasco Qual Dos Dois? Quincas Borba Tempo de Crise
Hotel da Europa	Onda	Almas Agradecidas A Mulher de Preto Nem Uma Nem Outra Tempo de Crise
Teatro Alcazar Lyrique	Canseiras em Vão	As Bodas de Luís Duarte O Caminho de Damasco Esaú e Jacó Luís Soares Miss Dollar Qual Dos Dois? O Segredo de Augusta
Teatro Lírico	Onda	Orai por Ele Uma Por Outra Astúcias de Marido O Carro Número 13 O Caso Barreto Cinco Mulheres Confissões de Uma Viúva Moça Dona Paula Esaú e Jacó Memorial de Aires A Mulher de Preto A Mão e a Luva Pobre Finoca! Questões de Marido Quincas Borba
Teatro de São Januário	Curiosidade	A Causa Secreta

Reflexões preliminares sobre a anotação das relações de coerência em um corpus de relatos sobre saúde mental do português brasileiro

Graziele Thainá Maciel Lima[1], Roana Rodrigues[1], Paula Christina Figueira Cardoso[2], Ewerson Dantas[1]

Universidade Federal de Sergipe (UFS), Universidade Federal do Pará (UFPA)

grazi.lima.15@gmail.com, roana@academico.ufs.br, pcardoso@ufpa.br, ewersonndantad@gmail.com

Abstract. *This study presents the preparation of a sample from the Amive corpus for discourse annotation based on Rhetorical Structure Theory (RST), focusing on the identification of coherence relations and their discourse markers in personal accounts of mental health experiences published on social media. To this end, a survey of the rhetorical relations and discourse markers described in the literature was conducted, with the aim of adapting the annotation protocol to the genre under analysis. The initial annotations reveal differences from the CSTNews journalistic corpus, previously annotated according to RST, particularly regarding the use of linguistic resources characteristic of digital communication. Thus, this research highlights the need to expand the typology of relations and discourse markers to accommodate this context.*

Resumo. *Este trabalho apresenta a preparação de uma amostra do corpus Amive para anotação discursiva baseada na Rhetorical Structure Theory (RST), com foco na identificação de relações de coerência e de seus sinalizadores discursivos em relatos pessoais sobre saúde mental publicados em redes sociais. Diante disso, foi realizado um levantamento das relações retóricas e dos sinalizadores discursivos descritos na literatura, com o objetivo de adaptar o protocolo de anotação ao gênero analisado. As primeiras anotações evidenciam diferenças em relação ao corpus jornalístico CSTNews, previamente anotado em RST, sobretudo no que diz respeito ao uso de recursos linguísticos característicos da comunicação digital. Assim, a pesquisa evidencia a necessidade de ampliação da tipologia de relações e sinalizadores para esse contexto.*

1. Introdução

Esta pesquisa insere-se na intersecção entre os projetos "Investigação de sinalizadores RST [*Rhetorical Structure Theory*] para o processamento do Português em um *corpus* multigênero de User Generated Content" e "AI-Based Support for Mental Health Communication (AIM-Health)". No âmbito do AIM-Health (aim-health.ufscar.br), o *corpus* Amive constitui a base de dados utilizada para a investigação de relatos relacionados à saúde mental publicados por usuários em rede social. O presente estudo busca, portanto, anotar as relações de coerência e seus sinalizadores discursivos em uma amostra do *corpus* Amive.

Trata-se de um desdobramento de estudos anteriores realizados com o *corpus* CSTNews [Cardoso *et al*., 2011], composto por textos jornalísticos. Acredita-se que o

trabalho com um gênero textual diferente, o qual é marcado pela espontaneidade, informalidade e proximidade com práticas comunicativas orais, apresentará dados linguísticos interessantes sobre o comportamento das relações de coerência textual para o português brasileiro (PB).

Nesse contexto, este trabalho concentra-se na preparação para a anotação discursiva de uma amostra do *corpus* Amive (amive.ufscar.br/publicações). Partindo dessas etapas preliminares, o trabalho objetiva apresentar os procedimentos adotados na preparação e primeiras impressões do *corpus* para anotação. Espera-se, assim, contribuir tanto para o avanço das pesquisas em processamento automático do PB quanto para a compreensão das estratégias discursivas utilizadas em contextos de comunicação sobre saúde mental em ambiente digital.

2. Noções gerais sobre a RST

A Rhetorical Structure Theory (RST) é uma teoria de base funcionalista, em nível discursivo, que visa a descrição das relações de coerência (também denominada relações retóricas) de um texto (Mann; Thompson, 1987). Embora os primeiros estudos da área tenham sido publicados na década de 1980, trata-se de uma teoria ainda muito utilizada [Souza; Cardoso; Rodrigues, 2023; Zeldes *et al*., 2024], principalmente no Processamento de Linguagem Natural (PLN), devido ao seu formalismo. Recentemente, publicou-se um estudo importante com novas diretrizes para a teoria, denominada e-RST (Enhanced Rhetorical Structure Theory), focando nos sinalizadores discursivos que contribuem para a interpretação das relações de coerência de um texto e nas relações de fenômenos discursivos mais complexos.

Na RST, a unidade mínima de análise é a EDU (Elementary Discourse Units), que se constitui, principalmente, por um núcleo (N), que contém a informação central que explicita a intenção do falante/escrito e por um satélite (S), entendido como a porção de texto que possui um dado adicional ao núcleo.

No PB, seguindo a taxonomia de Pardo (2005), tem-se 32 relações retóricas estabelecidas entre as EDUs, de tipo núcleo-satélite (mononuclear) ou núcleo-núcleo (multinuclear): Antithesis, Attribution, Background, Circumstance, Comparison, Concession, Conclusion, Condition, Contrast, Elaboration, Enablement, Evaluation, Evidence, Explanation, Interpretation, Joint, Justify, List, Means, Motivation, Non-Volitional Cause, Non Volitional Result, Otherwise, Parenthetical, Purpose, Restatement, Same-Unit, Sequence, Solutionhood, Summary, Volitional Cause e Volitional Result. Investiga-se, portanto, as restrições sobre N, sobre S, sobre a relação N-N ou N-S e o efeito gerado para ouvinte/leitor. Na Figura 1, exemplificamos uma EDU analisada, retirada do *corpus* CSTNews [Dantas *et al*., 2024].



Figura 1. Relação de coerência retirada de Dantas *et al.* (2024, p.15)

Acredita-se que este tipo de análise possibilita compreender os efeitos de coerência que constituem um texto e, em consonância com a agenda da área - proposta pela e-RST - parece ser possível aprofundar o estudo na medida em que se descreve não apenas a relação, mas os sinalizadores discursivos, entendidos, conforme Das e Taboada (2018), como recursos linguísticos e discursivos que orientam a interpretação das relações de sentido entre segmentos do texto e favorecem a construção da coerência textual. Os SDs podem ser do tipo marcadores discursivos (exemplo do "mas" na figura acima, que atua como MD do tipo conjunção), gênero, gráfico, lexical, morfológico, semântico, sintático etc.

2. A constituição do corpus e as etiquetas de anotação

O *corpus* Amive é composto por 604 textos do gênero relato pessoal, em meio digital, ou seja, trata-se de postagens textuais publicadas em rede social por usuários que abordam temas relacionados à saúde mental. Desse total, 336 postagens foram classificadas como PPD (perfil possível depressivo) e 268 como não PPD (perfil não depressivo). Esses textos contemplam discussões sobre sofrimento psíquico, transtornos mentais, experiências pessoais, busca por apoio e outras formas de expressão emocional em ambientes digitais.

Nesse sentido, para a anotação discursiva, foi realizado inicialmente um levantamento bibliográfico com o objetivo de identificar as relações retóricas e os sinalizadores discursivos descritos em estudos anteriores da RST. Esse percurso permitiu sistematizar as relações e os sinalizadores já documentados na literatura, para além do PB, assim como forneceu subsídios para a definição do conjunto de relações retóricas e dos sinalizadores discursivos que serão adotados no processo da nossa anotação do *corpus*.

Dessa atividade, foram consideradas as 32 relações retóricas previstas por Pardo (2005) para o PB, além das contribuições dos trabalhos de Mann e Thompson (1987) e Zeldes *et al*. (2024) para o inglês. Não cabe, neste trabalho, discutirmos os detalhes de cada relação, no entanto, parece-nos interessante ressaltar que: (i) algumas novas relações parecem ser produtivas em gêneros textuais distintos, para além do gênero acadêmico, como o trabalhado por Pardo (2005) ou jornalístico, que caracteriza o CSTNews, *corpus* previamente em RST para o PB [Dantas *et al*., 2024]. A título de exemplo, mencionamos Org-Preparation (relação que marca a organização gráfica do satélite para preparar o núcleo) e Org-Phatic (relação que marca a organização da interação do texto, como a manutenção, abertura ou encerramento do canal comunicativo); e (ii) o nível de

granularidade das etiquetas das relações retóricas é distinto: enquanto Pardo (2005), por exemplo, duplica as relações de causa e resultado segundo a volitividade da relação (Volitional Cause, Non-Volitional Cause, Volutional Result e Non-Volitional Result), diferente do observado em Zeldes *et al.* (2024); por sua vez, Zeldes *et al.* (2024) propõem que a relação Elaboration seja expandida em: Elab-Additional (em que há a elaboração de uma informação adicional) e Elab-Attribute (em que o efeito gerado se refere à elaboração especificamente de atributos, propriedades e características).

No que se refere aos sinalizadores discursivos, consideramos as classes e subclasses propostas também em trabalhos anteriores do PB [Dantas *et al.,* 2024; Gama e Souza, 2024] e do inglês [Das e Taboada, 2018; Zeldes *et al.,* 2024]. Com base em Dantas *et al.* (2024), temos os seguintes sinalizadores que dão indícios para a interpretação da relação de coerência entre duas porções de texto: *marcador discursivo* - MD (como preposição, conjunção e advérbio); *sinal gráfico* (como aspas, pontuação, travessão, vírgula, dois pontos, reticências, interrogação e caractere especial); *morfológico* (como tempo verbal e modo verbal); *sintático* (oração circunstancial, oração relativa e valência verbal); e *semântico* (como hiperonímia-hiponímia, campo semântico, sinonímia, sentido verbal, conhecimento de mundo).

Diante do exposto, estudamos as relações e sinalizadores discursivos não só comumente considerados para o PB, mas também para o inglês, tendo acrescentado à nossa tipologia outras classes e subclasses que podem ser relevantes na anotação do nosso *corpus* de relatos pessoais em rede social.

3. Primeiras impressões da anotação

Após a análise das etiquetas, foram selecionados 15 textos do corpus para a realização de um ensaio de anotação envolvendo quatro anotadores, sendo três experientes na anotação em RST e um atuando nessa função pela primeira vez. Após a anotação dos textos, a equipe deliberou uma taxonomia com 32 relações RST, sendo elas recorrentes nos autores investigados e/ou terem sido anotadas em nossa amostra. O mesmo procedimento foi realizado com relação aos sinalizadores discursivos: ao todo, consideramos 25 sinalizadores, distribuídos em 5 classes: *MD, sinal gráfico, morfológico, sintático, semânticos* e *referencial*.

A anotação ocorreu por meio da ferramenta rstWeb [Gessler; Liu; Zeldes, 2019]. Com uma interface amigável, é possível fazer a *segmentação* do texto, ou seja, organizá-lo em EDUs; anotar as *relações de coerência* entre as partes do texto; e ainda apontar os *sinalizadores* que colaboram com a interpretação das relações anotadas. A Figura 2 ilustra a anotação a partir de uma paráfrase baseada nos dados do corpus, já que os dados originais são de temática sensível e o projeto AIM-Health não permite a disponibilização na íntegra dos textos.

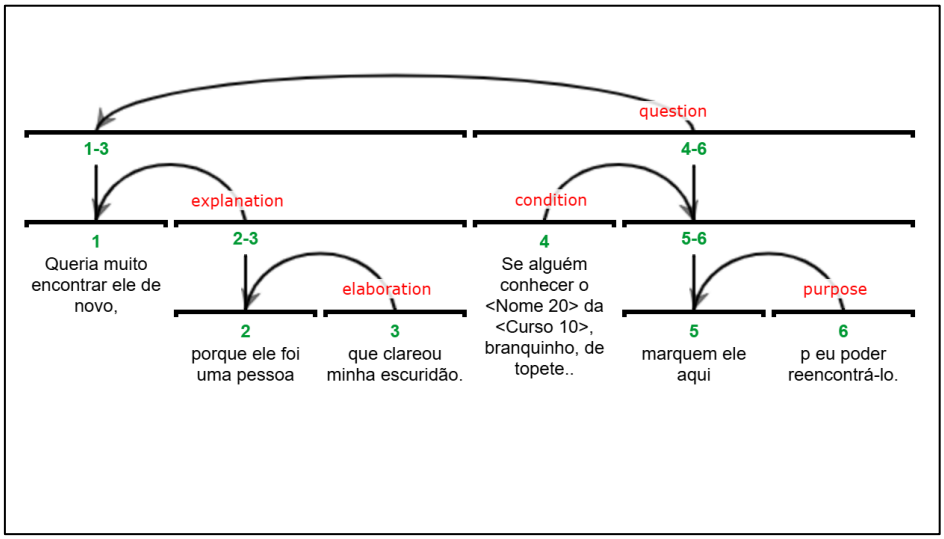


Figura 2. Exemplo de anotação realizada na ferramenta rstWeb

Diante disso, as análises iniciais já permitem observar algumas características linguísticas que distinguem o *corpus* Amive do *corpus* CSTNews previamente anotado. Enquanto este último é constituído por textos jornalísticos, marcados por maior formalidade e proximidade com a norma-padrão da língua, o primeiro apresenta traços típicos da comunicação em redes sociais, caracterizando-se por uma linguagem mais informal e espontânea. Assim, entre as ocorrências observadas, destacam-se o uso de fenômenos linguísticos característicos da comunicação presente no ambiente digital, como abreviações (como o uso do "p" para expressar a preposição para; "tá" para o verbo está), afastamentos da norma-padrão da língua (em "minha vida esta sendo desperdiçada") e usos frequentes de palavras informais, gírias (como em "cara" e "galera") etc.

Além dos aspectos apontados, destacamos: (i) o tamanho dos textos é, de maneira geral, menor que as notícias que compõem o CSTNews; e (ii) o aparecimento de relações que não eram consideradas, como Question (como pode ser visualizada na figura 2), além de casos que raramente apareciam no CSTNews, como Background (N- "Fui despedida, | S- iniciei acompanhamento com o especialista da saúde mental lá no <instituição 30>"), que oferece conhecimento de fundo que facilita a compreensão do núcleo. Essa mudança está certamente atrelada ao novo gênero trabalhado, já que o relato pessoal é uma narrativa e o meio (rede social) contribui para a interação entre o escritor e o leitor da publicação.

Portanto, as primeiras observações realizadas durante a fase de preparação do corpus já evidenciam diferenças significativas em relação aos textos jornalísticos anteriormente analisados. Nesse sentido, essas características demonstram um cenário desafiador para a anotação discursiva, haja vista que os recursos linguísticos utilizados para expressar relações de coerência podem manifestar-se de modo diferente, se comparado a gêneros mais formais.

4. Considerações Finais

As etapas desta pesquisa desenvolvidas até o momento evidenciam que a preparação do

corpus Amive para anotação discursiva requer adaptações metodológicas, as quais são frutos das especificidades do gênero *relato pessoal* analisado, mais especificamente no meio digital. Nesse sentido, o levantamento bibliográfico realizado permitiu sistematizar um conjunto de relações retóricas e categorias de sinalizadores discursivos descritos na literatura, fornecendo uma base teórica consistente para a elaboração do protocolo de anotação que será empregado no trabalho.

