ebralc

XVI ENCONTRO DE LINGUÍSTICA DE CORPUS
XIII ESCOLA BRASILEIRA DE LINGUÍSTICA COMPUTACIONAL
21 A 24 DE OUTUBRO, 2024 - UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA

ANAIS ELETRÔNICOS DO XVI ENCONTRO DE LINGUÍSTICA DE CORPUS E DA XII ESCOLA BRASILEIRA DE LINGUÍSTICA COMPUTACIONAL

BRASÍLIA-DF

2024

EXPEDIENTE

Comissão organizadora

Elisa Duarte Teixeira (Presidenta - UnB)

Andréa Geroldo dos Santos (IBFE-SP)

Célia Maria Magahães (UFMG / UnB)

Cláudio Corrêa e Castro Gonçalves (UnB)

Elaine Alves Trindade (PUC-SP)

Flávia Cristina Cruz Lamberti Arraes (UnB)

Joacyr Tupinambás (Unicamp)

Nilson Roberto Barros da Silva (UERN)

Patrícia Tuxi dos Santos (UnB)

Rafaela Araújo Jordão Rigaud Peixoto (DECEA-FAB)

Rodrigo Garcia Rosa (USP)

Rozane Rodrigues Rebechi (UFRGS)

Stella Esther Ortweiler Tagnin (USP)

Thiago Blanch Pires (UnB)

Vander Paula Viana (U. Edinburgh)

Comissão científica

Adriana Zavaglia (USP)

Adriane Orenha Ottaiano (UNESP)

Alessandra Matias Querido (UnB)

Ana Eliza Pereira Bocorny (UFRGS)

Andréa Geroldo dos Santos (IBFE-SP)

Angela Maria Tenório Zucchi (USP)

Ariel Novodvorski (UFU)

Camila Höfling (UFSCar)

Celia Maria Magalhães (UFMG)

Claudia Zavaglia (UNESP-IBILCE)

CAPES

POSTRAD

COMPLETT UnB
Corpus Multilíngue para Pesquisas em Línguas Estrangeiras, Tradução e Terminologia da Universidade de Brasília

TermiTraDiCo
Terminologia e Tradução Direcionadas por Corpus

ABraLC
Associação Brasileira de Linguística de Corpus

Cleci Regina Bevilacqua (UFRGS)

Cristiane Krause Kilian (Inst. Sup. de Ed. Ivoti - ISEI)

Deise Prina Dutra (UFMG)

Elaine Alves Trindade (PUC-SP)

Gleiton Malta (UFBA)

Guilherme Fromm (UFU)

Heliana Ribeiro de Mello (UFMG)

Heloísa Orsi Koch Delgado (La Salle / UFRGS)

Igor Antônio Lourenço da Silva (UFU / UFMG)

Joacyr Tupinambás de Oliveira (UNICAMP)

Luciana Carvalho Fonseca (USP)

Luciana Latarini Ginezi (Tradutora / Intérprete)

Luciane Leipnitz (UFPel)

Malila Carvalho de Almeida Prado (BNU-HKBU)

Marcos de Campos Carneiro (UnB)

Maria José Bocorny Finatto (UFRGS)

Nilson Roberto Barros da Silva (UERN)

Patrícia Tosqui Lucks (ICEA)

Paula Tavares Pinto (UNESP-IBILCE)

Rafaela Araújo Jordão Rigaud Peixoto (DECEA / USP)

Renato Rodrigues Pereira (UFMS)

Rodrigo Garcia Rosa (USP)

Rozane Rodrigues Rebechi (UFRGS)

Sandra María Pérez López (UnB)

Shirlene Bemfica de Oliveira (IFMG - Ouro Preto)

Simone Sarmento (UFRGS)

Stella Esther Ortweiler Tagnin (USP)

Thiago Alexandre Salgueiro Pardo (USP)

Vander Viana (University of Edinburgh)

APRESENTAÇÃO

A Linguística de Corpus (LC), ciência que estuda a linguagem por meio da análise de grandes quantidades de textos em formato eletrônico organizados na forma de corpora, avançou muito no mundo e também no Brasil desde os primeiros trabalhos publicados, que foram surgindo com a introdução do computador no ambiente acadêmico, na década de 1960. De lá para cá, é cada vez maior o número de campos do conhecimento que se valem da LC em suas pesquisas, seja como abordagem, seja como metodologia – inclusive para além das áreas de Letras e Linguística.

A Linguística Computacional, ou Processamento de Linguagem Natural (PLN), é um campo de estudos multidisciplinar que aplica a ciência da computação à análise e compreensão da linguagem humana. Os primeiros trabalhos publicados datam da década de 1950. Mas, assim como a LC, o PLN teve um crescimento exponencial nas últimas décadas, à medida que o uso da tecnologia foi evoluindo e se fazendo presente em praticamente todas as esferas da experiência humana.

O Encontro de Linguística de Corpus (ELC) teve sua primeira edição em 1999, na Universidade de São Paulo - em 2024, comemoramos 25 anos de sua criação! A Escola Brasileira de Linguística Computacional (EBRALC) surgiu alguns anos depois, em 2007, à medida que participantes do ELC constataram uma necessidade premente de ampliar os conhecimentos computacionais e tecnológicos de pesquisadora(e)s brasileira(o)s trabalhando com Linguística de Corpus nas áreas de Humanas. Assim, o objetivo da EBRALC é sempre oferecer diversas oficinas práticas, para as quais a(o)s interessada(o)s deverão se inscrever oportunamente. O ELC, por sua vez, é um evento científico que tem por tradição não possuir sessões paralelas, para que toda(o)s possam assistir a todas as comunicações orais, proporcionando um convívio mais profícuo entre a(o)s participantes das várias áreas envolvidas – um convite ao diálogo e à colaboração.

O ELC/EBRALC 2024, cujo tema será “Linguística de Corpus e Inteligência Artificial: interfaces com Letras e Linguística”, acontecerá de 21 a 24 de outubro, na Universidade de Brasília. Organizado por docentes do Instituto de Letras membros dos grupos de pesquisa TermiTraDiCo (Terminologia e Tradução Direcionadas por Corpus) e COMPLETT (Corpus Multilíngue para Pesquisas em Línguas Estrangeiras, Tradução e Terminologia), e colegas de várias outras universidades, o evento tem o apoio do Programa de Pós-Graduação em Tradução da UnB (POSTRAD). Visa fomentar discussões sobre os impactos da Inteligência Artificial na formação acadêmica e na pesquisa nas áreas de Letras e Linguística, com foco no papel que a Linguística de Corpus tem ocupado e pode ocupar, futuramente, no estabelecimento desse diálogo transdisciplinar.

Esperamos, assim, estimular parcerias e promover uma necessária valorização dos estudos da linguagem com o auxílio do computador, no ambiente acadêmico-científico e na sociedade brasileira.

LISTA COM NOME DE AUTORES
Ordem por sobrenome

ALENCAR, Leonel Figueiredo de

BARROS, Cláudia Dias de

BATISTA, Julia de Souza

BOCORNY, Ana Eliza Pereira

BOHORQUEZ, Carolina

FERNANDES BONALUMI, Emiliana

BRAGA, Jasper Vilan

CARDOSO DE CAMARGO, Diva

COSTA, Danilo Duarte

DELGADO, Heloísa Orsi Koch

DURAN, Magali Sanches

FEKETE, João

FONSECA, Luciana Carvalho

FREITAG, Patrícia Helena

FRANGIOTTI, Graziele Altino

FURTADO, Anna Beatriz Dimas

GAMONAL, Maucha Andrade

GIL, Cristina Borges

GUEDES DE SOUZA, Bianca Mara

KAUFFMANN, Carlos Henrique

KILIAN, Cristiane Krause

KUHN, Tanara Zingano

LIU, Ziyang

LOPES, Jhonatan Henrique

LOPES, Mauricio Jose Ferreira

MAIA-PIRES, Flávia de Oliveira

MARQUES, Carolina Godoi de Faria

MARCHESE, Giovana de Castro

MURTA, Lucas Renato dos Santos

NUNES, Wagner da Cunha

TELES OLIVEIRA, Helena Cid

MOREIRA DE OLIVEIRA, Isabela

O'CONNOR, Anne

OLIVEIRA, Simone

OLIVEIRA, Shirlene Bemfica de

ORENHA-OTTAIANO, Adriane

PAGANO, Adriana Silvina

PARDO, Thiago Alexandre Salgueiro

PINTO, Paula Tavares

RABELO, Iasmin Valéria Miranda

RAMOS, Camila Alves

RASO, Tommaso

REBECHI, Rozane Rodrigues

REBECHI, Rozane R.