Além disso, as primeiras anotações dos textos do corpus possibilitaram iniciar o processo de segmentação em EDUs, a identificação das relações retóricas e o levantamento dos sinalizadores discursivos presentes. Embora essa etapa ainda esteja em andamento, as observações iniciais já apontam diferenças relevantes entre o *corpus* Amive e o CSTNews, anteriormente utilizado em estudos do grupo de pesquisa. Dessa forma, as distinções sugerem que as relações de coerência podem ocorrer por meio de recursos linguísticos como abreviações, gírias, usos não convencionais da norma padrão, fatos esses que são distintos daqueles observados em textos jornalísticos. Isso reforça, portanto, a necessidade de investigar como os sinalizadores discursivos se manifestam nesse contexto específico.

Questões éticas

Este trabalho utiliza um exemplo adaptado do *corpus* Amive. Conforme relatado no texto, além de anônimos, os dados do Projeto AIM-Health são parafraseados devido ao caráter sensível de seu conteúdo. Informamos que nenhuma tecnologia de IA generativa ou assistida por IA foi utilizada na preparação deste trabalho. Além disso, cada autor participou ativamente no desenvolvimento deste trabalho: na anotação da amostra do *corpus*, na escrita e na revisão deste artigo.

Agradecimentos

Este trabalho integra o projeto *AI-Based Support for Mental Health Communication*, financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil (Processo nº 24/10233-7). Agradecemos também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil, pelo apoio financeiro concedido por meio de bolsa de doutorado.

Referências

Das, D.; Taboada, M. (2018) RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, v. 52, n. 1, p. 149-184.

Dantas, E. Cruz, G. S. B.; Santa Barbara, L. J.; Gama, N. S.; Pereira, M. A.; Almeida, T. J. A.; Souza, J. W. C.; Rodrigues, R.; Cardoso, P. C. F. (2024) *Manual de anotação de sinalizadores discursivos*. Relatório Técnico do NILC.

Gama, N. S.; SOUZA, J. W. C. (2024) Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários. In: *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology.*

Gessler, L., Liu, Y., & Zeldes, A. (2019). A discourse signal annotation system for RST

trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019* (pp. 56–61). Association for Computational Linguistics.

Mann, w. C.; Thompson, S. A. (1987) *Rhetorical structure theory: a theory of text organization.* Los Angeles: University of Southern California, Information Sciences Institute. (Technical Report ISI/RS-87-190).

Pardo, T. A. S. (2005) *Métodos para análise discursiva automática*. Tese (Doutorado em Ciência da Computação) - Universidade de São Paulo, São Carlos.

Rodrigues, R.; Souza, J. W. C.; Cardoso, P. C. F. (2023) Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTNews. In: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2023, Brasil. *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*.

Zeldes, A. *et al.* (2024) eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, v. 51, n 1.

Capacitismo e Ética em Processamento de Linguagem Natural: Reflexões Semânticas do Léxico Especializado na FrameNet Brasil

Iasmin Valéria Miranda RABELO[1], Maucha Andrade GAMONAL[2]

[1]Universidade Federal de Minas Gerais (UFMG) – Minhas Gerais – Brasil,
[2]Universidade Estadual de Campinas – São Paulo – Brasil

iasminvaleria@gmail.com, mauchaandrade@gmail.com

Resumo. *A Semântica de Frames [Fillmore, 1982] é uma teoria linguístico-cognitiva que investiga a concepção de significados considerando contextos socioculturais e a experiência humana na construção de sentido. Fundamentado nessa teoria, o projeto FrameNet Brasil [Torrent et al., 2022] desenvolve tecnologias linguísticas voltadas para tarefas de Processamento de Linguagem Natural. É nesse contexto que surge a iniciativa de expandir sua base de dados do português para o léxico especializado de acessibilidade. Este trabalho inicia uma discussão acerca da inclusão do léxico capacitista na FrameNet Brasil, fundamentando os conceitos iniciais necessários para a modelagem de frames e um conjunto lexical inicial para anotação semântica. Esta pesquisa evidencia a necessidade de incluir esse léxico na FrameNet Brasil a partir de discussões acerca de ética e vieses identificados na linguagem em tarefas de Processamento de Linguagem Natural e LLMs. Este trabalho ressalta a importância da linguagem inclusiva e a necessidade de desenvolver tecnologias linguísticas conscientes do capacitismo presente na língua portuguesa, contribuindo para a abordagem de questões éticas na área.*

Palavras-chave: Semântica de Frames; FrameNet Brasil; Acessibilidade.

Abstract. *Frame Semantics [Fillmore, 1982] is a cognitive-linguistic theory that investigates the conception of meaning by considering sociocultural contexts and human experience in the construction of meaning. Based on this theory, the FrameNet Brasil project [Torrent et al., 2022] develops linguistic technologies geared toward Natural Language Processing tasks. It is in this context that the initiative to expand its Portuguese database to include specialized accessibility terminology has emerged. This study initiates a discussion regarding the inclusion of ableist vocabulary in FrameNet Brasil, establishing the foundational concepts necessary for frame modeling and an initial lexical set for semantic annotation. This research highlights the need to include this lexicon in FrameNet Brasil based on discussions regarding ethics and biases identified in language used in Natural Language Processing tasks and Large Language Models (LLMs). This study emphasizes the importance of inclusive language and the need to develop linguistic technologies that are mindful of the ableism present in the Portuguese language, thereby contributing to addressing*

ethical issues in the field.

Keywords: Frame Semantics; FrameNet Brasil; Accessibility.

1. Semântica de *Frames* e FrameNet Brasil

A Semântica de *Frames* é uma teoria linguístico-cognitiva desenvolvida na década de 1980 por Charles Fillmore com o objetivo de compreender a construção de significado. Para isso, faz-se uso do conceito de *frames*: redes de significados relacionados e ancorados nos contextos socioculturais dos falantes.

A FrameNet é uma rede semântico-lexical baseada nos conceitos da Semântica de *Frames*, desenvolvida por Fillmore em 1997, no International Computer Science Institute (ICSI), em Berkeley, Califórnia. O projeto de lexicografia computacional registra as possibilidades sintáticas e semânticas da língua inglesa a partir dos *frames* [Fillmore et al., 2003]. A FrameNet Brasil[47] [Torrent et al., 2022], sediada na Universidade Federal de Juiz de Fora, Minas Gerais, é um laboratório de Linguística Computacional que aplica os conceitos da Semântica de *Frames* e as metodologias da FrameNet Berkeley às anotações semânticas do português brasileiro. Nos últimos anos, a FrameNet Brasil vem se expandindo para além da quantidade de ULs e *frames*, abordando diferentes aplicações da sua base de dados no desenvolvimento de tecnologias linguísticas e tarefas de Processamento de Linguagem Natural com foco na responsabilidade social [Gamonal et al., 2026], como será explorado neste trabalho.

2. Viés e preconceito em tarefas de Processamento de Linguagem Natural: a importância da linguagem inclusiva em tecnologias linguísticas

Ao analisar os *frames* semânticos propostos por Fillmore (1982), é possível entender os *frames* como estruturas cognitivas que organizam os contextos dos falantes e estabelecem relações de sentido através do léxico. Nesse contexto, Silva (2009) evidencia que contextos históricos e socioeconômicos influenciam a visão de mundo dentro de uma sociedade. Similarmente, Leal et. al (2024) apontam que a comunicação não acontece de forma isolada. A linguagem reflete crenças sociais e valores que influenciam o comportamento e perspectiva de uma comunidade de fala, sendo "um meio resultante da desigualdade produzida através das influências referidas e um eixo estrutural capaz de operar, ele mesmo, processos que podem resultar em diferença e desigualdade" [Leal et al., 2024].

Partindo das evidências que apontam a linguagem como recurso essencial no desenvolvimento de uma sociedade inclusiva e respeitosa, é necessário trazer essa questão para as tecnologias linguísticas. Caliskan et al. (2017) abordam as preocupações de que as Inteligências Artificiais e *Large Language Models* (LLMs) sejam usados, de forma intencional ou não, para perpetuar preconceitos intrínsecos na cultura humana. A semântica naturalmente reflete os vieses sociais e, consequentemente, ferramentas de PLN expostas à linguagem humana também serão enviesadas — o problema acontece quando esse viés carrega preconceitos que prejudicam grupos sociais marginalizados [Caliskan et al., 2017]. Através da linguagem, os grupos são categorizados e muitas vezes

[47] A FrameNet Brasil está disponível em: < https://webtool.frame.net.br/>. Acesso em 28 de maio de 2026.

estereotipados, reforçando desigualdades sociais transmitidas e reforçadas pela forma em que uma comunidade de fala se comunica [Beukeboom e Burgers, 2019]. Dessa forma, os estereótipos reforçados em tarefas de PLN causam um dano social ao representar diferentes grupos de forma negativa [Barokas et al., 2020].

Na área da Acessibilidade, o projeto AUTALIC [Rizvi et al., 2025] apresenta um dataset anotado de linguagem capacitista contra pessoas com autismo em sentenças em inglês extraídas do Reddit. A pesquisa ressalta o caráter subjetivo e sutil da linguagem capacitista, evidenciando a dificuldade das LLMs em identificar discursos de ódio ou preconceitos contra pessoas com deficiência. Similarmente, [Li et al. 2024] apontam a "natureza interseccional dos preconceitos" (p. 239): o capacitismo em LLMs também está conectado a outras identidades sociais, criando uma interseção de preconceitos interligados e mais difíceis de identificar. Consequentemente, é possível identificar valores negativos associados às deficiências em modelos de PLN, reforçando a necessidade de abordar a linguagem capacitista em tarefas de linguística computacional [Hassan et al. 2021].

3. Aplicações da FrameNet Brasil: recursos linguísticos com foco na responsabilidade social

A relação entre contextos sociais e a construção de sentido na Semântica de *Frames* define a conexão entre o significado e a forma como interpretamos o mundo, interligando os *frames* — redes de conceitos relacionados — e repertório sociocultural [Fillmore, 1982]. Nesse sentido, Gamonal et al. (2026) exploram as aplicações e contribuições da FrameNet Brasil para tarefas de PLN no que diz respeito a causas sociais. As metodologias de anotação da FrameNet Brasil, aliadas aos fundamentos extralinguísticos da construção de sentido da Semântica de *Frames*, permitem o desenvolvimento de sistemas computacionais mais capazes de identificar e interpretar diferentes contextos de uso da língua [Gamonal et al., 2026].

O projeto DataToStopGBV[48], baseado na pesquisa de Dutra (2024), utiliza a estrutura semântica da FrameNet Brasil para identificar padrões linguísticos indicativos de violência baseada em gênero através da análise de prontuários. Fazendo uso da anotação computacional para estudar casos notificados e identificar casos subnotificados, o projeto possibilita alertas precoces e acionamento das autoridades locais. Já o Audition[49] [Gamonal et al., 2025] constitui um *dataset* multimodal composto por curtas-metragens brasileiros. A anotação semântica da linguagem verbal e não verbal evidencia as vantagens de aplicação das metodologias da FrameNet Brasil com foco em acessibilidade, colocando o Audition como uma ferramenta valiosa na elaboração de roteiros de audiodescrição e na promoção da acessibilidade cultural [Gamonal et al. 2026].

3.1. O domínio da Acessibilidade na FrameNet Brasil

O trabalho de Rabelo et al. (2025) — focado na modelagem dos frames

[48] O projeto está disponível em <https://datatostopgbv.org/pt-br/>. Acesso em 28 de maio de 2026.

[49] A primeira versão do *dataset* está disponível em <https://huggingface.co/datasets/FrameNetBrasil/Audition>. Acesso em 28 de maio de 2026.

`Deficiência`[50], `Tecnologia_Assistiva` e `Pessoas_com_deficiência`, a partir da linguagem apropriada para se referir às deficiências identificada no *corpus* compilado — abre portas para a atual pesquisa, em desenvolvimento na FrameNet Brasil, que busca incluir, analisar e anotar o léxico capacitista na base de dados do projeto. Com base nas discussões apresentadas, é possível afirmar que a linguagem é uma ferramenta importante na manutenção de uma sociedade respeitosa e inclusiva. Além disso, o capacitismo presente na língua portuguesa, muitas vezes de forma subjetiva e sutil, também deve ser considerado no desenvolvimento de tecnologias linguísticas para evitar a propagação de preconceitos. A partir disso e das pesquisas já realizadas no projeto, a FrameNet Brasil apresenta aplicações relevantes para análise e anotação do caráter subjetivo da linguagem capacitista.

Atualmente, a pesquisa se dedica à definição dos conceitos principais sobre capacitismo e composição de *corpus* — baseando-se novamente em fontes especializadas, como órgãos governamentais e da área da saúde — para fundamentar a descrição e os Elementos de *Frame* do *frame* preliminar de `Capacitismo`. O *corpus* inicial é composto por cinco cartilhas elaboradas por instituições governamentais ou da área da saúde. Os documentos apresentam expressões e termos capacitistas que devem ser evitados por perpetuar preconceitos, discriminação e exclusão de pessoas com deficiência.

A análise preliminar do *corpus* permite a identificação de termos recorrentes apontados como inapropriados, como "retardado", "aleijado", "portador de deficiência", "inválido" e "necessidades especiais". Esses itens lexicais foram selecionados como ULs em potencial a serem anotadas, devido à frequência com que aparecem no *corpus* como termos inapropriados ou preconceituosos para se referir às pessoas com deficiência. É interessante apontar que algumas das expressões atualmente classificadas como capacitistas já foram consideradas corretas para se referir às deficiências. Por exemplo, o Glossário de Direitos Humanos da Polícia Federal (2025) aponta que, nas décadas de 1970 a 1980, os termos "pessoa portadora de deficiência" e "pessoa com necessidades especiais" surgiram como tentativas de tornar a língua mais respeitosa e abrangente, porém posteriormente foram revisados e caíram em desuso. Essas variações linguísticas e evoluções na forma de se falar das deficiência devem ser consideradas na análise do léxico, anotação e inclusão na base de dados da FrameNet Brasil. O conceito de *reframing* apresentado por Fillmore (1982) define esse "reenquadramento": ULs que evocam *frames* diferentes dentro de um contexto motivador que causa essas mudanças. Dessa forma, o *reframing* pode ser utilizado na anotação desses itens lexicais que foram revistos e reenquadrados como termos inapropriados. As sentenças abaixo, extraídas Sketch Engine[51], exemplificam diferentes contextos de uso da palavra "retardado":

(1) Esses obstáculos, de certa forma, têm **retardado** o desenvolvimento das estratégias aprovadas.

(2) Mudanças nos padrões de sono, mudanças no peso e no apetite, maior estresse percebido, metabolismo **retardado**.

(3) Você deve escrever um tutorial pensando no seu amigo mais **retardado** e fazer

[50] Por convenção, os *frames* são destacados na fonte Courier New.

[51] O *corpus* "Portuguese Web 2023" está disponível em <app.sketchengine.eu>. Acesso em 17 de junho de 2026.

uma maneira de ele entender.

(4) Esse reporter é **retardado**? Ultimamente os repórteres estão cada vez piores.

As sentenças (1) e (2) usam a palavra "retardado" como adjetivo para indicar algo adiado, atrasado ou demorado, sem atribuição de conotações negativas. Em contrapartida, os exemplos (3) e (4) apresentam "retardado" como sinônimo de falta de inteligência ou incapacidade, usando a palavra para criticar e descrever negativamente uma pessoa. Por isso, os contextos de uso devem ser considerados na anotação do léxico capacitista, avaliando quais *frames* uma mesma UL pode evocar na mente dos falantes.

O Guia de Linguagem Inclusiva em Direitos Humanos da Defensoria Pública da União ressalta a importância do contexto para determinar se uma palavra é ou não capacitista, usando como exemplo o termo "autista". Embora seja uma palavra frequentemente classificada como capacitista por definir a pessoa por sua condição, muitas pessoas da comunidade não veem o termo como pejorativo e apresentam preferência por essa nomenclatura. Como explicado no Guia, "quem faz essa escolha não vê o termo como condição patológica ou característica negativa, e sim como parte fundamental de sua identidade". Esse caráter subjetivo da linguagem capacitista é um fator relevante na modelagem do *frame*, análise de *corpus* e anotação lexicográfica. Nos exemplos abaixo, extraídos do Sketch Engine (sentenças 5 e 6) e do Corpus do Português[52] (sentenças 7 e 8), a mesma palavra possui conotações diferentes:

(5) Estranhamente, de uma hora pra outra, todo mundo tem tendência a ser psicopata, **autista**.