ROCHA, Bruno Neves Rati de Melo

ROCHA, Ísis Beber de Souza Fiorilo

SANTOS, Amanda Letícia Valadares dos

SARDINHA, Tony Berber

VALVERDE DA SILVA, Júlia Cristina

SILVA, Bruna Rodrigues da

SILVA, Luciano Franco da

TAVARES DA SILVA, Priscila

SOUSA, Luan Daniel dos Santos

SOUSA, Jackson Wilke da Cruz

SYDIO, Ursula Puello

TAGNIN, Stella Esther Ortweiller

TEIXEIRA, Elisa Duarte

TEIXEIRA, Gustavo Leal

TOLEDO, Gabriela Dias

TORRENT, Tiago Timponi

TAMAGNO, Júlia

VALE, Oto Araújo

VICTOR, Ana Clara Taborda de Paula

VIEIRA, Marcelo Augusto

VITAL, Átila Augusto Soares

PIRES, Thiago Blanche

ROCHA, Bruno Nevis Rati de Melo

RASO, Tommaso

WICK-PEDRO, Gabriela

Ordem por prenome

Adriana Silvina PAGANO

Amanda Letícia Valadares dos SANTOS

Ana Clara Taborda de Paula VICTOR

Ana Eliza Pereira BOCORNY

Anne O'CONNOR

Átila Augusto Soares VITAL

Bianca Mara GUEDES DE SOUZA

Bruna Rodrigues da SILVA

Bruno Neves Rati de Melo ROCHA

Carlos Henrique KAUFFMANN

Carolina BOHORQUEZ

Carolina Godoi de Faria MARQUES

Cláudia Dias de BARROS

Cristiane Krause KILIAN

Cristina Borges GIL

Danilo Duarte COSTA

Diva CARDOSO DE CAMARGO

Emiliana FERNANDES BONALUMI

Elisa Duarte TEIXEIRA

Flávia de Oliveira MAIA-PIRES

Giovana de Castro MARCHESE

Gabriela Dias TOLEDO

Gabriela WICK-PEDRO

Graziele Altino FRANGIOTTI

Heloísa Orsi Koch DELGADO

Helena Cid TELES OLIVEIRA

Iasmin Valéria Miranda RABELO

Isabela MOREIRA DE OLIVEIRA

Ísis Beber de Souza Fiorilo ROCHA

Jackson Wilke da Cruz SOUSA

Jhonatan Henrique LOPES

João FEKETE

Julia de Souza BATISTA

Júlia Cristina VALVERDE DA SILVA

Júlia TAMAGNO

Jasper Vilan BRAGA

Luan Daniel dos Santos SOUSA

Luciana Carvalho FONSECA

Luciano Franco da SILVA

Lucas Renato dos Santos MURTA

Mauricio José Ferreira LOPES

Magali Sanches DURAN

Marcelo Augusto VIEIRA

Maucha Andrade GAMONAL

Oto Araújo VALE

Patrícia Helena FREITAG

Paula Tavares PINTO

Rozane Rodrigues REBECHI

Rozane R. REBECHI

Shirlene Bemfica de OLIVEIRA

Simone OLIVEIRA

Stella Esther Ortweiller TAGNIN

Tanara Zingano KUHN

Thiago Alexandre Salgueiro PARDO

Thiago Blanch PIRES

Tiago Timponi TORRENT

Tommaso RASO

Tony Berber SARDINHA

Ursula Puello SYDIO

Wagner da Cunha NUNES

Ziyang LIU

LISTA COM PALAVRAS-CHAVE

Academic writing

Acessibilidade

Acessibilidade comunicacional

Agrarian science corpus

Agrarian sciences

Agrupamentos

Análise multidimensional

Anotação

Anotação de córpus

Antconc

Aposições predicativas

Argumentação

Artificial intelligence

Árvore de domínio

Audiodescrição

Automação com inteligência artificial

Blog de coworking

Brazilian television

Catálogo

Chavicidade

Ciências agrárias

Classificadores semânticos

Colocações

Colocações acadêmicas

Compilação

Compilação de corpus

Computational linguistics

Conceitos

Conceptual domain identification

Contraste

Convencionalidade

Corpora

Corpus

Corpus de aprendizes

Corpus linguistics

Corpus multimodal

Corpus oral

Corpus paralelo

Corpus

Corpus-based metaphor studies

Deficiência

Dicionário de colocações

Dilma Vana Rousseff

Discurso de ódio

Discursos presidenciais

Disfluências

Ditadura militar

Divulgação científica

DIY corpora

English language

Ensino de línguas estrangeiras

Ensino de línguas

Ensino médio técnico

Escolaridade limitada

Escrita acadêmica

Esquizofrenia

Estudos da Tradução

Estudos Descritivos da Tradução

Estudos Feministas da Tradução

Estupro

Expressões idiomáticas

Extração

Extração automática e semiautomática

Extração terminológica

Extratores automáticos e semiautomáticos de terminologia

Fala espontânea

Feminismo

Ferramentas de auxílio à tradução

Formas verbais

Fraseologia

Fraseologia português-inglês

Fraseologismos

Função adversativa

Gênero

Gêneros acadêmicos

Glossário

Glossário bilíngue

Idiomatic expressions

Ilocuções

Inclusão surda

Inglês

Inglês acadêmico

Inteligência artificial

Interview analysis

Introdução

Introdutor locutivo

Jair Messias Bolsonaro.

L-act

Legislação federal brasileira

Lei geral de proteção de dados pessoais

Lex-br-ius

Lexical bundles

Lexical frames

Léxico tabu

Limpeza de textos.

Língua inglesa

Língua italiana

Linguagem jurídica

Linguagem simples

Linguistic features

Linguística

Linguística computacional

Linguística de corpus

Linguística de corpus de aprendiz

Literatura brasileira traduzida

Literatura estrangeira

Luiz Inácio Lula da Silva

Machado de Assis

Memória

Metáfora

Metaphor identification

Modal verbs

Moderação de plataformas

Mulheres

Multi-dimensional analysis

Multimodal corpus

Multi-palavras

Natural language processing

N-grama de classe semântica.

Nheengatu

Nilc-metrix

Nominalization

Normas de tradução

O joio e o trigo

Objetivos de desenvolvimento sustentável (ODS)

Onomasiology

Padronização textual

Parsing sintático

Patriarcado

Pedagogia da tradução

Personal pronouns

Plataforma de dicionários

PLN

Português

Portuguese-english contrastive studies

Práticas discursivas

Pré-processamento de corpus

Processamento de linguagem natural

Prosódia

Python

Quadros e pacotes lexicais

Receitas

Recorrentes e preferenciais

Recursos lexicográficos

Redes sociais

Research articles

Research paper writing

Resenhas literárias

Resumo científico

Roda viva corpus

RST

Satire

Satirical news

Semantic fields

Semântica

Semântica de frames

Shape and stem disciplines

Simplificação textual e terminológica

Sintaxe

Sketch engine

Social media

Terminologia

Terminologia para tradução

Termos

Traços de normalização

Tradução

Tradução audiovisual acessível

Tradução automática

Tradução cultural

Tradução especializada

Tradução

Translator-oriented glossary

Transtorno do humor bipolar

Treebank

Tupinologia

UD

Universal dependencies

User- generated content

Variação

Variação diacrônica

Vocábulos

SUMÁRIO

RESUMOS
EBRALC-2024

RODA VIVA: UM CORPUS ORAL E A UNIVERSAL DEPENDENCIES

Cláudia Dias de BARROS[1]
Oto Araújo VALE[2]