(6) Eu digo 'eu era **autista**', só que meu pai tirou o **autista** na peia. Naquele tempo tirava **autista** era na chibata.

(7) O Governo Regional dos Açores tem tido uma incompreensível atitude **autista** e de ausência de diálogo.

(8) Cada **autista** tem sua própria personalidade, seu jeito de ser.

As sentenças (5), (6) e (7) apresentam a palavra "autista" em contextos pejorativos, associando a condição à algo negativo, prejudicial ou que deve ser "corrigido" através de violência física. Já a sentença (8) não indica uma conotação negativa, se referindo às pessoas com autismo de forma respeitosa e considerando suas individualidades. Esses vieses identificados em *corpora* genéricos de diferentes ferramentas evidenciam a presença da linguagem capacitista no português. Consequentemente, ao usar *corpora* do português em tarefas de PLN — como treinamento de LLMs, Inteligências Artificiais Generativas, tradutores automáticos, entre outros —, os vieses capacitistas são replicados, causando a perpetuação de estigmas e preconceitos através da linguagem.

O objetivo, ao longo do desenvolvimento da pesquisa, é modelar o *frame* `Capacitismo` para tarefas de anotação e identificação de linguagem capacitista a partir dos conceitos da Semântica de *Frames* e metodologias da FrameNet Brasil. Futuramente, a finalidade desse trabalho é continuar aplicando a base de dados da FrameNet Brasil no desenvolvimento de tecnologias linguísticas inclusivas, conscientes e cada vez mais

[52] O Corpus do Português está disponível em <https://www.corpusdoportugues.org/>. Acesso em 07 de junho de 2026.

capazes de identificar preconceitos subjetivos que ainda passam despercebidos em muitas LLMs.

4. Conclusão

Essa pesquisa apresenta uma discussão preliminar dos conceitos acerca do capacitismo na língua portuguesa, fundamentando-se na Semântica de *Frames* para abordar os fatores extralinguísticos — como contextos sociais, históricos e comunidades de fala — que impactam a forma de se falar das deficiências, e traz foco para o viés capacitista muitas vezes presente em tarefas de PLN. Essas discussões constituem as bases necessárias para a modelagem de *frames* e expansão da base de dados da FrameNet Brasil para compreender o léxico que diz respeito ao capacitismo. Essa pesquisa abre portas para um desenvolvimento da base de dados da FrameNet Brasil de forma consciente de questões capacitistas, com possibilidades de aplicações em tecnologias linguísticas e seguindo os objetivos do projeto de empregar os princípios da Semântica de *Frames* e suas metodologias na manutenção de uma sociedade mais inclusiva.

Referências

Amaral, S. e Paulista, B. (2021). Mudanças terminológicas na referência feita à pessoa com deficiência: contribuições da Análise do Discurso. In *Revista Philologus*, v. 27, n. 81, p. 1702-1715.

Barokas, S., Blodgett, S., Daumé, H. e Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5454–5476.

Beukeboom, C. e Burgers, C. (2019). How Stereotypes Are Shared Through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Frame work. In *Review of Communication Research*, p. 1–37.

Caliskan, A., Bryson, J. e Narayanan, J. (2017). Semantics derived automatically from language corpora contain human-like biases. In *Science*, v. 356, n. 6334, p. 183-186.

Campbell, K. Refusing Able(ness): A Preliminary Conversation about Ableism. In *M/C Journal*, v. 11, n. 3.

Dutra, L. (2024). Evaluating the contribution of framenet to gender-based violence identification: how semantic annotation can be used as a resource for identifying patterns of violence. 87 p. Tese (Mestrado em Language Technology). Department of Philosophy, Linguistics and Theory of Science, University of Gotherburg, Gotemburgo, Suécia.

Fillmore, C. (1982). Frame Semantics. In *Linguistics in the Morning Calm*, v. 27, n. 81, p. 1702-1715.

Fillmore, C., Johnson, C. e Petruck, M. (2003). Background to FrameNet. In *International Journal of Lexicography*, v. 16, n. 3, p. 235–250.

Gama, A., Paravidini, C., Guimarães, D. e Amaral, S. (2021). O anticapacitismo e o poder das palavras: a linguagem como ideal de emancipação das pessoas com deficiência. In *Revista Philologus*, v. 27, n. 81, p. 1762-1772.

Gamonal, M. et al. (2025). Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content. In *Proceedings of the 21st Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, p. 95-104.

Gamonal, M. et al. (2026). PLN e Responsabilidade Social: Aplicações da FrameNet-Br. In Caseli, H.M.; Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 4 ed., p. 330-369.

Hassan, S., Huenerfauth, M. e Alm, C. (2021). Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. In *Findings of the Association for Computational Linguistics*, p. 3116–3123.

Leal, D., Freitas, J., Magalhães, S. e Matias, M. (2024). Quem inclusivo fala, sempre acerta: uma discussão sobre linguagem inclusiva. In *ex æquo*, n. 49, p. 19-34.

Li, R., Kamaraj, A., Ma, J. e Ebling, S. (2024). Decoding Ableism in Large Language Models: An Intersectional Approach. In *Decoding Ableism in Large Language Models: An Intersectional Approach*, p. 232–249.

Mazzeto, A., Santos, A. e Hora, N. (2025). Interesses coletivos e o combate ao ódio contra pessoas com deficiência. In *XXV Encontro Nacional de Pesquisa em Ciência da Informação*.

Rabelo, I., Gamonal, M., Pagano, A. (2025). Disability and assistive technology: modeling accessibility domain frames on FrameNet Brasil. In *Cadernos de Estudos Linguísticos*,, v. 67, n. 00, p. e025032, 2025.

Rizvi, N. et al. (2025). AUTALIC: A Dataset for Anti-AUTistic Ableist Language In Context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, v. 1, p. 20999-21015.

Salomão, M. (2003). Razão, realismo e verdade: o que nos ensina o estudo sociocognitivo da referência. In *Cadernos de estudos linguísticos*, v. 44, p. 71-84.

Silva, M. (2009). Da Exclusão à Inclusão: Concepções e Práticas. In *Revista Lusófona de Educação*, v. 13, p. 135-153.

Torrent, T. et al. (2022). Representing Context in FrameNet: A Multidimensional, Multimodal Approach. In *Frontiers in Psychology*, v. 13.

Travis, M. (2002). Perceived disabilities, social cognition, and innocent mistakes. In *Vand. L. Rev.*, v. 55, p. 481-577.

Vendramim, C. (2019). Repensando mitos contemporâneos: o capacitismo. In *Simpósio Internacional Repensando Mitos Contemporâneos*, v. 2, p. 16-25.

CoRpMIA: Proposta de um Corpus de Representações em Mídias Culturais da Inteligência Artificial

Cibele Santos[1,2], Eduardo de Moura[1,3], Livy Real[1,4]

[1]Instituto Kunumi, [2]PPGCC — Universidade Federal de Minas Gerais (UFMG), [3]Universidade Estadual de Campinas (UNICAMP), [4]Instituto de Computação — Universidade Federal do Amazonas (UFAM)

cibelesimoesoliveira@gmail.com, eduardo.moura@kunumi.com, livyreal@gmail.com

Abstract. *This paper presents the initial design of CoRpMIA, a machine-readable corpus of cultural media representations that mention, represent or thematize artificial intelligence. The proposal combines Corpus Linguistics and Digital Humanities to document metadata, media types, representational roles and ethical-cultural themes associated with AI. Rather than assuming a causal link between cultural representations and technological behavior, we situate the corpus within the study of sociotechnical imaginaries, where collective visions of the future participate in shaping public perceptions of technology. We describe a pilot annotation scheme and discuss multilingual scope, access to primary material and annotator disagreement as methodological issues.*

Resumo. *Este artigo apresenta o desenho inicial do CoRpMIA, um corpus legível por máquina de representações em mídias culturais que mencionam, representam ou tematizam a inteligência artificial. A proposta articula Linguística de Corpus e Humanidades Digitais para documentar metadados, tipos de mídia, papéis representacionais e temas ético-culturais associados à IA. Em vez de supor uma relação causal entre representações culturais e comportamento tecnológico, situamos o corpus no estudo dos imaginários sociotécnicos, em que visões coletivas de futuro participam da formação de percepções públicas sobre a tecnologia. Descrevemos um esquema piloto de anotação e discutimos escopo multilíngue, acesso ao material primário e desacordo entre anotadores como questões metodológicas.*

1. Introdução

Com o crescimento do uso da inteligência artificial (IA) em diferentes esferas da vida cotidiana, torna-se relevante investigar de que forma as representações culturais da IA participam da formação de percepções públicas sobre essa tecnologia. Não se trata de afirmar uma relação causal direta entre uma obra cultural específica e um comportamento tecnológico contemporâneo, nossa hipótese é melhor situada no plano dos imaginários sociotécnicos, em que visões coletivas de futuro participam da fabricação do poder tecnológico (Jasanoff and Kim 2015). Nesse plano, representações culturais e tecnologia partilham um campo comum de linguagem e expectativa: robôs, sistemas autônomos e inteligências não humanas já habitavam o imaginário coletivo muito antes dos grandes modelos de linguagem e dos chatbots. Essas representações não determinam os usos contemporâneos da IA, mas podem tornar certas interações mais reconhecíveis, familiares ou aceitáveis.

Esse enquadramento ganha relevância diante de comportamentos antropomorfizados e bajuladores em sistemas conversacionais. Sistemas de diálogo são frequentemente antropomorfizados por projetistas e personificados por usuários, o que estimula confiança excessiva em seus resultados (Abercrombie et al. 2023). Modelos ajustados por preferências humanas podem apresentar sicofância, favorecendo respostas que concordam com o usuário em vez de respostas factuais (Sharma et al. 2024). Quanto aos agentes conversacionais, um estudo recente com 270 adultos de 29 países identificou que usuários recorriam ao ChatGPT tanto para necessidades de saúde mental quanto para demandas psicossociais, como companhia e orientação relacional (Luo et al. 2025). O Brasil ainda aparece como o país 'mais viciado em IA' em pesquisa recente[53]. É como resposta a esse contexto e tentativa de compreender a sociedade e sua relação com a IA que se enquadra o presente trabalho.

Este artigo apresenta a proposta inicial do CoRpMIA, um corpus de obras culturais que mencionam, representam ou tematizam a inteligência artificial. A partir de métodos das Humanidades Digitais e da Linguística de Corpus, pretende-se catalogar essas obras de forma estruturada e legível por máquina, reunindo informações como mídia, ano, país, língua original, tipo de IA representada, papel narrativo, tom da representação, temas associados e observações contextuais. Nesta fase piloto, o objetivo não é oferecer uma lista exaustiva, mas definir critérios de inclusão, testar uma ficha de catalogação e elaborar uma taxonomia preliminar. As categorias iniciais (*ameaça, companheira, trabalhadora, oráculo, ferramenta, consciência emergente, espelho do humano*) são derivadas exploratoriamente da literatura sobre narrativas culturais de IA (Cave, Dihal, and Dillon 2020; Hermann 2023) e assumidas como provisórias, a serem refinadas durante a anotação.

2. Trabalhos relacionados

A construção do CoRpMIA apoia-se na tradição da Linguística de Corpus e na discussão sobre datasets e corpora anotados para PLN. Em português brasileiro, Berber Sardinha (Berber Sardinha 2004) fundamenta a Linguística de Corpus como abordagem empírica apoiada em coleções organizadas de dados, e o trabalho "Dataset e corpus" (Freitas 2023) sistematiza procedimentos de criação de corpora anotados. Essas bases são centrais porque o projeto não visa apenas uma lista de obras, mas um conjunto estruturado de dados culturais, com critérios de inclusão, metadados, categorias de anotação e documentação.

O projeto dialoga também com as Humanidades Digitais, que mostram que dados culturais não são neutros: sua organização depende de escolhas de recorte, descrição e classificação (Burdick et al. 2012). Como discutem Zandavalle e Real (Zandavalle and Real 2023) sobre transdisciplinaridade em tecnologias da linguagem, um corpus desse tipo exige colaboração entre computação, linguística, estudos culturais, literatura, artes e debates sociais sobre IA.

O interlocutor mais direto do CoRpMIA é o campo que estuda narrativas e imaginários culturais da IA. Cave, Dihal e Dillon (Cave, Dihal, and Dillon 2020) mostram

[53]https://www.biia.com/checkpoint-the-countries-using-ai-the-most-might-surprise-you-new-study/

que a IA é organizada na imaginação pública por um repertório de narrativas e mitos que estruturam medos, expectativas e disputas sociais. Na mesma direção, Hermann (Hermann 2023) lê a IA na ficção não só como antecipação, mas como conjunto de metáforas, e Aktaş (Aktaş 2026) analisa dez anos de representações da IA no cinema a partir do IMDb. O CoRpMIA não se diferencia apenas por ampliar as mídias: sua contribuição própria está em duas decisões metodológicas, tratar o desacordo entre anotadores como evidência interpretativa, e não como ruído, e acompanhar o corpus de documentação responsável que explicita suas escolhas de recorte.

Por fim, o CoRpMIA dialoga com documentação responsável de dados e abordagens perspectivistas em PLN. *Datasheets for Datasets* (Gebru et al. 2021) defende acompanhar datasets de informações sobre motivação, composição, coleta, limitações e usos previstos. Basile (Basile 2026) e Frenda et al. (Frenda et al. 2025) apontam a importância de preservar diferentes interpretações humanas em vez de reduzir a anotação a um consenso único; ponto relevante aqui, pois uma mesma IA pode ser lida como ameaça, vítima, trabalhadora ou companheira conforme o foco.

3. Metodologia e desenho do corpus

A unidade de análise é a obra cultural, não apenas o texto verbal. Entram no escopo filmes, romances, contos, séries, peças, jogos, quadrinhos, pinturas, instalações e outras formas que mencionem, representem ou tematizem IA, sem se limitar a obras ficcionais canônicas; etapas futuras poderão incluir manifestações culturais urbanas, desde que com registro documentado e fonte identificável. Distinguimos três níveis de relevância: (i) IA como tema central; (ii) IA como elemento relevante para a narrativa, a estética ou o conflito; e (iii) IA como menção periférica. O piloto prioriza os dois primeiros.

Separamos ainda obras *sobre* IA de obras *feitas com* IA: estas só entram quando a IA também for objeto de representação ou reflexão. A distinção evita confundir duas perguntas: como a IA é representada culturalmente e como ela participa da produção cultural. As categorias não são fechadas; a taxonomia inicial é instrumento exploratório, refinado durante a anotação.

A língua das obras é uma questão metodológica relevante. Como o escopo é internacional, uma mesma obra circula em língua original e em traduções. Nesta fase, metadados e anotações são registrados em português brasileiro, mas a ficha preserva título original, título em português quando houver, língua original, edição ou tradução consultada e fonte. A tradução pode afetar escolhas lexicais, metáforas e a própria interpretação da representação da IA.

O CoRpMIA é, nesta etapa, um corpus anotado de metadados culturais com uma camada de evidências textuais ou multimodais. A inclusão de uma obra não exige redistribuir seu texto integral, sobretudo em obras protegidas por direitos autorais, mas cada item deve ter fonte verificável (domínio público, edição consultada, roteiro, trecho citável, sinopse autorizada, imagem documentada ou outro registro rastreável). Obras com texto integral legalmente disponível podem compor um subcorpus textual; nos demais casos, registram-se metadados, anotações e evidências selecionadas segundo critérios explícitos.

A ficha inclui: título original e em português, ano, tipo de obra, mídia, língua original, país ou região (de produção, histórica e atual), centralidade, tipo, papel e tom da

IA, temas principais, fonte, tipo de acesso e observações. A distinção entre região histórica e atual é necessária porque algumas obras foram produzidas em contextos geopolíticos diferentes; *R.U.R.*, por exemplo, é associada à Tchecoslováquia, embora hoje se relacione à tradição cultural tcheca.