Neste resumo é apresentado o trabalho sobre a construção de um subcorpus do Corpus Roda Viva (MIRANDA JR. et al., 2024), que é formado por 713 entrevistas de vários anos do programa Roda Viva da TV Cultura, transcritas por jornalistas de forma textualizada, nas quais há complementações das falas, por meio de inserções textuais, informações enciclopédicas, entre outros, o que faz com que percam o status de língua oral, passando a língua escrita. Dessa forma, nesta pesquisa tomou-se a decisão de construir o subcorpus com quatro entrevistas e, a fim de manter o status de língua oral, decidiu-se realizar a transcrição automática das entrevistas por meio de um ASR (Sistema de Reconhecimento Automático de Fala) chamado Whisper (RADFORD et al., 2023). Os textos transcritos apresentaram alguns problemas como transcrição equivocada de algumas palavras e erro de segmentação das sentenças, que precisaram ser corrigidos manualmente posteriormente. A escolha das quatro entrevistas se deu baseada na possível diversidade sintática apresentada pelos quatro entrevistados, sendo eles: uma governadora, um desenhista de história em quadrinhos, um jogador de futebol, e um rapper. A partir dos textos transcritos revisados foi realizada a anotação automática com o formalismo da Universal Dependencies (UD) (DE MARNEFFE et al., 2021), um projeto que tem como objetivo uma anotação gramatical consistente (etiquetas morfossintáticas, características morfológicas e dependência sintática), entre línguas humanas diferentes. Atualmente, a UD possui dezessete etiquetas morfossintáticas ou Part-of-Speech (PoS) tags e 37 etiquetas de relações de dependência – deprel (de dependency relation), que ligam dois a dois os elementos (tokens) de uma sentença. Um deles é chamado de head (núcleo), que é sempre uma palavra de conteúdo e o outro é chamado de dependente. A anotação UD foi realizada pelo parser PortParser (LOPES et al., 2024) e após isso, foi feita uma revisão manual por meio da ferramenta Arborator-Grew ElizIA (GUIBON et al., 2020) e foram identificados alguns fenômenos característicos da língua falada, como a presença de vocativos e marcas discursivas, como né, hein, entendeu, entre outros. A entrevista com o rapper foi a que apresentou menor formalidade e se mostrou desafiadora para o parser anotar corretamente as relações sintáticas. Na entrevista com a governadora do estado, observou-se uma grande quantidade de orações subordinadas e coordenadas, fruto de um discurso mais prolixo, característico de um político. A entrevista do jogador de futebol apresentou a ocorrência de muitas etiquetas dislocated, as quais marcam a presença de um item que poderia ser classificado como o sujeito da oração, mas, por estar longe do verbo, é substituído por um outro sujeito mais próximo, como ‘João’ em: “O João, ele sempre foi uma pessoa desconfiada”. O objetivo dessa

[1] Docente do Curso de Licenciatura em Letras, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Câmpus Sertãozinho

[2] Docente do Curso de Licenciatura em Letras e Bacharelado em Linguística, Universidade Federal de São Carlos – UFSCar.

anotação é fornecer um corpus de língua oral (a princípio as 5 entrevistas e, posteriormente, as outras 708 do Projeto Roda Viva) ao projeto Porttinari (PARDO et al., 2021), um grande corpus multigênero do Português do Brasil, composto por textos escritos, como artigos de jornal, tweets do mercado financeiro brasileiro, revisões de consumidores de e-commerce e revisões de livros.

Palavras-chave: Universal Dependencies; sintaxe; Linguística de Corpus; PLN.

ANÁLISE LINGUÍSTICA DE DISCURSOS PRESIDENCIAIS: UM ESTUDO BASEADO EM CORPUS

Ísis Beber de Souza Fiorilo ROCHA[3]
Bruno Neves Rati de Melo ROCHA[4]

Este trabalho mapeia temas relacionados ao uso das palavras “Brasil”, “Deus” e "governo" em discursos presidenciais do primeiro ano do primeiro mandato de Luiz Inácio Lula da Silva (2003), Dilma Vana Rousseff (2011) e Jair Messias Bolsonaro (2019). A análise visa entender a maneira pela qual cada presidente lida com questões evocadas por esses termos, usando como arcabouço teórico-metodológico a Linguística de Corpus (McEnery; Wilson, 1996; Sardinha, 2004). Para tanto, compilou-se um corpus de 1001 textos de discursos presidenciais, retirados do site da Biblioteca da Presidência, totalizando aproximadamente 1.600.000 tokens. O corpus é formado pelas transcrições de todos os discursos proferidos pelos presidentes no período estudado e exclui os discursos de vice-presidentes e outros representantes governamentais. Usando o Corpus Brasileiro (Sardinha, 2010) como corpus de referência, gerou-se listas de frequência (com/sem stopwords), listas de palavras-chave (com/sem stopwords), colocados das palavras alvo, linhas de concordância das palavras alvo e de colocados das palavras alvo. A escolha das palavras alvo se baseou em suas posições nas listas de frequência e de palavras-chave dos subcorpora: para todos os presidentes, “Brasil”, “Deus” e "governo" estão entre as cinco palavras lexicais mais frequentes e as cinco primeiras palavras-chave de cada mandato. Os dados analisados sugerem que cada presidente utiliza de maneira específica cada uma das palavras alvo. Para Lula e Dilma, "Brasil" aparece em colocados que indicam temas econômicos (“banco” e “risco” para Lula e “rico” para Dilma), políticas públicas e programas sociais (“analfabetismo” e “miséria” para Lula e “sem fronteiras” e “sem miséria” para Dilma), e termos que evocam relações exteriores (“ligação” e “Argentina” para Lula e “embaixador” e “Venezuela” para Dilma). Para Bolsonaro, "Brasil" ocorre sobretudo em contextos que expressam uma ideia de “resgate do Brasil” (como “Brasil que ressurge” e “colocar o Brasil no local de destaque”). A análise de “Deus” sugere que Bolsonaro relaciona sua fé pessoal a assuntos do governo (“agradeço”), evoca o movimento integralista (“Deus, família, Brasil”) e propaga o slogan de sua campanha eleitoral (“Brasil acima de tudo, Deus acima de todos”). Para Lula, “Deus” também é usado em contextos que apontam fé pessoal (“queira”, “peço”, “graças”). Para Dilma, a palavra “Deus” é pouco frequente, não tendo sido analisada detalhadamente. Quanto a “governo”, o primeiro colocado de Lula e Dilma é “federal”, ao passo que “meu” aparece em terceiro lugar, sugerindo que se referem ao próprio governo sobretudo de maneira institucional, mas também pessoal. Os demais colocados relacionam-se a temas como economia, relações exteriores e

[3] Bacharel em Letras-Estudos Linguísticos/ Ênfase em Linguística do Texto e do Discurso pela Universidade Federal de Minas Gerais (UFMG), membro do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da UFMG.

[4] Professor efetivo no curso de Letras da UFMG. Doutor em Estudos Linguísticos pela Faculdade de Letras da UFMG.

programas sociais (“sociais” e “programa” para Lula e “compromisso” e “bolsas de estudo” para Dilma). Para Bolsonaro, o primeiro colocado de “governo” é “nosso”, evidenciando uma postura pessoal, enquanto “federal” não figura entre primeiras vinte posições. Além disso, Bolsonaro co-relaciona "governo" a governos militares ditatoriais (“Médici” e “Figueiredo”) e civis não ditatoriais (“Sarney” e “anterior”, referindo-se a Michel Temer) e também a expressões como “respeita a família”, "adora a Deus" e “honra os militares”.

Palavras-chave: Linguística de Corpus; Discursos Presidenciais; Compilação de Corpus; Luiz Inácio Lula da Silva; Dilma Vana Rousseff; Jair Messias Bolsonaro.

IDENTIFICAÇÃO SEMIAUTOMÁTICA DE EQUIVALENTES ENTRE EXPRESSÕES IDIOMÁTICAS COM FRUTAS EM PORTUGUÊS-INGLÊS: DESCASCANDO ESSE ABACAXI

Julia de Souza BATISTA[5]
Isabela MOREIRA DE OLIVEIRA[6]
Stella Esther Ortweiler TAGNIN[7]
Elisa Duarte TEIXEIRA[8]
Rozane Rodrigues REBECHI[9]

As expressões idiomáticas (EIs) apresentam grande dificuldade para a tradução e a aprendizagem de línguas, seja por sua opacidade, seja pelo fato de que seu aprendizado só se dá por meio de repetidas exposições (MATTOS, 2003). Levando em conta as dificuldades enfrentadas por pesquisadores que trabalham com essas unidades (p. ex. XATARA, 2001; XATARA et al., 2001; SAG et al., 2002; RIVA, 2009; TAGNIN, 2013; PINNAVAIA, 2018; REBECHI e TRINDADE, 2021; SILVA e TEIXEIRA, 2021; ADEWUMI et al., 2022; ORTIZ ALVAREZ 2022), pensou-se em um projeto que visa, inicialmente, coletar o maior número possível de EIs contendo palavras relacionadas à grande área da alimentação, em português, inglês, italiano, chinês e espanhol, por enquanto. Planeja-se, em seguida, desenvolver estratégias de classificação e correlação das EIs entre as línguas, de modo a permitir sua identificação (semi-)automática em corpora, bem como a de possíveis equivalentes para diferentes contextos de tradução e de aprendizado de língua estrangeira. O objetivo do projeto é criar um sistema informatizado online para a identificação, coleta e consulta de EIs utilizando uma abordagem que permita uma busca tanto pela expressão em si, quanto pelos seus sentidos, ou seja, onomasiológica. O presente trabalho relata o planejamento e testagem das abordagens metodológicas utilizadas até o momento e os resultados obtidos no subconjunto de EIs contendo palavras do campo semântico FRUTAS, no par de línguas português-inglês. Esse primeiro recorte teve como objetivo principal chegar a uma lista de classificadores que pareça razoável e que permita a posterior identificação de EIs equivalentes em duas ou mais línguas de forma (semi-)automática. Tendo como base uma lista criada a partir da junção de: i) classificadores resultantes de um trabalho de mestrado sobre EIs da área da alimentação (MOREIRA DE OLIVEIRA, 2022); ii) o conjunto de dados para classificação refinada de emoções do Go Emotions

---

[5] Graduada em Línguas Estrangeiras Aplicadas (LEA-MSI) pela Universidade de Brasília (UnB); Graduanda em Letras Tradução Inglês também pela UnB.

[6] Mestra em Estudos da Tradução pela Universidade de Brasília (UnB), docente temporária do Departamento de Línguas Estrangeiras e Tradução (LET) da UnB e doutoranda do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS).

[7] Docente associada da Universidade de São Paulo (USP), atua nos Programas de pós-graduação em Estudos Linguísticos e Literários em Inglês e LETRA (USP).

[8] Docente da área de Tradução - Inglês do Departamento de Línguas Estrangeiras e Tradução (LET) da Universidade de Brasília (UnB), membro do Programa de Pós-Graduação em Estudos da Tradução - POSTRAD da UnB.

[9] Docente do Departamento de Línguas Modernas e Professora Permanente do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS).

(DEMSZKY et al., 2020); iii) uma lista de emoções humanas elaborada pelo Chat GPT; iv) e os classificadores listados no Themes of Oxford Dictionary of Idioms Index (AYTO, 2020), utilizados no trabalho de Rafatbakhsh e Ahmadi (2019), cinco pesquisadoras da equipe classificaram uma lista em português e outra em inglês contendo cada mais de 60 EIs do referido campo semântico. Os resultados individuais foram comparados e a lista, refinada. As EIs foram, então, classificadas novamente por cada pesquisadora, utilizando-se essa versão, contendo cerca de 90 classificadores. Depois de chegarem a uma classificação majoritariamente consensual das EIs em ambas as línguas, foi feito o cruzamento dos dados para verificar o quanto essa metodologia se mostraria eficaz na identificação (semi-)automática de possíveis equivalentes interlinguais. Observamos, por exemplo, que do total de 90 classificadores, 60 foram usados e "facilidade" nos permitiu identificar uma possível equivalência entre "mamão com açúcar" e "as easy as apple pie", "easy peasy lemon squeezy" e "low hanging fruit"; e entre "a cereja do bolo" e "the cherry on top", ambas classificadas com "excelência". Concluído este balão de ensaio, nosso próximo passo será expandir a classificação para outras categorias, procurando refinar ainda mais a lista de classificadores, sempre pensando em outras formas de correlacionar esses dados.

Palavras-chave: Expressões idiomáticas; classificadores semânticos; fraseologia português-inglês; PLN; Linguística de Corpus.

NOT EVERY B!7CH IS A B1TC|-|: TESTANDO A MODERAÇÃO DE DISCURSO DE ÓDIO EM ESPAÇOS VIRTUAIS COM PADRÕES TEXTUAIS

Priscila TAVARES DA SILVA[10]

Em espaços virtuais de interação, é comum que usuários perpetradores de discursos de ódio driblem a moderação (Cristani, 2022) utilizando Leetspeak, que são escritas alternativas com caracteres especiais (Froud, 2021), e Algospeak, neologismos ou combinações de palavras que fazem alusão ao que se pretende dizer (Huyghe, 2022). A presente pesquisa descreve um projeto piloto para propor e testar a efetividade de moderação de espaços virtuais a partir da análise de combinações de palavras. Espera-se demonstrar que existem padrões textuais recorrentes nesses discursos de ódio e que a moderação por agrupamentos de palavras pode ser mais efetiva do que por palavras isoladas. O objetivo é chegar a uma metodologia que possa ser aplicada com eficácia para diminuir os casos em que os usuários conseguem driblar a moderação. Utilizando ferramentas da Linguística de Corpus, pode-se analisar a co-ocorrência entre itens lexicais em um contexto próximo (Teixeira, 2008, p. 170). A análise é feita com um software capaz de comparar textos organizados em um corpus, que é um conjunto extenso o suficiente para ser representativo do fenômeno que se busca avaliar (Berber Sardinha apud Teixeira, 2008, p. 159). Para a análise proposta, o primeiro passo é identificar em quais contextos os itens lexicais são utilizados como discurso de ódio, entendido como um discurso disciplinador que busca manter ou reforçar o status quo de poder de um grupo sobre outro (Foucault apud Silva, 2016, p. 40). O objeto da análise é o corpus TwitterHateSpeech , que contêm tweets do X. Os dados serão analisados com o software AntConc. Pretende-se levantar e classificar as palavras-chave desse corpus em duas categorias: “sim” e “não”, em que “sim” corresponde a “palavra utilizada em contexto de discurso de ódio”. Em seguida, observaremos quais correlações surgem da análise de clusters e n-grams das palavras levantadas na pesquisa. Caso existam padrões reconhecíveis em seu uso, checaremos se buscas com esses padrões retornam somente tweets classificados como discurso de ódio, mesmo sem conter as palavras-chave, o que permitiria combater os artifícios utilizados pelos usuários. Nossa hipótese é de que a análise dos padrões permitiria chegar a uma lista de clusters e n-grams cujo uso, em um corpus não rotulado como discurso de ódio, permita identificar sua ocorrência mesmo sem a presença das palavras-chave. A título de ilustração, escolhemos a palavra bitch, que ocorre 8.353 vezes no corpus. Nas 500 primeiras ocorrências somente 92 foram classificados como discurso de ódio. Dentre os padrões mais comuns estão as expressões look like a bitch e be a bitch, usados como sinônimos de “covarde” ou “difícil”, como em “@BarackObama looks like a bitch in foreign policy” e “Life is a bitch”. Esses casos são mais numerosos do que o uso em discursos de ódio, como em “Every nigga can make a female act like a bitch #ThatsAFact”.

[10] Mestranda em Estudos da Tradução na Universidade de Brasília

Palavras-chave: Linguística de Corpus; discurso de ódio; moderação de plataformas; padronização textual; AntConc.

ABSTRACT REGISTER VARIATION BETWEEN HUMAN AND AI

João FEKETE[11]
Deise Prina DUTRA[12]

In the little time chatbots powered by Large Language Models (LLMs) and Artificial Intelligence (AI) have been available to the general public, different uses of them have been created (Holmes & Tuomi, 2022; Islam et al., 2020). The use of these technologies in writing has brought concerns to many areas, including academic writing (Thorp, 2023; Nature Editorials, 2023). In order to understand the real effect of such usages, studies have addressed AI detectors effectiveness (Hu et al., 2023; Lu et al., 2023; Walters, 2023), human judgment for detecting AI generated texts (Ma et al., 2023), and linguistic analysis of these texts (Berber Sardinha, 2024). In our research, we focus on expanding the linguistic analysis on the similarities between human-authored and AI-generated texts, taking into account lexicogrammatical features and their communicative functions in abstracts, which are understood as an academic register .The experiment was undertaken using two different corpora, 450 abstracts from high impact journals from 3 different areas (Reference corpus) and the second corpus 900 abstracts created by ChatGPT3.5, using as resource the content from the 450 articles of the reference corpus. The AI-generated corpus is split into two different subcorpora: Simple Query and Complex Query. The two corpora were created using LangChain (LangChain, 2024) and ChatGPT3.5-turbo (OpenAI, 2024). First, we gathered all the sections (except the abstract) from the articles from which the reference corpus was extracted, then, we used a splitter to segment the text into chunks and embedded the information to be used through Chroma db. From Langchain's retrieval chain, we provide all chunks to the AI, which then selects the ones which will be sent to ChatGPT at the moment of answering the query. For the Simple Query corpus we generated the abstracts with a simple prompt, which required only an abstract to be created based on the provided information (Simple Query) and, for the Complex Query, we created a prompt with the persona approach (White et al., 2023), which asked the AI to behave as an experienced article publisher and professor. To analyze the data, we make use of the Additive Multidimensional Analysis (AMDA) (Berber Sardinha et al., 2019) with the Dimensions 1, 2, and 5 from Biber (1988). Results have shown statistical differences between the AI-generated abstracts and human-authored for Dimensions 1, 2, and 5, for both AI corpora. From the output data, we can conclude that: human-authored texts display a greater variety of features; using the persona prompt does not guarantee a more human-like output from the AI model; and some communicative purposes are better defined for artificial intelligence than others, such as understanding abstracts as an information production type of text rather than an involved one. Furthermore, we encourage researchers to take into account the creation of testing data using authentic scenarios of AI use for text creation.