A anotação será feita por pelo menos dois pesquisadores por obra. Em vez de reduzir os rótulos a uma decisão majoritária, a consolidação se dá por revisão conjunta dos casos divergentes: os anotadores confrontam suas leituras, e o corpus registra tanto o rótulo consolidado quanto o desacordo original, quando este for baseado em justificativa e não mero erro de anotação. Um item pode receber múltiplos rótulos, pois muitas obras representam a IA de modo ambivalente; essa multiplicidade é parte do fenômeno observado, tratada como evidência interpretativa e não como ruído. Nesta fase, com diferentes anotadores e escopo reduzido, não se calcula ainda uma medida formal de concordância; o objetivo é tornar o procedimento rastreável e preparar diretrizes para tal medida em etapas posteriores.

4. Amostra inicial

A Tabela 1 apresenta duas linhas de exemplo do corpus piloto: *Her*, em que a IA é assistente virtual e companheira afetiva, e *R.U.R. – Rossum's Universal Robots*, de Karel Čapek, associada à difusão do termo "robô" e a debates sobre trabalho artificial e revolta (Kussi 2019).

Tabela 1. Amostra inicial de campos do CoRpMIA

Campo	*Her*	*R.U.R. – Rossum's Universal Robots*
Título original	*Her*	*R.U.R. – Rossum's Universal Robots*
Ano	2013	1920
Tipo/mídia	Filme / cinema	Peça de teatro / teatro
Língua original	inglês	tcheco
País/região histórica	Estados Unidos	Tchecoslováquia
País/região atual	Estados Unidos	Tchéquia
Centralidade da IA	central	central
Tipo de IA	assistente virtual	robô; androide; outro
Papel da IA	companheira afetiva; espelho do humano	trabalhadora; ameaça; espelho do humano
Tom	melancólico; filosófico; ambíguo	distópico; crítico; filosófico
Fonte/acesso	roteiro consultável; não redistribuível no corpus aberto	texto integral em domínio público; pode compor subcorpus textual
Observações	Sistema operacional conversacional e afetivo; aproxima a interação humano-IA de uma relação emocional.	Genealogia anterior à IA computacional: seres artificiais ligados a trabalho, servidão e revolta.

As duas linhas demonstram a ficha, não a cobertura pretendida: ambas são obras ocidentais e canônicas, escolhidas pela facilidade de validação, enquanto a diversidade de origem e língua será buscada na consolidação. A amostra evidencia duas propriedades: o mesmo conjunto de categorias se aplica a mídias distintas, permitindo comparações futuras, e a coluna de fonte/acesso separa obras que compõem subcorpus textual integral das que só podem ser anotadas, por direitos autorais. A versão piloto pretende reunir 20 a 30 obras, conjunto que permitirá testar a consistência da ficha, identificar categorias

problemáticas e observar coocorrências iniciais.

5. Considerações finais

Este artigo apresentou a proposta inicial do CoRpMIA. Sua contribuição é metodológica, com um núcleo distintivo: em vez de reduzir a anotação a um consenso único, o corpus registra o desacordo interpretativo como evidência do próprio fenômeno cultural, em diálogo com abordagens perspectivistas em PLN (Gebru et al. 2021; Basile 2026; Frenda et al. 2025). A isso somam-se uma unidade de análise que atravessa mídias, uma ficha testável, uma taxonomia preliminar derivada da literatura de narrativas de IA, o tratamento explícito da língua e a documentação responsável do corpus. Como próximo passo, pretendemos consolidar a versão piloto, ampliar a amostra em origem e língua, produzir seu *dataset card* e avaliar a publicação aberta dos metadados, respeitando direitos autorais.

Agradecimentos

Este trabalho foi realizado no âmbito do Instituto Kunumi e do Instituto Nacional de Ciência e Tecnologia em Tecnologias de Linguagem para Inteligência Artificial Responsável (INCT-TILDIAR).

Referências

Abercrombie, Gavin et al. (2023) Mirages: On Anthropomorphism in Dialogue Systems. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4776–90. https://doi.org/10.18653/v1/2023.findings-emnlp.319.

Aktaş, Seda. (2026) The Evolution of the Representations of Artificial Intelligence in Cinema: An Analysis of the Last 10 Years Based on the IMDb Dataset. *Erciyes İletişim Dergisi* 13 (1): 227–48. https://doi.org/10.17680/erciyesiletisim.1756059.

Basile, Valerio. (2026) *Modeling Human Perspectives in Natural Language Processing*.

Berber Sardinha, Tony. (2004) *Linguística de Corpus*. Barueri, SP: Manole.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. (2012) *Digital_humanities*. Cambridge, MA: MIT Press.

Cave, Stephen, Kanta Dihal, and Sarah Dillon, eds. (2020) *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford: Oxford University Press.

Freitas, Cláudia. (2023) Dataset e Corpus. In: *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações Em Português*, edited by Helena de Medeiros Caseli and Maria das Graças Volpe Nunes. Brasileiras em PLN.

Frenda, Simona et al. (2025 Perspectivist Approaches to Natural Language Processing: A Survey. *Language Resources and Evaluation* 59 (2): 1719–46. https://doi.org/10.1007/s10579-024-09766-4.

Gebru, Timnit et al. (2021) Datasheets for Datasets. *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

Hermann, Isabella. (2023) Artificial Intelligence in Fiction: Between Narratives and Metaphors. *AI & Society* 38: 319–29. https://doi.org/10.1007/s00146-021-01299-6.

Jasanoff, Sheila, and Sang-Hyun Kim, eds. (2015) *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: University of Chicago Press.

Kussi, Peter (2019) The Czech Play That Gave Us the Word 'Robot'.

Luo, Xiaochen et al. (2025) Seeking Emotional and Mental Health Support from Generative AI: Mixed-Methods Study of ChatGPT User Experiences. *JMIR Mental Health* 12: e77951. https://doi.org/10.2196/77951.

Sharma, Mrinank et al. (2024) Towards Understanding Sycophancy in Language Models. In: *International Conference on Learning Representations (ICLR)*.

Zandavalle, Ana Claudia, and Livy Real (2023) Transdisciplinaridade Na Tecnologia Da Linguagem. In *XV Encontro de Linguística de Corpus: O Alcance Da Linguística de Corpus: Do Léxico Ao Discurso*, edited by Regiani A. S. Zacarias and Stella E. O. Tagnin, 145–55. São Paulo: Ed. dos Autores.

Uma proposta de agenda data-driven para o ensino de italiano em contexto universitário

Graziele Altino Frangiotti

Universidade Federal de Santa Catarina

grazielefrangiotti@gmail.com

Resumo. *O presente trabalho propõe o desenvolvimento de uma agenda de pesquisa orientada por dados para o ensino de italiano em contexto universitário. Parte-se da hipótese de que a abordagem Data-driven learning (DDL) pode promover aprendizagem mais ativa e significativa, estimulando a reflexão sobre o funcionamento da linguagem e o desencadeamento de processos inferenciais complexos, fundamentais para a aprendizagem crítica e autônoma. Diante da escassez de pesquisas no Brasil que investiguem os efeitos do DDL no ensino de italiano, propõe-se uma agenda de pesquisa que prevê: a análise linguística de usos concretos do italiano a partir de corpora disponíveis, a transposição didática dos resultados da descrição linguística para o contexto de ensino e a elaboração de sequências didáticas DDL. A premissa central é que a integração entre descrição linguística orientada por corpora e a produção de materiais didáticos orientados por dados pode fomentar novas perspectivas para o ensino de italiano e desenvolver a capacidade dos aprendizes de refletir criticamente sobre a linguagem e seus usos.*

Palavras-chave: Data-Driven Learning; Linguística de Corpus; Ensino de italiano.

Abstract. *This investigation proposes the development of a data-driven research agenda for teaching Italian in university contexts. It is based on the hypothesis that this approach can promote more active and meaningful learning by encouraging reflection on how language works and fostering complex inferential processes that are essential to critical and autonomous learning. Given the scarcity of research in Brazil investigating the effects of Data-Driven Learning (DDL) on Italian language teaching, this study proposes a research agenda encompassing the linguistic analysis of actual uses of Italian based on available corpora, the didactic transposition of linguistic description findings into the teaching context, and the development of DDL-based instructional sequences. The central premise is that integrating corpus-informed linguistic description with the development of data-driven teaching materials can foster new perspectives on Italian language teaching and enhance learners' ability to reflect critically on language and its uses.*

Keywords: Data-Driven Learning; Corpus Linguistics; Italian language teaching.

1. Introdução

Nos últimos anos, e de forma ainda mais aguda após a pandemia de Covid-19, a educação básica no Brasil vem atravessando uma profunda crise. Cotidianamente surgem notícias

relativas a níveis de alfabetização ainda insatisfatórios nas escolas, a dificuldades ligadas à capacidade de resolver problemas e desenvolver formas de raciocínio lógico-matemático, bem como a dificuldades crescentes de concentração por parte dos estudantes. A esse cenário soma-se uma verdadeira invasão das tecnologias digitais e daquelas baseadas em inteligência artificial, que, não raramente, são utilizadas de maneira pouco consciente e crítica, contribuindo para tornar ainda mais complexa uma questão que, aliás, nunca foi simples no país.

A constatação da situação atual da educação brasileira evidencia que a universidade, enquanto espaço privilegiado de produção do conhecimento e de reflexão crítica sobre a sociedade, não pode se furtar à necessidade de propor novas formas de pensar e organizar as práticas pedagógicas na tentativa de enfrentar tais dificuldades. Essa responsabilidade se mostra particularmente relevante nos cursos de Letras, que, se por um lado, têm como cerne justamente a reflexão sobre o papel da linguagem na constituição dos indivíduos e na atividade social, por outro, constituem uma etapa fundamental na formação de muitos profissionais que atuarão na educação básica e que, portanto, farão parte desse sistema que hoje apresenta importantes limitações. Logo, se a transformação das práticas educacionais se mostra necessária, é também no espaço universitário, e particularmente na formação inicial de professores, que parte desse processo precisa ser construída.

Nesse contexto, o ensino de línguas estrangeiras representa um terreno particularmente fértil, uma vez que o trabalho com línguas distintas favorece processos de observação e interpretação, os quais podem ser ainda mais potencializados se associados a atividades que estimulem a comparação, a análise atenta e o teste de hipóteses. Assume-se que atividades dessa natureza podem contribuir não apenas para o desenvolvimento da competência comunicativa em língua estrangeira, mas também para a criação de um ambiente de aprendizagem que favoreça o exercício da atenção, do raciocínio analítico e da leitura crítica**,** aspectos que dialogam diretamente com alguns dos desafios apontados no início deste texto.

Levando isso em consideração, a presente proposta busca refletir sobre a construção de uma agenda orientada por dados, voltada, especificamente, ao ensino de italiano em contexto universitário. Procura-se investigar possibilidades de integração da análise de dados linguísticos ao ensino dessa língua desde os níveis iniciais de aprendizagem, bem como compreender como esse processo pode contribuir para reconfigurar as práticas didáticas, mediante a articulação entre Linguística de Corpus, métodos quantitativos e qualitativos de descrição linguística e *Data-Driven Learning*. De modo mais específico, pretende-se, com essa agenda, responder à seguinte pergunta: Como a análise sistemática de dados linguísticos do italiano contemporâneo pode informar o ensino de italiano como língua estrangeira e orientar práticas pedagógicas de natureza investigativa?

2. O emprego de dados no ensino de línguas

Vivemos em uma sociedade cada vez mais orientada por dados, a ponto de a atual fase histórica ser definida por muitos como a era do Big Data (cf. Mayer-Schönberger; Ramge; Maugeri, 2018; Leonelli, 2018). A chamada *datafication* (ou datificação), que constituiu o processo por meio do qual aspectos cada vez mais amplos da vida individual e coletiva são transformados em dados quantificáveis e analisáveis por meio de ferramentas

computacionais (Mayer-Schönberger; Cukier, 2013: 78), representa uma das transformações mais significativas da contemporaneidade. Tanto que se prevê que esse fenômeno terá um impacto comparável ao produzido pela Revolução Industrial, incidindo profundamente sobre a forma como os indivíduos pensam, produzem conhecimento, tomam decisões, interagem com o ambiente e organizam a vida social.

As transformações geradas pelo enorme volume de dados hoje disponível não se limitam a um único campo disciplinar. Na informática, por exemplo, áreas como a inteligência artificial e o *machine learning* tornam-se cada vez mais centrais. Na economia e nas finanças, os dados favorecem o crescimento do uso de modelos preditivos cada vez mais sofisticados. No marketing, a datificação dos hábitos e dos estilos de vida dos consumidores transformou profundamente as estratégias de análise do comportamento de consumo. Desenvolver a *data literacy,* portanto, entendida como a capacidade de observar, interpretar e analisar criticamente dados, constitui hoje uma necessidade concreta para qualquer percurso formativo que pretenda dialogar com a realidade contemporânea e confrontar-se com as múltiplas dimensões de sua complexidade.

Em decorrência disso, para que a Educação Linguística esteja em coerência com o seu tempo, é necessário que estimule também essa nova forma de letramento, dando espaço a práticas mais ativas de observação, interpretação e construção autônoma do conhecimento. Essa necessidade é ainda mais premente no contexto brasileiro, onde as experiências pedagógicas com dados permanecem restritas ao âmbito do ensino de inglês e configuram-se mais como iniciativas isoladas ou experimentais do que como práticas sistemáticas.

Nesse sentido, a ideia de utilizar dados no ensino de línguas certamente não constitui uma novidade absoluta. Já na década de 1990, Tim Johns questionava como levar dados linguísticos autênticos para a sala de aula sem sobrecarregar os recursos cognitivos e linguísticos dos aprendizes de uma segunda língua. Como resposta a esse desafio, propôs o Data-Driven Learning (DDL), uma perspectiva didática baseada no trabalho com *corpora* digitais interrogáveis por meio de ferramentas computacionais específicas, que se fundamenta na premissa de que os aprendizes devem assumir um papel mais protagonista no processo de aprendizagem por meio da observação direta dos dados linguísticos.

Embora a revisão da bibliografia internacional evidencie um aumento do número de pesquisas desenvolvidas no âmbito do DDL (Boulton; Vyatkina, 2021; Boulton, 2017), ao se concentrar a atenção no contexto brasileiro, é possível identificar duas lacunas principais: uma de natureza linguística e outra de natureza metodológica. No que se refere à primeira, as experiências didáticas inspiradas no DDL ainda se encontram predominantemente circunscritas ao ensino de inglês (Pinto; Marques; Garcia, 2025; Tagnin; Murakami, 2014; Santos, 2011; Dayrell, 2011), havendo algumas exceções representadas por estudos dedicados ao francês (Zavaglia; Celli, 2011), ao espanhol (Pagano; Figueredo, 2011), ao alemão (Aquino, 2018, 2019; Finatto; Zilio; Migotto, 2011), ao português (Oliveira, 2005; Zhang; You, 2024; Martins, 2022) e ao italiano (Frangiotti, 2021). Do ponto de vista metodológico, muitos desses trabalhos se limitam a oferecer propostas para a integração do DDL nas aulas de línguas, sendo muito menos frequentes os estudos que descrevem e analisam experiências efetivamente realizadas em

sala de aula (Mareco; Maracci, 2010; Pinto; Souza, 2011; Frangiotti, 2021; Oliveira; Ottaiano, 2017; Garcia, 2002; Zuppardi, 2017).

Considerando esse conjunto de informações preliminares e a ausência, até o presente momento, de estudos sistemáticos que permitam compreender a presença e as potencialidades do uso de *corpora* linguísticos no ensino de italiano no Brasil, verifica-se a necessidade de inserir uma agenda de pesquisa específica que abarque investigações que, além de ampliar a presença do DDL nesse contexto, permitam indagar e documentar os efeitos e as potencialidades do uso de *corpora* para a aprendizagem e para a reflexão linguística.

3. Proposta metodológica

Buscando contribuir para a diminuição das lacunas apresentadas na seção anterior, o presente trabalho propõe uma perspectiva na qual o dado linguístico assume o papel de princípio organizador do percurso de pesquisa e de intervenção didática. A proposta fundamenta-se na ideia de que a observação sistemática de dados linguísticos pode constituir o ponto de partida para a descrição dos fenômenos da língua estudada, para a elaboração de materiais didáticos, para a organização das ações pedagógicas e para a posterior verificação da eficácia das práticas de ensino.

Parte-se do pressuposto de que o ensino de uma língua estrangeira pode se beneficiar de uma compreensão mais sistemática dos usos efetivamente atestados na língua em diferentes contextos comunicativos. Fenômenos lexicais, morfossintáticos, pragmáticos e discursivos apresentam padrões de distribuição e funcionamento que nem sempre são plenamente contemplados nas descrições presentes em materiais didáticos tradicionais. Nessa direção, a primeira etapa da agenda de pesquisa aqui proposta consiste em uma investigação do italiano contemporâneo a partir de *corpora* linguísticos disponíveis, selecionados de acordo com sua representatividade, composição e possibilidades de exploração. Serão privilegiados, inicialmente, *corpora* que permitam observar diferentes modalidades e contextos de uso, como o CORIS/CODIS, o PAISÀ e, quando pertinente, *corpora* de língua falada, como o KIParla.

A partir dos resultados obtidos por meio dessas observações, serão selecionados fenômenos linguísticos que apresentem relevância simultaneamente descritiva e pedagógica e que ofereçam potencial para o desenvolvimento de atividades orientadas por dados. Em vez de se limitar à apresentação de exemplos artificialmente construídos, as atividades serão desenvolvidas segundo os princípios do *Data-Driven Learning*, colocando o aprendiz em contato com evidências linguísticas e incentivando-o a observar, comparar, formular hipóteses e identificar regularidades no uso da língua. Em particular, poderão ser exploradas linhas de concordância, listas de frequência, informações sobre coocorrência e outras formas de representação dos dados, adaptadas ao nível de proficiência e aos objetivos de aprendizagem. Desse modo, os dados do corpus não funcionarão apenas como fonte de exemplos, mas como objeto de investigação e ponto de partida para a construção do conhecimento linguístico.

A etapa seguinte prevê a experimentação dos materiais em contextos reais de ensino. Por meio da observação sistemática das aulas, da análise das produções dos estudantes e da coleta de dados relativos às suas percepções e experiências, será possível avaliar os efeitos das práticas data-driven sobre o processo de aprendizagem. Essas

informações constituirão uma base importante para a reflexão crítica sobre a eficácia do percurso desenvolvido e para eventuais reformulações dos materiais e das técnicas didáticas empregadas. As diferentes etapas descritas darão origem a um processo cíclico e dinâmico que integrará pesquisa, desenvolvimento de material didático DDL, experimentação em sala de aula e revisão contínua.

4. Considerações finais

Em um contexto em que os aprendizes têm demonstrado contínua dificuldade com leitura e com o desenvolvimento de raciocínios complexos, torna-se necessário oferecer ocasiões para uma aprendizagem ativa que envolva observação atenta, levantamento de hipóteses, comparação e discussão. Nesse cenário, parece plausível levantar a hipótese de que o trabalho pedagógico orientado por dados possa representar uma possibilidade pedagógica a ser experimentada.

Já se tem para línguas como o inglês conhecimento informado a respeito, confirmando os efeitos positivos desse tipo de abordagem. No caso do italiano, especificamente em contexto brasileiro, não se tem ainda grande volume de investigações nesse sentido. Logo, uma agenda data-driven para esse idioma pode representar não apenas uma possibilidade de inovação metodológica, mas também uma oportunidade concreta de integrar a descrição linguística à prática pedagógica.

Mais do que introduzir novas tecnologias em sala de aula, uma agenda dessa natureza busca promover práticas capazes de desenvolver a autonomia, a reflexão crítica e a consciência linguística dos aprendizes, contribuindo para uma formação profissional mais alinhada às demandas da realidade contemporânea.

Questões de ética

No presente resumo, declara-se que o uso da IA se restringiu ao emprego da ferramenta ChatGPT para correção de eventuais erros ortográficos e sintáticos, para a pesquisa de sinônimos e para a realização da tradução do resumo para o inglês. A autoria do artigo é de inteira responsabilidade da pesquisadora propositora.

Referências

Aquino, M. A contribuição da ferramenta de corpus para o ensino das partículas modais denn, doch, halt e wohl. Diálogo das Letras, v. 8, n. 1, p. 144-158, 2019. DOI: https://doi.org/10.22297/dl.v8i1.3500.

Boulton, A. Research timeline: Corpora in language teaching and learning. Language Teaching, v. 50, n. 4, p. 483-506, 2017. DOI: https://doi.org/10.1017/S0261444817000167.

Boulton, A.; Vyatkina, N. Thirty years of data-driven learning: Taking stock and charting new directions over time. Language Learning & Technology, v. 25, n. 3, p. 66-89, 2021.

Dayrell, C. Corpora no ensino do inglês acadêmico: padrões léxico-gramaticais em abstracts de pós-graduandos brasileiros. In: VIANA, V.; TAGNIN, S. (org.). Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2010. p. 137-171.

Finatto, M. J. B.; Zilio, L.; Migotto, E. J. Artigos de cardiologia em português e alemão: contribuições da pesquisa em corpus para o ensino de leitura instrumental. In: VIANA, V.; TAGNIN, S. (org.). Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2010. p. 205-234.

Frangiotti, G. A. O efeito da Linguística de Corpus na formulação de material didático e na promoção de output significativo em aulas de italiano/LE. Tradterm, São Paulo, v. 37, p. 119-147, 2021. DOI: https://doi.org/10.11606/issn.2317-9511.v37p119-147.

Johns, T. From printout to handout: Grammar and vocabulary teaching in the context of Data-Driven Learning. In: ODLIN, T. (org.). Perspectives on pedagogical grammar. Cambridge: Cambridge University Press, 1994. p. 28-45.

Leonelli, S. La ricerca scientifica nell'era dei Big Data. Milano: Meltemi, 2018.

Mareco, R. T. M.; Maracci, C. C. Linguística de corpus e ensino de inglês instrumental. Colloquium Humanarum, v. 7, n. 1, p. 1-7, 2010. DOI: https://doi.org/10.5747/ch.2010.v07.n01.h072.

Martins, F. Corpus computadorizado contribuindo para o ensino e pesquisa no campo da língua portuguesa e literatura. Brazilian Journal of Development, Curitiba, v. 8, n. 9, p. 64888-64902, 2022. DOI: https://doi.org/10.34117/bjdv8n9-297.

Mayer-Schönberger, V.; Cukier, K. Big Data: a revolution that will transform how we live, work and think. London: John Murray, 2013.

Mayer-Schönberger, V.; Ramge, T.; Maugeri, G. Reinventare il capitalismo nell'era dei Big Data. Milano: Egea, 2018.

Mejias, U. A.; Couldry, N. Datafication. Internet Policy Review, v. 8, n. 4, 2019. DOI: https://doi.org/10.14763/2019.4.1428.

Oliveira, E. N. A Linguística de Corpus no ensino-aprendizagem de línguas: um estudo de caso das práticas de professores da educação básica de Dourados (MS). 2005. Tese (Doutorado) – Universidade Federal de Santa Catarina, Florianópolis, 2005.

Oliveira, E. C.; Ottaiano, A. O. Harry Potter e a Linguística de corpus em aulas de língua inglesa para o ensino fundamental I. Desenredo, v. 13, n. 3, p. 609-627, 2017.

Pagano, A. S.; Figueredo, G. P. Gramaticalização da dor em português e espanhol: uma abordagem comparada com subsídios da Linguística de Corpus e da linguística sistêmico-funcional. In: VIANA, V.; TAGNIN, S. (org.). Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2010. p. 173-187.

Pinto, M.; Souza, R. C. Filmes comerciais: uma perspectiva da aplicação de pesquisa em corpus na sala de aula. In: VIANA, V.; TAGNIN, S. (org.). Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2010. p. 173-187.

Pinto, P. T.; Marques, S. M. K.; Garcia, D. N. M. Corpus e escrita acadêmica: uma estratégia data-driven para a internacionalização da produção científica. Trabalhos em Linguística Aplicada, v. 64, p. 1-11, 2025.

Plonski, G. A. Inovação em transformação. Estudos Avançados, São Paulo, v. 31, n. 90, p. 7-21, 2017. DOI: https://doi.org/10.1590/S0103-40142017.3190002.

Santos, A. G. Working closely with corpus: análise de colocações adverbiais em inglês para negócios. In: VIANA, V.; TAGNIN, S. (org.). Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2010. p. 97-136.

Tagnin, S.; Murakami, D. Linguística de Corpus e ensino: a compilação de um corpus de especialidade para preparação e implementação de um curso preparatório rápido para exame de proficiência. Letras & Letras, v. 30, n. 2, p. 349-366, 2014.

Zavaglia, A.; Celli, M. Aplicação do conceito de transcategorialidade ao ensino de francês como língua estrangeira: o caso da marca "mas". In: VIANA, V.; TAGNIN, S. (org.). Corpora no ensino de línguas estrangeiras. São Paulo: Hub Editorial, 2010. p. 235-267.

Zhang, J.; You, M. Corpus de aprendizes de português da Universidade de Macau e ensino de português L2. Texto Livre, Belo Horizonte, v. 17, e47754, 2024. DOI: https://doi.org/10.1590/1983-3652.2024.47754.

Zuppardi, M. C. O uso de corpus especializado no ensino público de língua inglesa. Caderno Seminal Digital, n. 27, p. 304-338, 2017.

Análise de convenções disciplinares em *abstracts* gerados por inteligência artificial

Gabriela Carvalho Escobar, Ana Eliza Pereira Bocorny

Universidade Federal do Rio Grande do Sul (UFRGS) - Instituto de Letras

gabrielacarvalhoescobar@gmail.com, ana.bocorny@gmail.com

Abstract. *This study investigates whether generative AI systems reproduce linguistic conventions typical of the Humanities. A corpus of 160 abstracts from CORHUM (InPACT-CNPq) is compared with abstracts generated by ChatGPT, Gemini, and Claude, focusing on format, extension and lexical bundles, and commenting rhetorical functions. Preliminary findings suggest divergences and similarities to human-authored texts and different degrees of disciplinary alignment across models.*

Resumo. *O estudo investiga se sistemas de IA generativa reproduzem convenções retóricas e fraseológicas das Humanidades. Foram comparados 160 abstracts do corpus CORHUM (InPACT-CNPq) e resumos produzidos por ChatGPT, Gemini e Claude. A análise contempla formato, extensão e pacotes lexicais, comentando aspectos de função retórica. Resultados preliminares apontam divergências e semelhanças entre textos humanos e gerados por IA, além de diferentes níveis de adequação disciplinar entre os modelos.*

1. Introdução

Desde os trabalhos seminais de Swales (1990), os gêneros acadêmicos têm sido compreendidos como práticas discursivas vinculadas a comunidades disciplinares e a propósitos comunicativos específicos. Entre os elementos que materializam tais convenções, destacam-se os padrões recorrentes de linguagem formulaica, especialmente os *lexical bundles (*pacotes lexicais), cuja distribuição varia conforme disciplina, gênero e função discursiva (Biber; Barbieri, 2007; Cortes, 2004; Hyland, 2008).

Paralelamente, a expansão recente de sistemas de Inteligência Artificial generativa vem transformando práticas de produção textual em contextos acadêmicos. Sistemas baseados em Large Language Models (LLMs), como ChatGPT (OpenAI, 2026), Gemini (Google, 2026), e Claude (Anthropic, 2026), passaram a ser utilizados para revisão, reformulação e produção de textos científicos, provocando discussões sobre autoria, ética, confiabilidade e qualidade linguística (Harati, 2024; William, 2024). Apesar disso, ainda permanecem pouco exploradas questões relacionadas à capacidade que esses sistemas têm de reproduzir os padrões retóricos e linguísticos convencionalmente usados em diferentes disciplinas.

Do ponto de vista da Linguística de Corpus, a questão torna-se particularmente relevante: seriam os modelos capazes de reproduzir convenções linguísticas próprias de diferentes áreas do conhecimento ou tenderiam a uniformizar padrões discursivos? Considerando que estudos anteriores demonstram que convenções disciplinares se manifestam em diversos aspectos dos gêneros acadêmicos, como formato, extensão linguagem formulaica (Moreno; Swales, 2018; Gray; Cotos; Smith, 2020), investigar a produção textual de sistemas de IA torna-se fundamental para avaliar sua capacidade de

reproduzir padrões discursivos característicos de diferentes disciplinas acadêmicas.

Nesse contexto, o presente estudo investiga se *abstracts* gerados por IA reproduzem convenções linguísticas (formato, extensão e *lexical bundles)* relacionadas a retóricas características das Humanidades ou se apresentam padrões mais próximos de modelos generalizados de escrita científica. Traçamos também uma discussão sobre a função retórica que os *lexical bundles* podem desempenhar nesse contexto.

2. Metodologia

O estudo adota uma abordagem baseada em corpus e articula a Linguística de Corpus, os estudos sobre gêneros acadêmicos e a pesquisa sobre linguagem formulaica. Seu objetivo é comparar *abstracts* produzidos por autores humanos e por sistemas de IA generativa quanto ao formato, extensão e uso de *lexical bundles*.

O corpus de referência deriva do Corpus das Humanidades (CORHUM), desenvolvido no âmbito do projeto InPACT-CNPq (Bocorny et al., 2024). O CORHUM compreende corpora especializados compilados com o programa AntCorGen (Anthony, 2026) a partir da plataforma PLOS ONE, organizados em oito disciplinas - Antropologia, Arqueologia, Linguística, Língua e Artes, Psicologia, Educação, Ciências Políticas, Direito e Ciências Jurídicas e Economia - e subdivididos segundo as seções de artigos científicos.

Para o presente estudo, selecionaram-se vinte *abstracts* de cada disciplina, a partir de uma amostragem aleatória (*seed=42*), totalizando 160 textos no nosso corpus, que denominamos CAAH (Corpus Amostral de Abstracts das Humanidades). A escolha dos *abstracts* se dá por seu papel central como gênero de apresentação da pesquisa científica e por sua relevância para análises de linguagem formulaica (Swales, 1990).

A coleta e o processamento dos dados foram automatizados por meio da aplicação CompareAI, desenvolvida para esta pesquisa. A ferramenta recupera através do DOI os 160 artigos correspondentes aos *abstracts* selecionado, envia os textos a diferentes modelos de IA generativa (ChatGPT (OpenAI, 2026), Gemini (Google, 2026) e Claude (Anthropic, 2026)) via API, e organiza os *outputs* gerados em um banco de dados para análise comparativa. Para cada artigo, o texto integral foi submetido aos três modelos mediante um mesmo *prompt*, resultando em um corpus de resumos produzidos por IA para comparação com os *abstracts* originais.

A análise mobiliza três dimensões complementares. Primeiramente, examina-se o formato dos *abstracts,* caso sejam estruturados ou escritos em bloco. Em um segundo momento, é realizada a análise da extensão dos textos autênticos e dos gerados por IA. Em seguida, analisam-se os *lexical bundles* que ocorrem nos textos autênticos e aqueles que ocorrem nos textos gerados, com os seguintes critérios: extensão de 4-6 palavras, frequência mínima de 8 ocorrências e dispersão em no mínimo 5 textos diferentes. Por fim, são traçados comentários sobre a função retórica dos *lexical bundles* de acordo com estudos anteriores (Escobar, 2024). A análise dos padrões formulaicos apoia-se tanto em listas previamente produzidas pelo projeto InPACT-CNPq quanto em novas extrações realizadas sobre o corpus gerado por IA, em consonância com abordagens da Linguística de Corpus (Berber-Sardinha, 2004; Flowerdew, 2009).