---

[11] Bacharel em Inglês pela Universidade Federal de Minas Gerais

[12] Professor titular da Universidade Federal de Minas Gerais

Palavras-chave: Artificial Intelligence; Multi-dimensional Analysis; Corpus Linguistics; Computational Linguistics; Academic Writing

CRIAÇÃO DE UM GLOSSÁRIO LIBRAS-PORTUGUÊS DO BRASIL: UMA INICIATIVA DE ACESSIBILIDADE COMUNICACIONAL

Gabriela WICK-PEDRO[13]

Esta pesquisa tem como objetivo a criação de um glossário Libras-Português do Brasil, visando promover a acessibilidade comunicacional, especialmente na divulgação científica. A proposta foca no desenvolvimento de conteúdos e infraestrutura voltados para a editoração e publicação de materiais científicos acessíveis, com especial atenção à comunidade surda. A comunicação de resultados de pesquisa é parte essencial do ciclo de vida da atividade científica, conectando a pesquisa à aplicação dos resultados, com potencial para melhorar a qualidade de vida e fomentar novas investigações. A divulgação científica é considerada um meio informal de comunicação científica, utilizando uma linguagem acessível a diversos públicos e incorporando novos elementos ao processo de circulação de informações científicas e tecnológicas (BJÖRK, 2005; BUENO, 2010). A linguagem pública na divulgação científica deve alcançar o maior número possível de pessoas, ultrapassando o círculo restrito dos especialistas acadêmicos (SILVA; LAZZAROTTI FILHO; SILVA, 2011). A acessibilidade comunicacional, dentro desse contexto, é essencial. A divulgação científica, entendida como um meio informal de comunicação, utiliza uma linguagem acessível e visa ampliar o público alcançado, superando o círculo de especialistas acadêmicos. Neste contexto, a acessibilidade comunicacional, especialmente para usuários da Língua Brasileira de Sinais (Libras), é crucial, considerando a diversidade da comunidade surda. O processo de criação do glossário envolve: i) a preparação e limpeza de corpus, ii) a extração automática de termos por meio de ferramentas de Processamento de Linguagem Natural (PLN) e Inteligência Artificial (IA), iii) a validação com especialistas da área e iv) a categorização dos termos. Além disso, há uma ênfase em como o projeto se ancora espacialmente em instituições e eventos, como indicam estudos sobre a relação entre território e produção científica (CERTEAU, 1994). A pesquisa, em andamento, visa criar um glossário bilíngue com o intuito de incluir mais efetivamente a comunidade surda na comunicação científica, promovendo uma inclusão mais abrangente. Resultados preliminares apontam para a viabilidade do uso de técnicas de Linguística de Corpus e PLN para a criação desse recurso, oferecendo um importante avanço no campo da acessibilidade comunicacional.

Palavras-chave: acessibilidade comunicacional; divulgação científica; glossário bilíngue; processamento de linguagem natural; inclusão surda.

[13] Pesquisadora do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília/DF, bolsista FINATEC.

RODA VIVA CORPUS: STRUCTURING A MULTIMODAL LINGUISTIC RESOURCE FROM BRAZILIAN TELEVISION INTERVIEWS

Gabriela WICK-PEDRO[14]
Cláudia Dias de BARROS[15]
Oto Araújo VALE[16]

The Roda Viva Corpus aims to formalize and structure a multimodal linguistic resource derived from interviews broadcasted on Roda Viva, a long-running interview show on TV Cultura, a staple of Brazilian television since 1986. The corpus includes both textual transcriptions and corresponding video recordings, with the initial dataset consisting of 713 interviews conducted between 1986 and 2009. These interviews feature prominent figures such as politicians, artists, scientists, and intellectuals, and were made publicly available through the Memória Roda Viva portal, a project initiated by FAPESP in 2007 (FAPESP, 2024). This study seeks to transform the raw data available on the portal into a linguistically structured corpus for use in Corpus Linguistics (CL) and Natural Language Processing (NLP). The primary aim is to provide a resource that enables detailed linguistic analyses, such as the investigation of discourse markers, interactional features, and pragmatic phenomena in Brazilian Portuguese as it is spoken in formal interview contexts. In particular, this corpus offers a unique opportunity to analyze multimodal data, combining textual transcriptions with corresponding video recordings, which are essential for understanding non-verbal cues in communication (Botin, 2016; Pacheco, 2020). The construction of the Roda Viva Corpus involved extensive data cleaning, and two preliminary versions of the corpus are currently available. Version 0.1 maintains the original transcriptions with minimal cleaning, such as the removal of hyperlinks and special characters, while Version 0.2 offers a more refined dataset that excludes non-verbal transcriber interventions (e.g., "coughing," "sighing"). Both versions are available in CSV and JSON formats, allowing for flexibility in computational processing. The corpus is also annotated with metadata, including the date of the interview, the name of the interviewee, the speaker's identity, and the order of each utterance. Although the Roda Viva material has been academically cited in only a few studies , this project aims to fill the gap by offering a formally structured resource that can support a wide range of linguistic and computational research. The project represents a step forward in the integration of CL and NLP, providing a rich, multimodal dataset that bridges the gap between traditional linguistic analysis and modern computational methods. The corpus is designed to promote interdisciplinary research at the intersection of Corpus Linguistics, Artificial Intelligence, and media studies. This contribution aims to not only enhance the availability of Brazilian Portuguese

[14] Pesquisadora em Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília/DF, bolsista FINATEC

[15] Docente Efetiva de Português/Inglês no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Campus Sertãozinho, Sertãozinho/SP

[16] Docente Associado do Departamento de Letras (DL) na Universidade Federal de São Carlos (UFSCAR), São Carlos/SP

linguistic resources but also provide a robust tool for exploring multimodal communication in high-stakes, formal interview settings.

Palavras-chave: Roda Viva corpus; multimodal corpus; corpus linguistics; brazilian television; interview analysis.

CONSTRUÇÃO DO DOMÍNIO DA ACESSIBILIDADE POR MEIO DE FRAMES SEMÂNTICOS: UMA CONTRIBUIÇÃO PARA A FRAMENET BRASIL

Iasmin Valéria Miranda RABELO[17]
Maucha Andrade GAMONAL[18]
Adriana Silvina PAGANO[19]

A teoria linguístico-cognitiva da Semântica de Frames (FILLMORE, 1982) tem por objetivo compreender a construção de significado através de contextossócio-culturais. Para isso, a teoria se fundamenta nos frames, uma estrutura designificados relacionados que evidencia estruturas cognitivas para a representaçãode conceitos específicos. Nesse contexto, a FrameNet Brasil, desenvolvida naUniversidade Federal de Juiz de Fora (UFJF), é um projeto de lexicografiacomputacional que busca desenvolver tecnologias linguísticas com base nSemântica de Frames e na Linguística de Corpus. Dessa forma, este trabalho busca expandir a base de dados da FrameNet Brasil a partir do léxico especializado da área da acessibilidade, através da construção de um domínio específico. Esta pesquisa se ancora, primeiramente, na teoria da Semântica de Frames (FILLMORE, 1982), suas aplicações léxicas e sintáticas (GAWRON, 2008) e o projeto FrameNet Brasil (SALOMÃO, 2009). As diretrizes para construção de um domínio e criação de frames seguem os trabalhos de Dutra (2024) e Gamonal (2013), assim como os critérios de compilação de corpus definidos por Sardinha (2004). Os conceitos de semântica lexical abordados têm como base o trabalho de L'Homme (2020) e a relevância da construção de frames em domínios especializados (L'HOMME ET AL, 2014). Ademais, esta pesquisa também abrange o histórico da acessibilidade noBrasil (COSTA ET AL, 2005) e o desenvolvimento da Tecnologia Assistiva (RODRIGUES e ALVES, 2013). Este projeto se desenvolve a partir de uma pesquisa quantitativa de abordagem bottom-up, partindo dos dados para a criação de frames. A primeira etapa, já em desenvolvimento, consiste na investigação do léxico especializado através da compilação de um pequeno corpus sobre a área da acessibilidade (cartilhas, glossários e outros recursos) com, até então, 37774 tokens e 6454 types. O corpus foi analisado através do software concordanceador AntConc, identificando as terminologias específicas mais frequentes nos documentos e seus contextos de uso. A partir disso, as possíveis unidades lexicais foram extraídas. O próximo passo, também em progresso, compreende o mapeamento dessas unidades lexicais candidatas com os dados já presentes na FrameNet Brasil. Após isso, é possível sugerir a criação de novos frames para a construção do domínio da acessibilidade, com definições mais específicas e abrangentes dos conceitos e terminologias da área. Por fim, através das relações entre frames

[17] Graduanda da Faculdade de Letras da Universidade Federal de Minas Gerais. Universidade Federal de Minas Gerais, Minas Gerais.