3. Resultados Preliminares

Para apresentar os resultados preliminares desta pesquisa, utilizaremos os dados coletados das disciplinas de Psicologia, Ciências Políticas e Linguística, Língua e Artes. De acordo com a Tabela 1, observa-se que, nos textos autênticos, *abstracts* das Humanidades são mais frequentemente escritos em bloco, especialmente na disciplina de Linguística, Língua e Artes. Nos *abstracts* gerados por IA, apenas 10% foram escritos com rótulos, como estruturados, e apenas foram gerados pelo modelo da IA Claude na disciplina de Ciências Políticas. Isso pode sugerir que a inteligência artificial opta pelo formato mais consolidado em geral, que é o *abstract* escrito em bloco, refletindo um padrão de escrita das Humanidades.

Tabela 1. Comparação do Formato dos *abstracts*

	Formato CAAH		Formato CompareAI	
	Est.	Bloco	Est.	Bloco
Psicologia	8 (40%)	12 (60%)	0	20 (100%)
Ciências Políticas	7 (35%)	13 (65%)	2 (10%)	18 (90%)
Linguística, Língua e Artes	2 (10%)	18 (90%)	0	20 (100%)

No que diz respeito à extensão dos textos, identificou-se variação entre os sistemas analisados. A Tabela 2 demonstra de forma nítida que o modelo do Gemini apresenta uma maior correspondência com o número de palavras dos *abstracts* originais - o que também é consequência do fato de os outros sistemas de IA analisados escreverem textos muito mais extensos. Isso pode indicar tanto que o Gemini utiliza os padrões das Humanidades, quanto que sua concepção de *abstract* é de um texto de pequena extensão. Enquanto ChatGPT e Gemini costumam manter uma média de palavras similar ou pouco maior à convencional para um *abstract* (200-300 palavras), o sistema Claude produz textos mais longos, fugindo da convenção e levantando questões sobre a diferença do treinamento de cada um dos sistemas.

Tabela 2. Comparação da Extensão dos *abstracts*

	CAAH	CompareAI		
		Gemini	ChatGPT	Claude
Psicologia	229	226	280	378
Ciências Políticas	269	238	300	404
Linguística, Língua e Artes	229	212	280	363

Por fim, no nível de ocorrência de *lexical bundles,* os resultados são de relevância para análise, como indica a Tabela 3. Dos 180 *bundles* da lista do CORHUM para Psicologia, 12 são utilizados em *abstracts* da área pelo Claude, 4 pelo ChatGPT e 3 pelo Gemini. Dos 37 *bundles* de Ciências Políticas, nenhum é utilizado pelas IAS. Dos 78 do campo de

Linguística, Língua e Artes, apenas 1 é utilizado pelo Gemini e 2 pelo Claude. Também foram extraídos os *lexical bundles* dos corpora final de IA de cada disciplina, seguindo os critérios informados na metodologia. Apenas na área de Linguística a frequência mínima estipulada não foi alcançada, limitando-se a 7 ocorrências do *bundle* no corpus de IA.

Tabela 3. Comparação da Ocorrência de *lexical bundles*

	Ocorrências no CompareAI contidas no CORHUM		Bundles extraídos do corpus de IA com os critérios estabelecidos
Psicologia	Claude	the present study aimed to aim of the present study aim of the present study was of the present study was of the present study was to the aim of the present the aim of the present study the present study was to little is known about how the presence or absence of these findings suggest that the study aimed to evaluate the	these findings suggest that this study examined whether these findings demonstrate that
	Gemini	the psychometric properties of the the results showed that the these findings suggest that the	
	ChatGPT	were more likely to be consistent with the idea that little is known about the results are consistent with the	
Ciências Políticas	-		these findings suggest that
Linguística, Língua e Artes	Claude	the effectiveness of the proposed these findings suggest that the	these findings suggest that*
	ChatGPT	little is known about how	

Diante disso, percebe-se que a maioria dos modelos recorre a sequências lexicais que não coincidem com os padrões identificados no CORHUM. Tal comportamento pode refletir a composição dos dados de treinamento e a predominância de determinadas convenções discursivas. O fato de a disciplina de Psicologia aproximar-se mais da área da Saúde do que as demais disciplinas, pode ser um fator causal de seus elementos convencionais terem sido reproduzidos com maior adequação. No que tange às funções retóricas performadas pelos *lexical bundles*, pode-se discutir a ocorrência de sequências ligadas à apresentação dos objetivos de pesquisa e a maior frequência no corpus de IA de *bundles* relacionados à interpretação de resultados (Escobar, 2024).

Do ponto de vista pedagógico, os achados dialogam com estudos que defendem o

uso de corpora e ferramentas computacionais como suporte à escrita acadêmica (Flowerdew, 2009; Frankenberg-Garcia et al., 2019), podendo ser incorporados a aplicações como as desenvolvidas pelo Projeto Inpact para o apoio a pesquisadores. Os resultados também apontam para a necessidade de ferramentas de IA mais sensíveis à variação disciplinar, particularmente em contextos de Inglês para Fins Acadêmicos.

Questões Éticas e Uso de IA

A pesquisa utiliza exclusivamente artigos científicos publicados em plataforma de acesso aberto. Os sistemas de IA foram empregados como objeto de investigação e como apoio técnico no desenvolvimento da aplicação computacional utilizada na coleta e organização dos dados. Todas as decisões relativas à fundamentação teórica, metodologia, interpretação dos resultados e redação científica permaneceram sob responsabilidade das autoras.

Referências

ANTHROPIC. (2026). Claude API. https://docs.anthropic.com/

BERBER-SARDINHA, T. Linguística de Corpus. Barueri: Manole, 2004.

BHATIA, V. Analysing Genre: Language Use in Professional Settings. London: Longman, 1993.

BHATIA, V. Genre-mixing in academic introductions. English for Specific Purposes, v.16, n.3, p.181-195, 1997.

BIBER, D.; BARBIERI, F. Lexical bundles in university spoken and written registers. English for Specific Purposes, v.26, n.3, p.263-286, 2007.

CORTES, V. Lexical bundles in published and student disciplinary writing: examples from history and biology. English for Specific Purposes, v.23, n.4, p.397-423, 2004.

ESCOBAR, G. C. Lexical bundles in the writing of research article abstracts: proposing a corpus-based pedagogical application. 2024. Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2024.

FLOWERDEW, L. Applying corpus linguistics to pedagogy: a critical evaluation. International Journal of Corpus Linguistics, v.14, n.3, p.393-417, 2009.

FRANKENBERG-GARCIA, A. et al. Developing a writing assistant to help EAP writers with collocations in real time. ReCALL, v.31, n.1, p.23-39, 2019.

GOOGLE. (2026). *Gemini API*. https://ai.google.dev/

GRAY, B.; COTOS, E.; SMITH, S. Combining rhetorical move analysis with multi-dimensional analysis: research writing across disciplines. In: RÖMER, U.; CORTES, V.; FRIGINAL, E. (org.). Advances in Corpus-Based Research on Academic Writing. Amsterdam: John Benjamins, 2020.

HARATI, K. ChatGPT and AI-powered writing tools: unveiling risks and ethical challenges in scientific writing. Journal of Reviews in Medical Sciences, v.4, n.1, p.1-6, 2024.

HYLAND, K. As can be seen: lexical bundles and disciplinary variation. English for Specific Purposes, v.27, n.1, p.4-21, 2008.

MORENO, A. I.; SWALES, J. M. Strengthening move analysis methodology towards bridging the function-form gap. English for Specific Purposes, v.50, p.40-63, 2018.

OPENAI. (2026). *OpenAI API*. https://platform.openai.com/docs

SWALES, J. M. Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press, 1990.

WILLIAM, F. K. A. AI in academic writing: ally or foe. International Journal of Research Publications, v.148, n.1, p.353-358, 2024.

A prosódia semântica na linguagem das Ciências Agrárias: contribuições da Linguística de *Corpus* para o desenvolvimento de materiais didáticos de inglês para fins acadêmicos

Ana Carolina Martins Párcio, Deise Prina Dutra, Ludmila da Silva Marques, Mariana Gonzaga Silva, Marjorie Lourdes Almeida Corrêa, Paula Tavares Pinto, Shirlene Bemfica de Oliveira

Abstract. *This paper investigates semantic prosody in lexical clusters associated with the node "animal" in the Corpus of Research Papers in Agricultural Sciences (CorAgrarian). The study aims to identify recurrent lexical patterns that may inform the development of English for Academic Purposes (EAP) materials grounded in ecolinguistic principles. A keyword analysis was conducted using AntConc (Anthony, 2024), and "animal" was selected for further investigation. Four clusters—"animal care", "animal production", "animal husbandry", and "animal protein"—were examined through their frequencies, distributions, and concordance contexts. The preliminary qualitative analysis suggests that three clusters are predominantly associated with positive evaluative contexts, particularly those involving animal welfare and production, whereas "animal husbandry" is associated with negative environmental impacts in the examples discussed. These findings illustrate how corpus-based analysis can identify lexical and discursive patterns relevant to academic language teaching and may provide a basis for materials that encourage critical engagement with sustainability-related issues in Agricultural Sciences.*

Keywords: *ecolinguistics; corpus linguistics; english for academic purposes.*

Resumo. *Este artigo investiga a prosódia semântica de agrupamentos lexicais associados ao nódulo "animal" no Corpus de Artigos Científicos das Ciências Agrárias (CorAgrarian). O estudo tem como objetivo identificar padrões lexicais recorrentes que possam subsidiar a elaboração de materiais de Inglês para Fins Acadêmicos (IFA) fundamentados em princípios ecolinguísticos. Para isso, realizou-se uma análise de palavras-chave com o AntConc (Anthony, 2024), selecionando-se "animal" para investigação. Quatro agrupamentos—"animal care", "animal production", "animal husbandry" e "animal protein"—foram examinados considerando sua frequência, distribuição e contextos de concordância. A análise qualitativa preliminar sugere que três agrupamentos se associam predominantemente a contextos avaliativos positivos, especialmente aqueles relacionados ao bem-estar animal e à produção, enquanto "animal husbandry" aparece associado a impactos ambientais negativos nos exemplos discutidos. Os resultados ilustram como a análise baseada em corpus pode identificar padrões lexicais e discursivos relevantes para o ensino de língua acadêmica e oferecer subsídios para materiais que promovam o engajamento crítico com questões de sustentabilidade nas Ciências Agrárias.*

***Palavras-chave**: ecolinguística; linguística de corpus; inglês para fins acadêmicos.*

1. Introdução

As Ciências Agrárias constituem uma importante área de pesquisa no Brasil, um país que desempenha um papel fundamental diante do rápido crescimento da população mundial, projetada para atingir 8 bilhões de pessoas até 2030 (United Nations, 2022). Como os pesquisadores das Ciências Agrárias são incentivados a publicar em inglês para promover a internacionalização de suas pesquisas, os cursos de Inglês para Fins Acadêmicos (IFA)[54] devem incorporar uma perspectiva ecolinguística em um ecossistema de produção de conhecimento complexo, interdisciplinar e de grande impacto (Poole, 2022).

Sugerimos a perspectiva da Ecolinguística (Couto, 2009; do Couto, 2013; Fill, 2015; Poole, 2022), uma disciplina emergente, que explora o papel da linguagem nas interações que sustentam a vida, incluindo a mediação das percepções ambientais e das questões ecológicas. As pesquisas em Ecolinguística têm analisado os discursos sobre a mudança climática, representações de animais e o tratamento discursivo do meio ambiente em diversos contextos, como textos políticos, religiosos e corporativos. Esta área busca identificar como o uso da linguagem pode normalizar e perpetuar atitudes e ações que contribuem para a crise climática e a degradação ecológica, enquanto promove o uso de discursos que favorecem a justiça ambiental, o bem-estar e a sustentabilidade.

A integração da Ecolinguística na Educação pode proporcionar uma abordagem inovadora para ensinar e sensibilizar os alunos sobre questões ambientais. Utilizar textos orais e escritos para disseminar conteúdos ecolinguísticos pode não apenas melhorar a competência linguística dos alunos, mas também conscientizá-los sobre a importância da linguagem na construção de atitudes sustentáveis.

Neste estudo, trabalharemos os conceitos da prosódia semântica para a análise de termos e agrupamentos lexicais que sejam utilizados por autores de textos científicos. De acordo com Partington (1998, p. 66) a prosódia semântica está vinculada ao padrão formado entre itens lexicais e a conotação produzida no texto, que pode ser positiva, negativa ou neutra. De acordo com o autor, a conotação não se encontra apenas em um item, e sim em sua associação com os demais encontrados ao seu redor (colocações[55]). Na Linguística de *Corpus*, Berber Sardinha (2004) define este fenômeno da seguinte forma:

> A prosódia semântica, assim como a colocação, cria uma relação de expectativa para com o ouvinte ou leitor. Na tradução, a quebra de padrões entre uma língua e outra pode trazer implicações relacionadas à fidedignidade, aceitabilidade e legibilidade do texto traduzido ou vertido. A quebra da prosódia semântica, por sua vez, implica na mudança da conotação intencionada pelo falante ou escritor (Berber Sardinha, 2004, p.238).

No estudo de Lopes (2011), a autora analisa um *corpus* de textos jornalísticos e destaca a importância de estudos bilíngues que descrevem padrões lexicogramaticais, como a

[54] Em inglês: *English for Academic Purposes – EAP*

[55] A colocação é definida por Sinclair (1991, p. 170) como sendo "a ocorrência de duas ou mais palavras distantes num pequeno espaço de texto entre si. ([...] *the ocurrence of two or more words within a short space of each other in a text*").

prosódia semântica, uma vez que seus resultados demonstram que tradutores podem utilizar colocações que não imprimem a prosódia semântica correta no texto traduzido.

Este trabalho investiga o uso do termo "animal" como modificador de outros substantivos em artigos acadêmicos das Ciências Agrárias. A descrição dos contextos de ocorrência dos agrupamentos lexicais permite observar padrões de uso e discutir sua relevância para o ensino de inglês acadêmico sob uma perspectiva ecolinguística.

O estudo utiliza o Corpus de Artigos Científicos das Ciências Agrárias (*CorAgrarian*) para identificar agrupamentos lexicais em textos acadêmicos em língua inglesa. O foco deste artigo está na análise dos padrões selecionados; a elaboração e a aplicação de atividades didáticas constituem desdobramentos posteriores da pesquisa.

Assim, os objetivos desta pesquisa são:

1. Identificar, no *CorAgrarian*, os agrupamentos lexicais estatisticamente significativos para as Ciências Agrárias com o termo "animal";
2. Analisar a prosódia semântica presente nos contextos dos agrupamentos lexicais observados, sendo "animal" o modificador de outros substantivos;
3. Discutir as possíveis contribuições dos padrões lexicogramaticais identificados para a elaboração de materiais didáticos de Inglês para Fins Acadêmicos fundamentados em princípios ecolinguísticos.

Para tanto, consideramos os estudos sobre a prosódia semântica na Linguística de *Corpus* (Partington, 1998; Berber Sardinha, 2004; Lopes, 2011) e sobre a Ecolinguística (Couto, 2009; do Couto, 2013; Fill, 2015).

Este estudo faz parte de duas pesquisas mais amplas voltadas para a produção escrita e oral de textos acadêmicos e no reconhecimento de termos e contextos que envolvem os objetivos de desenvolvimento sustentável (ODS) das Nações Unidas.

2. Procedimentos metodológicos

O *CorAgrarian* é um *corpus* de artigos científicos das Ciências Agrárias composto por 447 artigos, compilados de periódicos internacionais indicados por pesquisadores de uma das instituições envolvidas, totalizando 2.502.932 ocorrências (Teixeira et al., 2022; Marques, 2024). Para este estudo, foi utilizado um subcorpus de 1.073 textos, com 974.872 ocorrências. Esse subcorpus foi comparado a um *corpus* de referência composto por 10.440 arquivos e 4.880.115 ocorrências, constituído por artigos da revista PLOS ONE e compilado com o AntCorGen (Anthony, 2021). O *corpus* de referência é aproximadamente cinco vezes maior que o subcorpus de estudo, conforme a orientação discutida por Berber Sardinha (2004).