[18] Residente de pós-doutoramento no Programa de Pós-Graduação em Linguística na Universidade Federal de Minas Gerais, Minas Gerais, atualmente é bolsista Capes.

[19] Professora Titular de Linguística Aplicada da Universidade Federal de Minas Gerais, Minas Gerais, bolsista de produtividade em Pesquisa IC do CNPq.

é possível conectar o domínio da acessibilidade à rede léxico-semântica de FrameNet Brasil. Com as ULs e frames criados, a última etapa na modelagem de um domínio é a anotação de unidades lexicais realizada de acordo com a metodologia de anotação lexicográfica da FrameNet Brasil. As etapas já em andamento desta pesquisa revelam a riqueza lexical da área da acessibilidade. As investigações e análises terminológicas indicam um domínio vasto e com diversas possibilidades de aplicação dentro da FrameNet Brasil. Expandir a base de dados com esse léxico especializado permite uma compreensão aprofundada dos termos e conceitos relevantes para uma parcela discriminada da população brasileira. Além disso, a amplificação das unidades lexicais proporciona mais informações para o desenvolvimento de tecnologias linguísticas e possibilita futuras pesquisas na área.

Palavras-chave: Semântica de Frames; criação de frames; domínio especializado; acessibilidade; deficiência.

ENSINO DE ITALIANO EM CONTEXTO UNIVERSITÁRIO: POR UMA REFLEXÃO LINGUÍSTICA DIRECIONADA POR CORPUS

Graziele Altino FRANGIOTTI[20]

Esta proposta tem como objetivo contribuir para a discussão sobre possíveis caminhos para um ensino de línguas estrangeiras direcionado por corpus, focalizando de maneira especial a aprendizagem do italiano em contexto universitário. Parte-se do pressuposto de que estudos que dialoguem com a Linguística de Corpus são pouco numerosos na área do ensino de italiano no Brasil (FRANGIOTTI, 2019; FRANGIOTTI, 2021), lacuna essa que leva à indagação sobre quais seriam os efeitos do ensino direcionado por dados no aprendizado dessa língua por brasileiros. De modo mais específico, pretende-se compreender se e de que maneira um percurso didático formulado a partir de uma obra literária escrita na Idade Média e de um conjunto de suas traduções pode fomentar a compreensão sobre mudanças linguísticas na língua italiana. Como organização geral da apresentação, pretende-se, em um primeiro momento, descrever o caminho teórico-metodológico que orienta a formulação e a aplicação de uma sequência didática em uma disciplina oferecida no curso de graduação em Letras Italiano na Universidade Federal de Santa Catarina (UFSC). Para tanto, parte-se dos trabalhos de Berber Sardinha (2010), Morales (2008), Condi de Souza (2005), Barbosa (2004), entre outros, para a apresentação das potencialidades do data-driven learning. Já com base em Antunes (2014, 2009) e Carter e McCarthy (1995) expõe-se uma concepção de ensino de línguas estrangeiras centrada na análise textual e na reflexão metalinguística indutiva como catalisadores do desenvolvimento de competência comunicativa. Finalmente, levando em conta a taxonomia de Bloom e sua revisão (respectivamente BLOOM [1972] e KRATHWOHL [2002]), sustenta-se uma perspectiva de ensino que diversifique as atividades propostas em sala de aula e que coloque a comparação entre as línguas como uma técnica didática relevante para a promoção do noticing (SCHMIDT, 1990). Após essa etapa teórica, será caracterizada a sequência didática propriamente dita, que se ancora, sobretudo, na obra italiana Il Principe (1532) de Nicolau Maquiavel e em algumas de suas principais traduções para o português. Posteriormente, serão discutidos os resultados dos participantes da pesquisa em atividades didáticas onde a identificação de semelhanças e diferenças entre as línguas foi estimulada e ponderados os efeitos desse procedimento para a sensibilização quanto à variação linguística diacrônica.

Palavras-chave: Linguística de corpus; ensino de línguas estrangeiras; língua italiana; variação diacrônica.

[20] Docente do Departamento de Língua e Literatura Estrangeiras da Universidade Federal de Santa Catarina, Florianópolis/SC, bolsista ADC-1C do CNPq.

AS CARACTERÍSTICAS PROSÓDICAS TEMPORAIS DA UNIDADE INFORMACIONAL DE INTRODUTOR LOCUTIVO

Gabriela Dias TOLEDO[21]
Marcelo Augusto VIEIRA[22]
Tommaso RASO[23]

Segundo a Language into Act Theory (L-Act), teoria corpus driven que tem como foco a organização da fala espontânea, as unidades informacionais podem ser identificadas a partir da sua funcionalidade, do seu perfil prosódico e da sua distribuição em relação ao Comentário (unidade com força ilocucionária) e agrupadas a partir de sua macro-funcionalidade textual ou dialógica (CRESTI, 2000). O Introdutor Locutivo (INT) é uma unidade informacional textual cuja função é sinalizar que os elementos que o sucedem devem ser interpretados pelo ouvinte em um plano pragmaticamente diferente do resto do enunciado, evidenciando um salto para outro nível hierárquico. Seu perfil prosódico não possui foco ou forma definida, mas é descendente, com frequência fundamental e intensidade menor que o elemento seguinte e taxa de articulação maior que o resto do enunciado (CRESTI, 2000; MAIA ROCHA; RASO, 2011). O objetivo da pesquisa é investigar as características prosódicas temporais do INT seguido do discurso reportado (metailocução mais realizada após a unidade), buscando descrever um aspecto formal que pode ser crucial para veicular a função dessa unidade informacional e diferenciá-la das demais com maior acurácia. Metodologia desenvolvida para a pesquisa: seleção de 58 INTs de minicorpora compilados e tratados no projeto C-ORAL-BRASIL (RASO; MELLO, 2012), agrupados de acordo com o seu tamanho (número de palavras prosódicas); segmentação e anotação manual dos INTs e dos contextos adjacentes a eles (unidade que precede o INT e o discurso reportado) pelo Praat (BOERSMA; WEENINK, 2023); extração automática das medidas de duração normalizada, taxa de articulação e proporção de apagamento silábico do INT e dos seus contextos adjacentes a partir da adaptação do script SGdetector (BARBOSA, 2006); tratamento estatístico e modelagem linear de efeitos mistos com o auxílio do R (R CORE TEAM, 2023) das características prosódicas temporais do INT. Resultados das análises estatísticas: o INT possui menor duração, maior taxa de articulação e maior proporção de apagamento silábico que seus contextos adjacentes; quanto maior o INT, menor sua taxa de articulação e proporção de apagamento silábico; o INT possui maior proporção de apagamento silábico em oxítonas que seus contextos adjacentes, enquanto que em paroxítonas não há diferença significativa entre as estruturas; as palavras mais frequentemente realizadas no INT são as oxítonas, principalmente o verbo 'falar' e o advérbio 'assim', enquanto que as mais frequentemente realizadas nos contextos

[21] Aluna de doutorado do Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Minas Gerais, Belo Horizonte/Minas Gerais.

[22] Aluno de doutorado da School of Communication Sciences and Disorders da Universidade McGill, Montreal/Quebec.

[23] Docente da Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte/Minas Gerais.

adjacentes são as paroxítonas. Ainda, análises preliminares (sem tratamento estatístico) sugerem que INTs com duas ou mais palavras prosódicas tendem a acelerar do começo da unidade até o alongamento pré-fronteiriço ao discurso reportado, sendo a maior parte de sua estrutura acelerada.

Palavras-chave: L-Act; fala espontânea; prosódia; Introdutor Locutivo; Linguística de Corpus.

DESAFIOS METODOLÓGICOS NA COMPILAÇÃO DO CORPUS DE TEXTOS ACADÊMICOS DAS CIÊNCIAS AGRÁRIAS

Deise Prina DUTRA[24]
Ana Eliza Pereira BOCORNY[25]
Danilo Duarte COSTA[26]
Gustavo Leal TEIXEIRA[27]
Carolina Godoi de Faria MARQUES[28]

No campo dos estudos em Inglês para fins Acadêmicos (IFA), corpora especializados desempenham um importante papel em investigações linguísticas de base empírica. No que se refere à escrita acadêmica, atenção tem sido dada para o fato de que os textos das diferentes áreas do conhecimento apresentam especificidades linguísticas (HYLAND, 2004; 2006). No campo das ciências agrárias, uma área estratégica para o desenvolvimento nacional, observou-se uma carência de estudos que se proponham a descrever a escrita acadêmica de brasileiros produzida em inglês como língua adicional. Para preencher esta lacuna e, a fim de compreender a linguagem utilizada nesta área no Brasil, visando propor instrumentos de apoio ao desenvolvimento linguístico da comunidade acadêmica, identificamos a necessidade da compilação de dois corpora. O primeiro contém textos produzidos por autores pouco experientes - dissertações e teses - e o segundo por artigos publicados em revistas de alto impacto internacionais para posterior comparação. Neste trabalho descrevemos os procedimentos adotados para a compilação do corpus de teses e dissertações (Corpus de Textos Acadêmicos das Ciências Agrárias) arquitetado de maneira a ser representativo da comunidade acadêmica de autores pouco experientes das ciências agrárias. Os procedimentos metodológicos adotados para a compilação seguem os pressupostos trazidos pela Linguística de Corpus, sendo embasados sobretudo nos conceitos de representatividade (BIBER, 1993) e balanceamento (SINCLAIR, 2004). Os textos foram obtidos a partir de buscas em repositórios institucionais de universidades brasileiras das cinco regiões do país. O corpus foi dividido em subcorpora, definidos a partir dos cursos ofertados por um instituto de ciências agrárias de uma universidade federal: Agronomia, Engenharia Agrícola, Engenharia de Alimentos, Engenharia Florestal e Zootecnia. Selecionados os textos, procedeu-se à etapa de limpeza, realizada manualmente com o objetivo de removerem-se os caracteres especiais, gráficos, tabelas, imagens e números de páginas. Dividiu-se os textos em seções: resumo, introdução, metodologia, resultados e conclusão, uma vez que possuem funções específicas, com estrutura e traços linguísticos distintos. Por fim, aos textos foi atribuído um código para facilitar a organização do corpus e a localização das

[24] Docente do Programa de Pós-graduação em Estudos Linguísticos (POSLIN) da Faculdade de Letras da UFMG

[25] Professora - Universidade Federal do Rio Grande do Sul, Porto Alegre/RS

[26] Professor - Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina/MG

[27] Professor - Universidade Federal de Minas Gerais, Montes Claros, Minas Gerais/MG

[28] Doutoranda - Programa de Pós-grafuação em Estudos Linguísticos, Universidade Federal de Minas Gerais, Belo Horizonte/MG

informações. Três fatores principais impactaram a compilação do Corpus de Textos Acadêmicos das Ciências Agrárias. O primeiro foi a ausência de Trabalhos de Conclusão de Curso escritos em inglês por brasileiros, resultando na exclusão dessa categoria do corpus. O segundo fator foi a disparidade no número de trabalhos disponíveis nos repositórios de cada região do país, especialmente na região Centro-Oeste, onde foram encontradas apenas duas teses. Esse dado contrasta com os resultados das outras regiões, destacando-se a região Sudeste, onde a maioria dos textos foi coletada. O terceiro fator foi a identificação de teses e dissertações compostas por artigos científicos em sua estrutura, o que parece ser uma prática comum nas ciências agrárias. Portanto, foi necessário criar dois subcorpora: um com textos no formato canônico (organizados por seções, como introdução e metodologia) e outro com textos não-canônicos (compostos por artigos). Foram obtidos um total de 180 trabalhos acadêmicos, dos quais 153 são do formato não-canônico e apenas 27 no formato canônico, totalizando aproximadamente 2.300.000 palavras.

Palavras-chave: Compilação de corpus; Escrita acadêmica; Ciências Agrárias.

DO FRASEOLOGISMO À METÁFORA NA EXPLORAÇÃO DO CORPUS JORNALÍSTICO DE O JOIO E O TRIGO

Bianca Mara GUEDES DE SOUZA[29]

Neste pôster, apresentamos uma exploração inicial de um corpus, parcialmente coletado, composto por textos de diferentes gêneros, extraídos do jornal O joio e o trigo. Portanto, estas são as primeiras análises realizadas para uma tese de doutorado em andamento. O joio e o trigo é um projeto de jornalismo investigativo que defende o papel central da prática jornalística como ferramenta de mudança social, especialmente tratando-se de temas como o combate às grandes corporações, com destaque para as do ramo alimentício e do agronegócio (O JOIO E O TRIGO, 2017). A busca exploratória realizada no corpus teve como objetivo a identificação e descrição de unidades fraseológicas e/ou unidades fraseológicas especializadas, seguida da identificação e descrição de expressões metafóricas com detalhamento da metáfora conceptual, relações de domínios (fonte e alvo), mapeamentos e desdobramentos. As unidades fraseológicas são definidas como combinações estáveis de pelo menos duas palavras, cujo limite superior é a oração composta, são caracterizadas pela fixação e/ou idiomaticidade (CORPAS PASTOR, 2010). Já as unidades fraseológicas especializadas são unidades sintáticas (não lexicais) de um domínio especializado, compostas por mais de um lexema sendo altamente frequentes (CABRÉ et al., 1996). Para o estudo, coletamos os primeiros seis meses de publicações do jornal, a saber de outubro de 2017 a março 2018, em português brasileiro. Empreendemos uma análise fundamentada teórico-metodologicamente na Linguística de Corpus (LC) (PARODI, 2010), com a qual articulamos os estudos de Fraseologia (CORPAS PASTOR, 2010) e Metáfora (LAKOFF; JOHNSON, 2002; BERBER SARDINHA, 2009). Na análise utilizamos o Sketch Engine (2016), um programa pago que permite a análise de textos online. Realizamos a análise em duas partes, primeiro, por meio de uma análise impressionística (BERBER SARDINHA, 2004) para a qual selecionamos cinco textos aleatórios para leitura completa, a partir dela notamos a presença de: metáforas do futebol; metáforas da guerra; e a construção da indústria ligada ao léxico das emoções. Para a segunda parte da análise, retornamos ao corpus total, com o auxílio das ferramentas Wordlist, Keywords, Concordance e Word Sketch do SE. Os principais resultados estão relacionados à metáfora conceptual ALIMENTAÇÃO É GUERRA, mapeada a partir do uso de unidades fraseológicas especializadas como conflito de interesses, sinais de advertência, ligar o alerta vermelho, sair em defesa e fazer uma defesa. Ademais, entre os resultados importantes identificamos, na leitura de linhas de concordância geradas com indústria, como essa é caracterizada a partir de emoções e sentimentos, nessa esteira, inferimos a metáfora conceptual INDÚSTRIA É ENTE HUMANO.

---

[29] Doutoranda em Estudos Linguísticos no Programa de Pós-graduação em Estudos Linguísticos da Universidade Federal de Uberlândia (PPGEL/UFU), bolsista CAPES.

Palavras-chave: Linguística de Corpus; Fraseologia; Metáfora; O joio e o trigo

AS PRÁTICAS TRADUTÓRIAS PRESENTES NO LIVRO *HARRY POTTER AND THE CHAMBER OF SECRETS*: UMA COMPARAÇÃO ENTRE AS VERSÕES BRASILEIRA E JAPONESA

Júlia Cristina VALVERDE DA SILVA[30]