Os *corpora* foram processados no AntConc (Anthony, 2024). Inicialmente, foi gerada uma lista de palavras-chave para identificar itens estatisticamente relevantes no subcorpus em comparação com o corpus de referência. O item "animal" foi selecionado para análise por sua relevância entre as palavras-chave identificadas. Em seguida, foram examinados os agrupamentos lexicais associados a esse nódulo por meio das ferramentas *Cluster*, *Rank*, *Frequency* e *Range*.

A partir do levantamento, foram selecionados quatro agrupamentos para análise: “animal care”, “animal production”, “animal husbandry” e “animal protein”. Para cada agrupamento, foram consultadas linhas de concordância em contexto expandido. A análise qualitativa considerou os contextos de ocorrência e as associações avaliativas observadas, classificando a prosódia semântica como positiva, negativa ou neutra.

Os agrupamentos e seus contextos poderão ser utilizados, em etapas subsequentes da pesquisa, em disciplinas de graduação e pós-graduação de IFA, com vistas à elaboração de atividades de leitura e escrita acadêmica que permitam aos estudantes examinar como temas das Ciências Agrárias são discutidos em artigos científicos.

3. Resultados preliminares

A análise de palavras-chave identificou o termo “*animal*” como central, dada a sua relevância para a produção sustentável, segurança alimentar e a conservação ambiental. O levantamento dos agrupamentos lexicais, realizado com o software *AntConc* (Anthony, 2024), permitiu o levantamento das combinatórias junto ao nódulo selecionado conforme evidenciado na nuvem de palavras da Figura 1:



Figura 1. Nuvem de palavras com as combinatórias junto ao nódulo “*animal*” no *CorAgrarian*

Uma vez levantados os agrupamentos lexicais com a palavra “animal” no *CorAgrarian*, selecionamos quatro termos para serem discutidos neste trabalho, três com uma prosódia semântica positiva e um com a prosódia semântica negativa.

Tabela 1. Informações e ranqueamento de agrupamentos lexicais com o termo “animal”

Classificação	Agrupamento lexical	Frequência	Distribuição
1.	Animal care	27	10
2.	Animal production	16	10
3.	Animal husbandry	10	10
4.	Animal protein	10	9

Para que os contextos junto aos agrupamentos lexicais fossem analisados, realizamos a leitura das linhas de concordância com o intuito de verificar se a prosódia semântica era positiva, negativa ou neutra.

3.1 Animal care

O primeiro termo observado foi “animal care”, conforme apresentado na Figura 2.

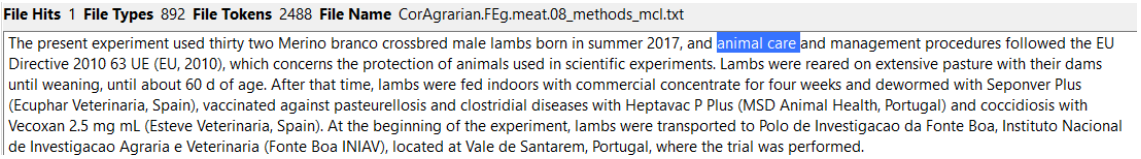


Figure 2. Excerto com o termo “animal care”

“Animal care” pode ser traduzido como “cuidado animal” e, no manuscrito de origem, é relacionado ao “Princípio das Cinco Liberdades”, formulado em 1965, que orienta discussões sobre cuidado e bem-estar animal (Elischer, 2019). Nos contextos examinados, a expressão aparece principalmente em referências a comitês de ética de universidades e ao programa canadense “proAction Animal Care”, voltado ao bem-estar animal. Outros exemplos a relacionam a procedimentos de criação que priorizam a proteção dos animais. Com base nesses contextos, a expressão foi interpretada como apresentando prosódia semântica predominantemente positiva.

3.2 Animal production

O segundo termo analisado foi “animal production” conforme apresentado na Figura 3:

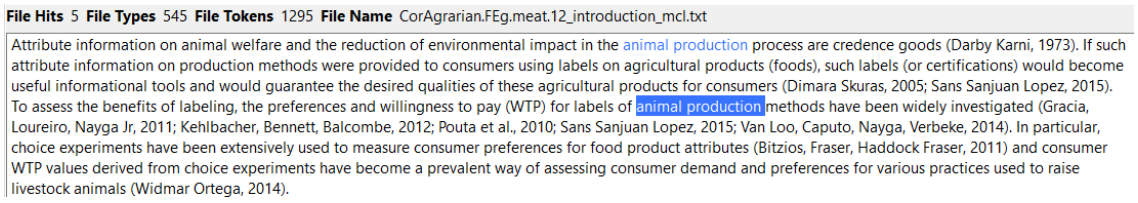


Figura 3. Excerto com o termo “animal production”

“Animal production” aparece associado, nos contextos analisados, a demandas por maior sustentabilidade e bem-estar nos processos de produção animal. A expressão se refere à atividade agropecuária de criação de animais para obtenção de carne, matérias-primas e outros produtos ou serviços. Nos exemplos descritos, as associações com sustentabilidade e bem-estar sustentam uma interpretação de prosódia semântica predominantemente positiva. Essa classificação deve ser entendida como relativa aos contextos examinados, e não necessariamente ao uso da expressão em todos os textos do *corpus*.

3.3 Animal husbandry

O terceiro termo apresentado na Figura 4 foi “animal husbandry”, como vemos a seguir:

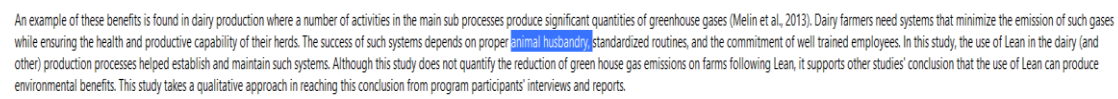
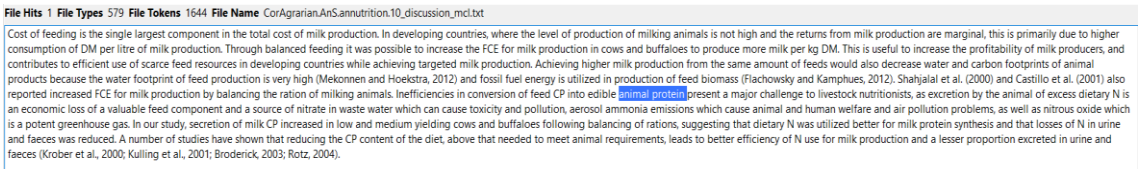


Figura 4. Excerto com o termo “animal husbandry”

"Animal husbandry", traduzido no manuscrito como "pecuária" ou criação animal, foi interpretado como apresentando prosódia semântica negativa nos contextos analisados. Essa avaliação decorre de sua associação, nos exemplos descritos, a práticas de queima, emissão de gases de efeito estufa e apropriação de áreas de vegetação nativa. Assim, os contextos apontam para uma representação da atividade vinculada a impactos ambientais.

3.4 Animal protein

Ao analisarmos o termo "animal protein" em seu contexto, é possível observar que ele se apresenta com uma prosódia semântica positiva, conforme visualizado na Figuras 5, à seguir:


File Hits 1 File Types 579 File Tokens 1644 File Name CorAgrarian.AnS.annutrition.10_discussion_mcl.txt

Cost of feeding is the single largest component in the total cost of milk production. In developing countries, where the level of production of milking animals is not high and the returns from milk production are marginal, this is primarily due to higher consumption of DM per litre of milk production. Through balanced feeding it was possible to increase the FCE for milk production in cows and buffaloes to produce more milk per kg DM. This is useful to increase the profitability of milk producers, and contributes to efficient use of scarce feed resources in developing countries while achieving targeted milk production. Achieving higher milk production from the same amount of feeds would also decrease water and carbon footprints of animal products because the water footprint of feed production is very high (Mekonnen and Hoekstra, 2012) and fossil fuel energy is utilized in production of feed biomass (Flachowsky and Kamphues, 2012). Shahjalal et al. (2000) and Castillo et al. (2001) also reported increased FCE for milk production by balancing the ration of milking animals. Inefficiencies in conversion of feed CP into edible animal protein present a major challenge to livestock nutritionists, as excretion by the animal of excess dietary N is an economic loss of a valuable feed component and a source of nitrate in waste water which can cause toxicity and pollution, aerosol ammonia emissions which cause animal and human welfare and air pollution problems, as well as nitrous oxide which is a potent greenhouse gas. In our study, secretion of milk CP increased in low and medium yielding cows and buffaloes following balancing of rations, suggesting that dietary N was utilized better for milk protein synthesis and that losses of N in urine and faeces was reduced. A number of studies have shown that reducing the CP content of the diet, above that needed to meet animal requirements, leads to better efficiency of N use for milk production and a lesser proportion excreted in urine and faeces (Krober et al., 2000; Kulling et al., 2001; Broderick, 2003; Rotz, 2004).

Figura 5. Excerto com o termo "animal protein"

Nos contextos apresentados, "animal protein" aparece associado à alimentação, ao valor de mercado e aos desafios envolvidos em sua produção. A expressão é discutida em relação à demanda de mercado e aos cuidados requeridos na criação dos animais. Com base nesses exemplos, a prosódia semântica foi interpretada como positiva. Entretanto, a valência avaliativa da expressão deve ser delimitada aos contextos efetivamente analisados, uma vez que referências ao valor de mercado ou à demanda não implicam, por si só, uma avaliação positiva.

3.5 Síntese dos resultados

Considerando as interpretações apresentadas para os quatro agrupamentos, três foram classificados como predominantemente positivos ("animal care", "animal production" e "animal protein"), enquanto um foi interpretado como negativo ("animal husbandry").

A identificação desses padrões pode contribuir para a compreensão de como textos científicos das Ciências Agrárias lexicalizam questões relacionadas à produção, ao bem-estar animal e aos impactos ambientais. Do ponto de vista ecolinguístico e pedagógico, os agrupamentos oferecem exemplos autênticos que podem ser explorados em atividades de leitura crítica e análise lexical em IFA.

4. Considerações Finais

A análise preliminar dos agrupamentos lexicais associados ao termo "animal" no *CorAgrarian* indicou predominância de contextos interpretados como positivos, especialmente aqueles relacionados ao cuidado e ao bem-estar animal, à produção e à alimentação. Também foi identificado um agrupamento cujos contextos foram associados a impactos ambientais negativos. Os resultados evidenciam, portanto, que as expressões examinadas participam de diferentes formas de representar a produção animal e suas relações com questões socioambientais.

A investigação demonstra o potencial da Linguística de *Corpus* para identificar padrões lexicogramaticais em textos acadêmicos e para subsidiar a seleção de exemplos autênticos destinados ao ensino de Inglês para Fins Acadêmicos. A articulação com a Ecolinguística pode contribuir para que atividades didáticas explorem não apenas aspectos linguísticos, mas também as representações discursivas de temas relacionados à sustentabilidade.

Como desdobramento, pretende-se aprofundar a análise dos contextos de ocorrência dos agrupamentos, incluindo a quantificação das categorias avaliativas quando os dados estiverem disponíveis, e desenvolver atividades de leitura e escrita acadêmica para estudantes de graduação e pós-graduação. Essas etapas futuras poderão avaliar de que modo o trabalho com padrões identificados em corpus contribui para a aprendizagem linguística e para o engajamento crítico com questões socioambientais.

Agradecimentos

As autoras agradecem o apoio da FAPEMIG (Proc. APQ-01173-22), do CNPq (Proc. 308195/2025-6) e da FAPESP (Proc. 2022/05848-7).

Referências

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Anthony, L. (2021). AntCorGen (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntCorGen

Berber Sardinha, T. (2004) Linguística de Corpus São Paulo: Editora Manole. 2004.

Couto, Hildo Honório do. (2009) Ecolinguística. Cadernos de Linguagem e Sociedade, Brasília, v. 10, n. 1, p. 125-149, 2009.

Do Couto,Hildo Honório. (2013) O que vem a ser ecolinguística, afinal?. Cadernos de linguagem e sociedade, v. 14, n. 1, p. 275-312, 2013.

Elischer, Melissa. (2019) *The Five Freedoms*: A history lesson in animal care and welfare. MSU Extension. Michigan, 2019.

Fill, Alwin. (2015) Ecolinguística: a história de uma ideia verde para o estudo da linguagem. Ecolinguística: *Revista brasileira de ecologia e linguagem* (ECO-REBEL), v. 1, n. 1, p. 07-21.

Lopes, Maria Cecília. (2011) Tradução, padrões e nuances: um estudo de Linguística de Corpus sobre diferentes prosódias semânticas na língua fonte e na língua alvo. *Revista Brasileira de Linguística Aplicada*, v. 11, p. 747-772, 2011.

Marques, C. G. de F. CorAgrarian: um corpus de artigos científicos de ciências agrárias. *In:* SEMANA DE EVENTOS DA FACULDADE DE LETRAS (SEvFALE), XV, 2024, Universidade Federal de Minas Gerais. *Apresentação oral.* Belo Horizonte: 2024.

Partington, A. (1998) *Patterns and Meanings* - Using Corpora for English Language Research and Teaching. Amsterdam/Philadelphia: John Benjamins. 1998.

Poole, R. (2022). *Corpus-assisted ecolinguistics*. Bloomsbury Academic.

Sinclair J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Teixeira, G.; Bocorny, A. E. P.; Dutra, D. P. Mapping the language needs of the Brazilian researchers in the Agrarian Sciences: a corpus-based study of abstracts. In: AMERICAN ASSOCIATION FOR CORPUS LINGUISTICS CONFERENCE (AACL), 15, 2022, Flagstaff. Anais [...]. Flagstaff: Northern Arizona University, 2022.

The Sustainable Development Goals. Disponível em <https://sdgs.un.org/goals> Acesso em: 23 mai. 2021.

OFICINAS
EBRALC 2026

A. Introdução aos Grandes Modelos de Linguagem: história, funcionamento e desafios da Inteligência Artificial contemporânea

Livy Real[1,3], Cibele Simões de Oliveira Santos[1,2]

[1]Instituto Kunumi, [2]PPGCC — Universidade Federal de Minas Gerais (UFMG),
[4]Instituto de Computação — Universidade Federal do Amazonas (UFAM)

livyreal@gmail.com, cibelesimoesoliveira@gmail.com

Palavras-chave: Inteligência Artificial; Transformers; Processamento de Linguagem Natural.

1. Ementa

Este curso é introdutório, de fácil acesso e teórico. Nosso objetivo é empoderar a comunidade de linguística com conhecimento técnico sobre o funcionamento de Grandes Modelos de Linguagem (LLMs). Apresentaremos uma introdução histórica à Inteligência Artificial, com ênfase na evolução dos sistemas de Processamento de Linguagem Natural até o surgimento dos modelos baseados em Transformers e LLMs. Partindo dos primeiros debates sobre IA, discutiremos as mudanças trazidas pelo aprendizado de máquina, pelas redes neurais profundas e pela disponibilidade massiva de dados para a área. Serão apresentados os principais componentes de uma LLM relacionando o funcionamento dos modelos aos níveis de análise linguística. O curso também abordará limitações desses modelos, incluindo vieses, alucinações, explicabilidade, alinhamento e desafios éticos e políticos relacionados ao uso da IA contemporânea.

2. Objetivos

Apresentar uma visão histórica da evolução da IA aplicada à linguagem.
Explicar os princípios básicos de funcionamento dos modelos Transformer.
Capacitar participantes a compreender criticamente o funcionamento e os limites das LLMs.
Promover o diálogo entre computação e linguística.

3. Conteúdos abordados

- História da IA e do Processamento de Linguagem Natural.
- Sistemas simbólicos e baseados em regras.
- Aprendizado de máquina e representações vetoriais.
- Redes neurais profundas e Big Data.
- Arquitetura Transformer e mecanismos de atenção.
- Funcionamento das LLMs.
- Vieses, alucinações, explicabilidade e questões éticas.

4. Metodologias e Ferramentas

Aulas expositivas, análise de exemplos práticos, demonstrações de ferramentas e discussão estudos de caso.

5. Público-alvo estimado

Estudantes, pesquisadores e profissionais das áreas de linguística, letras, ciências humanas e áreas afins interessados em compreender os fundamentos da IA contemporânea. Não é necessário ter conhecimento de programação.

B. O uso de corpora de referência por profissionais de Letras para a aprendizagem de língua inglesa

Rachel Lourenço, Thiago Blanch Pires

Instituto de Letras – Universidade de Brasília

rachellourenco@unb.br, thiagocomaga@gmail.com

Palavras-chave: corpora de referência, aprendizagem de língua inglesa, estudo de léxico

1. Ementa

Este trabalho propõe uma aplicação prática dos corpora COCA (Corpus of Contemporary American English) e BNC (BriӨsh NaӨonal Corpus), por meio do acesso à plataforma english-corpora.org, ao estudo de língua inglesa por parte de profissionais em formação e/ou pesquisadores da área de Letras. A primeira parte consistirá de discussões teórico-metodológicas e a segunda parte será composta por exercícios orientados por corpora (corpus-driven) e de exercícios baseados em corpora (corpus-based) aplicados ao estudo do léxico da língua inglesa.

O uso de corpora de referência como ferramenta de aprendizagem tem como objetivo desenvolver estratégias de aprendizagem do idioma de modo a possibilitar mais autonomia no estudo da língua inglesa para profissionais que desempenham tarefas que exigem nível avançado do idioma, tais como: ensino, pesquisa, tradução, legendagem, interpretação, revisão, curadoria linguística.

2. Objetivos

Este trabalho, de nível introdutório, tem como objetivo familiarizar os participantes com as funcionalidades oferecidas pelos corpora COCA e BNC para usá-los como ferramentas de aprendizagem do léxico da língua inglesa.

3. Conteúdos abordados

Além de explorar as funcionalidades dos corpora COCA e BNC, os participantes farão análises de falsos cognatos e de sinônimos parciais na língua inglesa.

4. Metodologias e Ferramentas

Com duração total de 4 (quatro) horas, o trabalho propõe que os participantes analisem falsos cognatos e sinônimos parciais ao explorarem as seguintes funcionalidades dos corpora de referência, entre outras: linhas de concordância, colocados, aglomerados (clusters), gêneros/registros textuais, frequência de palavras.

5. Público-alvo estimado

O trabalho tem como público-alvo professores da rede pública, estudantes de graduação

e pós-graduação da área de Letras que tenham como interesse ou objetivo o uso de corpora de referência no ensino-aprendizagem da língua inglesa. Não é necessário ter conhecimento prévio da ferramenta, mas é desejável ter nível (pelo menos) intermediário do idioma.

C. Anotação morfossintática de corpora de línguas indígenas: da Plataforma Tycho Brahe ao modelo Universal Dependencies

Leonel F. de Alencar[1,2], Maria Filomena S. Sandalo[1], Charlotte Marie C. Galves[1]

[1]Instituto de Estudos da Linguagem – Universidade Estadual de Campinas (Unicamp)

Campinas – SP – Brasil [2]Centro de Humanidades – Universidade Federal do Ceará

Fortaleza – CE – Brasil

leonel.de.alencar@ufc, {sandalo,galvesc}@unicamp.br

Palavras-chave: Línguas indígenas brasileiras; Universal Dependencies; parsing de constituintes.

1. Ementa

O minicurso introduz fundamentos para construção de corpora morfossintaticamente anotados de línguas com baixo suporte digital por meio da Plataforma Tycho Brahe e do modelo Universal Dependencies (UD). Com base em exemplos da língua indígena kadiwéu, falada no Mato Grosso do Sul, serão demonstradas funcionalidades da plataforma para o parsing de constituintes, análise morfológica, tratamento de tokens multivocabulares e manutenção de um léxico computacional.

Em seguida, serão apresentados scripts em Python para inspeção e conversão, para o formato CoNLL-U adotado por UD, de um corpus do kadiwéu da Plataforma Tycho Brahe no estilo do Penn Treebank. Finalmente, será explicado o uso de scripts para comparação de diferentes versões de um treebank, aperfeiçoamento do conversor automático com base num padrão-ouro bem como validação e avaliação de treebanks no modelo UD.

2. Objetivos

- Apresentar funcionalidades da Plataforma Tycho Brahe para construção de um corpus de uma língua com baixo suporte digital.
- Demonstrar o parsing de constituintes, a análise vocabular e a manutenção de um léxico computacional.
- Introduzir o modelo Universal Dependencies e o formato CoNLL-U.
- Apresentar scripts em Python para conversão automática de um corpus do kadiwéu anotado no estilo do Penn Treebank para o modelo UD.
- Demonstrar scripts para aperfeiçoamento iterativo do conversor a partir de um padrão-ouro, bem como para validação, avaliação e comparação de diferentes versões de um treebank.

3. Conteúdos abordados

- Suporte digital das línguas indígenas brasileiras

- Plataforma Tycho Brahe
- Parsing de constituintes
- Segmentação morfológica
- Tratamento de tokens multivocabulares
- Léxico computacional
- Modelo Universal Dependencies
- Formato CoNLL-U
- Edição de CoNLL-U
- Scripts em Python para conversão automática em UD de anotações no estilo do
- Penn Treebank
- Comparação de diferentes versões de um treebank com um padrão-ouro
- Validação e avaliação de treebanks no modelo UD

4. Metodologias e Ferramentas

Serão combinadas exposição teórica, demonstrações práticas e atividades guiadas. Serão utilizados a Plataforma Tycho Brahe, scripts desenvolvidos em Python para conversão e comparação de treebanks, bem como ferramentas oficiais do ecossistema Universal Dependencies para validação e avaliação de treebanks.

5. Público-alvo estimado

Estudantes de graduação e pós-graduação, docentes e pesquisadores das áreas de linguística de corpus, linguística computacional e processamento de linguagem natural. Não se exigem conhecimentos prévios sobre a Plataforma Tycho Brahe ou Universal Dependencies.

D. Como os Grandes Modelos de Linguagem representam a linguagem? Uma introdução aos Transformers para linguistas com Orange Data Mining

Marcos Moisés Crisóstomo de Oliveira

Instituto Federal da Bahia (IFBA)

Palavras-chave: Grandes Modelos de Linguagem; Arquitetura Transformer; Processamento de Linguagem Natural

1. Ementa

Os Grandes Modelos de Linguagem (Large Language Models – LLMs) transformaram profundamente o Processamento de Linguagem Natural (PLN) e passaram a integrar o cotidiano da pesquisa linguística. Entretanto, seus fundamentos permanecem pouco conhecidos por pesquisadores das Ciências da Linguagem, que frequentemente utilizam essas tecnologias sem compreender como representam palavras, contexto e relações de sentido. Este minicurso propõe uma introdução conceitual e prática aos fundamentos da arquitetura Transformer, tendo como questão norteadora: como os Grandes Modelos de Linguagem representam a linguagem e quais são as possibilidades e os limites dessa representação para a pesquisa linguística?

Na primeira parte serão apresentados, em linguagem acessível e sem conhecimentos prévios de programação, os conceitos de tokenização, embeddings, mecanismos de atenção (attention e self-attention), representações contextuais e funcionamento geral de modelos como BERT, GPT e Llama. Também serão discutidas aproximações entre essas representações computacionais e categorias tradicionalmente utilizadas pelos estudos linguísticos e discursivos, evidenciando convergências, diferenças e limites epistemológicos.

Na segunda parte, os participantes utilizarão o Orange Data Mining para experimentar técnicas contemporâneas de mineração textual sem necessidade de programação. Serão construídos fluxos envolvendo preparação de corpus, geração de embeddings, análise de similaridade semântica, agrupamento automático de documentos, classificação textual e busca semântica. Além das atividades práticas, serão apresentados exemplos reais de aplicações baseadas em Transformers, como organização automática de grandes corpora, classificação de documentos, recuperação semântica de informações e detecção automática de desinformação.

Ao final do minicurso, espera-se que os participantes compreendam como funcionam os Grandes Modelos de Linguagem e sejam capazes de utilizá-los de forma crítica e fundamentada em pesquisas linguísticas.

2. Objetivos

Geral: Introduzir pesquisadores das Ciências da Linguagem aos fundamentos conceituais e práticos dos Grandes Modelos de Linguagem, promovendo uma compreensão crítica da arquitetura Transformer e de suas aplicações em pesquisas linguísticas por meio do

Orange Data Mining.

Específicos: Ao final do minicurso o participante deverá ser capaz de

- compreender a evolução dos modelos computacionais de representação da linguagem;
- compreender o funcionamento geral da arquitetura Transformer;
- explicar os mecanismos de atenção e autoatenção;
- compreender o conceito de embeddings e representações contextuais;
- interpretar como os LLMs representam palavras e textos;
- construir fluxos básicos de mineração textual utilizando o Orange Data Mining;
- explorar aplicações contemporâneas dos Grandes Modelos de Linguagem;
- avaliar criticamente potencialidades e limitações da Inteligência Artificial aplicada às Ciências da Linguagem.

3. Conteúdos abordados:

Primeiro encontro

Módulo 1 - Por que surgiram os Grandes Modelos de Linguagem?

- Breve história do PLN
- Evolução dos modelos de representação textual
- O surgimento dos Transformers

Módulo 2 - Como um Transformer representa a linguagem

- Tokens
- Embeddings
- Positional Encoding
- Attention
- Self-Attention
- Representações contextuais
- BERT
- GPT
- Llama

Módulo 3 - LLMs e pesquisa linguística

- O que os modelos realmente aprendem
- Contexto linguístico e contexto estatístico
- Potencialidades
- Limitações
- Aspectos éticos

Segundo encontro – Oficina prática

Construção de um fluxo completo utilizando Orange Data Mining. Os participantes irão:

- importar um corpus;
- preparar os dados;

- gerar embeddings;
- explorar similaridade semântica;
- realizar clusterização automática;
- visualizar espaços vetoriais;
- classificar textos;
- realizar buscas semânticas.

Demonstrações: Serão apresentados protótipos desenvolvidos pelo ministrante ilustrando aplicações dos Grandes Modelos de Linguagem, incluindo:

- organização automática de grandes corpora;
- classificação automática de documentos;
- busca semântica;
- detecção automática de desinformação (fake news);
- comparação semântica entre textos.

Encerramento: Mesa de discussão "Como incorporar os Grandes Modelos de Linguagem à pesquisa em Linguística?"

4. Metodologia e Ferramentas

O minicurso será realizado em dois encontros presenciais de quatro horas, totalizando oito horas. A metodologia combinará exposição dialogada, demonstrações práticas e atividades orientadas.

Os conceitos computacionais serão apresentados utilizando linguagem acessível, analogias e exemplos linguísticos, sem utilização de fórmulas matemáticas e sem necessidade de conhecimentos prévios de programação.

Durante a oficina prática, cada participante desenvolverá em seu próprio computador um fluxo completo de mineração textual utilizando o Orange Data Mining, permitindo experimentar diferentes aplicações dos Grandes Modelos de Linguagem em corpora reais. Ao longo do minicurso serão discutidas as potencialidades e limitações dessas tecnologias para pesquisas em Linguística, incentivando uma postura crítica diante das ferramentas de Inteligência Artificial.

Ferramentas:

- Orange Data Mining
- Orange Text Mining
- Modelos Transformer pré-treinados
- Corpus em língua portuguesa
- Projetor multimídia

5. Público-alvo

Pesquisadores, professores e estudantes de graduação e pós-graduação em Letras, Linguística, Linguística de Corpus, Linguística Computacional, Humanidades Digitais e áreas afins. O minicurso foi planejado especialmente para iniciantes, não sendo exigidos conhecimentos prévios em programação, Inteligência Artificial ou Processamento de Linguagem Natural.

E. Linguística de Corpus, modelos de linguagem e Inteligência Artificial: técnicas e ferramentas para simplificar bulas de medicamentos em português do Brasil

Maria José Bocorny Finatto[1], Bianca Franco Pasqualini[1]

[1] Programa de Pós-Graduação em Letras – Universidade Federal do Rio Grande do Sul – Porto Alegre – RS – Brasil

Palavras-chave: Linguística Computacional; Modelos de Linguagem; Simplificação Textual

1. Ementa

A oficina aborda a interseção entre Linguística de Corpus (LCorp), Linguística Computacional e Inteligência Artificial (IA) no contexto da Acessibilidade Textual e Terminológica (ATT) e letramento em saúde. Discutem-se Large Language Models (LLMs) como ferramentas de auxílio à simplificação de bulas de medicamentos, sob a ótica da Tradução Intralinguística (TI).

Serão desenvolvidos exercícios de simplificação manual e automática, com o apoio de ferramentas digitais e LLMs, e validados por linguistas com base em insumos da LCorp.

2. Objetivos

Apresentar recursos de Linguística Computacional para apoio à simplificação de textos em saúde; capacitar para a identificação de barreiras terminológicas e sintáticas em bulas; e debater o potencial e os limites do uso de LLMs na reescrita de textos complexos.

3. Conteúdos abordados

Recursos e ferramentas de linguística computacional; fundamentos de Tradução Intralinguística e Acessibilidade Textual e Terminológica; o que são e como funcionam LLMs; estratégias de simplificação e marcos regulatórios da Linguagem Simples (Lei 15.263/2025).

4. Metodologias e Ferramentas

A oficina terá dois encontros presenciais de 2h. Combinará exposição dialogada sobre ferramentas digitais (como CorPop e Vidya) e o funcionamento de LLMs, com exercícios de análise e reescrita de bulas da Anvisa e cartilhas de saúde.

5. Público-alvo estimado

Estudantes de pós-graduação e tradutores com base em Linguística de Corpus ual.

Referência

SCALERCIO, Arthur; BERTOTTO, Eduarda; JESUS, Silvana; FINATTO, Maria José; PAES, Aline. (2026) Annotation Guidelines and Challenges for Automatic Simplification of Portuguese Drug Leaflets. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)* - Vol. 2, p. 121–127, Salvador, Brazil. Association for Computational Linguistics.

F. Introdução à Mineração de Texto com Orange Data Mining: fundamentos e aplicações práticas

Crysttian Paixão, Eduardo Grizenti

Universidade Federal da Bahia

Palavras-chave: Mineração de texto; Aprendizado de máquina; Análise de dados

1. Ementa

O curso apresenta os fundamentos da mineração de texto e do aprendizado de máquina com foco no software Orange Data Mining. Aborda a transformação de textos em dados analisáveis, explorando técnicas de análise exploratória, pré-processamento textual e visualização interativa. Inclui atividades práticas com widgets para construção de workflows de análise, permitindo ao participante realizar tarefas como tokenização, filtragem, criação de n-gramas, bag of words e introdução à modelagem de tópicos. Ao final, são apresentados conceitos iniciais de clusterização, redução de dimensionalidade e classificação.

2. Objetivos

Introduzir os participantes aos princípios da análise de texto automatizada; capacitar para uso básico do Orange; demonstrar como transformar corpus textual em dados estruturados; desenvolver habilidades iniciais de análise exploratória e modelagem simples.

3. Conteúdos abordados

Conceito de texto como dado; análise quantitativa de texto; interface e lógica de funcionamento do Orange (widgets, workflow, canvas); importação e visualização de corpus; pré-processamento (tokenização, normalização, filtragem); n-gramas; *bag of words*; introdução à modelagem de tópicos; noções de clusterização e redução de dimensionalidade.

4. Metodologias e Ferramentas

Aula expositiva dialogada (2h) combinada com oficina prática (2h) em laboratório. Uso do software Orange Data Mining com exercícios guiados, construção de fluxos analíticos e exploração de corpora de exemplo.

5. Público-alvo estimado

Estudantes de graduação e pós-graduação e pesquisadores das áreas de ciências sociais, comunicação e áreas afins, interessados em análise de dados textuais sem necessidade de conhecimento prévio em programação.



SALVADOR - BAHIA - BRASIL

9 A 12 DE NOVEMBRO DE 2026