A série de livros “Harry Potter”, de autoria da escritora inglesa J.K Rowling, foi publicada entre 1997 e 2007 e foi traduzida para ao menos 79 línguas. Apesar de terem como ponto inicial o mesmo texto de partida, são as normas tradutórias existentes no país e cultura de recepção que determinam as estratégias de tradução empregadas. Essas normas são restrições socioculturais particulares a uma dada cultura e período que regulam quais obras são traduzidas e de que maneira o são (Munday, 2008, p.112). Tendo isso em consideração, ao analisar comparativamente, a tradução da obra *Harry Potter and the chamber of secrets* para a língua japonesa e para o português do Brasil, buscou-se identificar as tendências tradutórias de cada obra e criar hipóteses em relação à posição ocupada por obras traduzidas e pelos tradutores nos respectivos países para poder, então, realizar generalizações acerca de estratégias de tradução predominantes em um dado gênero literário nos diferentes países. Partindo dos pressupostos de Gideon Toury (2012, p.63), para quem normas são valores ou ideias gerais compartilhadas por uma dada comunidade acerca do que é adequado ou inadequado ao se traduzir, objetivou-se depreender quais são as normas presentes nas traduções de *Harry Potter and the chamber of secrets* por meio da análise de fatores linguísticos (com o uso de um corpus paralelo) e extralinguísticos (investigação de paratextos com base no modelo de descrição de Lambert e van Gorp, 2014). Dessa maneira, a partir do estabelecimento de categorias de análise, selecionadas após a investigação das palavras mais representativas (*keywords*) do texto de partida e de suas respectivas traduções, observou-se como os itens lexicais em crivo foram traduzidos nas línguas em estudo e como essas estratégias potencialmente apontavam para comportamentos de tradução predominante nas duas culturas. As categorias analisadas foram as de “Onomásticos/Antropônimos”; “Tradução de itens lexicais relacionados ao mundo da magia” e “Nível de formalidade/*Yakuwarigo.* As estratégias de tradução encontradas em fase inicial apontam para a tendência estrangeirizante da versão japonesa que, em muitas passagens, apenas reproduz as palavras em língua inglesa com a transliteração para o silabário katakana, empregando adaptação fonológica. Na contramão dessa tendência, a versão brasileira apresentou como estratégia a criação de neologismos para itens lexicais relacionados à magia e a tradução de nomes próprios em português. A terceira categoria de análise revela como o *yakuwarigo*—conjunto de marcadores linguísticos usados para acentuar características de determinados personagens e criar estereótipos—foi utilizado como recurso narrativo e tradutória na versão japonês, adicionando, inclusive, nuances ausentes no texto de partida.

[30] Doutoranda do Programa de Pós-Graduação em Linguística (PPGL) da UnB.

Palavras-chave: Estudos Descritivos da Tradução; corpus paralelo; literatura estrangeira; normas de tradução.

PRONOUNS ON SOCIAL MEDIA: PRACTICES AND OTHERING

Ziyang LIU[31]

Personal pronouns and gender identities in recent years have gained considerable attention as the visibility of nonbinary gender identities is increasing especially on social media. Among these personal pronouns, singular they has been an interesting one which has led to wide-ranging discussions because of its nonbinary use that personal pronoun they is increasingly used to refer to a person with a nonbinary gender identity. The establishment of the nonbinary use of personal pronoun they has been confirmed by both American Dialect Society (ADS) and Merriam-Webster, as singular they has been voted as the word of the Decade 2010-2019 by ADS (2020) and the Word of the Year for 2019 by Merriam-Webster (2019). Both the studies of Richards and Barker (2013) and of Brown et al. (2020) have confirmed the positive effect of using correct pronouns. Referring to someone by the wrong pronouns that does not correctly reflect their gender identity, also called misgendering (Yarbrough, 2018; Brown et al., 2020; Cordoba, 2022), can be harmful to the mental health of the referee (Brown et al., 2020). Given the paucity of quantitative research on use of pronouns on social media, this study attempts to fill the gap by using a combination of corpus linguistics techniques and critical discourse analysis to investigate the innovative use of personal pronouns on social media. Especially, Twitter (now X) is selected as the source of data. This study constructs a small-sized Twitter-specific corpus and uses the twitter scraping tool named Twint to collect posts that are relevant to the discussion of pronouns and importing the preprocessed data into AntConc. By utilizing the quantitative tools in AntConc—n-gram tools, word frequency list, concordance analysis—this study finds two significant practices of personal pronouns disclosure and two discursive strategies of Othering, supported with concordances lines and discourse analysis. Two practices to express pronouns include active pronouns articulation when one takes the initiative to share one's pronouns and pronouns display when one puts one's pronouns on bio. Negative uses of pronouns include that pronouns as a identifying factor to judge and ostracize an individual and that stereotypes about nonbinary people and those who use pronouns lead to Othering. Although it is clear that the discourse on social media is constantly changing, it is of great significance to capture the dynamics of different groups when it comes to personal pronouns, and to confirm the use of pronoun practices and the existence of ostracism and stereotyping towards pronoun users. The data and findings will raise people's attention of the current situation of social media discourse and hopefully will lead to discussions on how to build pronouns-inclusive communities online.

Palavras-chave: personal pronouns; social media; corpus linguistics.

[31] Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

MULHERES E LÉXICO TABU: UMA ANÁLISE DE FRASEOLOGISMOS BASEADA EM CORPUS

Mayra Natanne Alves Marra[32]

Este trabalho é parte de uma pesquisa de doutorado em andamento, de base empírico-descritiva. Neste recorte, são apresentadas as primeiras análises que resultaram da exploração inicial de uma parte do corpus da pesquisa. O objetivo deste estudo é identificar e descrever diferentes tipos de fraseologismos em torno dos vocábulos mulher/es e do léxico tabu vagina e sinônimos, buscando investigar como foram utilizados no contexto de uma plataforma de vídeos, na internet. Assim, o corpus de estudo deste trabalho é composto por transcrições de episódios do videocast Mini Saia, publicados no canal da emissora GNT, no Youtube, sendo este um produto derivado do programa de TV Saia Justa. O corpus analisado é composto por episódios publicados entre os anos de 2020-2021 e seus respectivos comentários e abordam temas sobre os corpos femininos, feminilidades e feminismos. As participantes do programa buscam dialogar, expressando diferentes opiniões, compartilhando informações e relatando experiências. As transcrições foram realizadas com o auxílio do software Transkriptor. Neste estudo, realizamos uma análise fundamentada teórico-metodologicamente na Linguística de Corpus (LC) adotando seus princípios para a compilação do corpus, identificação, extração e análise dos dados (BERBER SARDINHA, 2009; PARODI, 2010; NOVODVORSKI, 2008). Foi utilizado o programa computacional WordSmith Tools, versão 6.0 (SCOTT, 2015) para identificação de padrões e análises, especialmente, as ferramentas “lista de palavras” e “lista de concordâncias”, assim como recursos em corpora disponíveis online para verificação de diferentes aspectos linguísticos. Esta análise também está ancorada à fundamentação teórica existente na área de Lexicologia (TAGNIN, 2013), Fraseologia (CORPAS PASTOR, 1996; 2010) e tabuísmos linguísticos (GUÉRIOS, 1979; PRETI, 2010; NOVODVORSKI E LIMA, 2020). Os resultados encontrados demonstram a criatividade lexical e apontam para um contexto propicio à utilização de tabus linguísticos, fraseologias da língua comum, inclusive aquelas do registro coloquial e vulgar e, também, demonstram que nesses contextos de utilização da língua, é frequente o uso de manipulações fraseológicas, disfemismos e eufemismos.

Palavras-chave: Mulheres; Léxico Tabu; Fraseologismos; Linguística de Corpus.

[32] Professora de Ensino Básico, Técnico e Tecnológico no Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro (IFTM), campus Ituiutaba, MG. Estudante de doutorado do programa de pós-graduação em Estudos Linguísticos (PPGEL) do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU), Uberlândia, MG. mayra@iftm.edu.br

O USO DE N-GRAMAS DE CLASSE SEMÂNTICA EM UM CORPUS DE APRENDIZ

Cristina Borges GIL[33]

O objetivo deste trabalho é analisar o uso de n-gramas de classe semântica (BERBER SARDINHA, 2023; RIBEIRO, 2023) na produção escrita e oral de aprendizes de inglês como língua estrangeira. Um n-grama de classe semântica (NGCS) é uma sequência contínua de classes semânticas, sendo que cada classe semântica expressa uma ideia ou conceito e representa uma palavra no texto (BERBER SARDINHA, 2023). Dito de outra forma, um NGCS reúne sequências de palavras que compartilham as mesmas categorias semânticas. Ou seja, as sequências de palavras abarcadas em um mesmo NGCS possuem sentido semelhante. Tal fato nos possibilita investigar uma proporção maior de agrupamentos lexicais no corpus (idem, ibidem). Com esta pesquisa, avaliamos se variação no uso dos n-gramas de classe semântica pode ser explicada pelo fato de o texto ser escrito ou falado, pela tarefa atribuída ao aprendiz, pelo seu nível de proficiência, pela sua língua materna, pela sua idade ou pelos anos de estudo do idioma inglês. O corpus empregado neste estudo foi o COREFL, cujo acrônimo significa Corpus de Inglês como Língua Estrangeira (Corpus of English as a Foreign Language), disponibilizado para a comunidade de modo gratuito por pesquisadores da Universidade de Granada, sob os termos da Creative Commons (LOZANO; DÍAZ NEGRILLO; CALLIES, 2020). Primeiramente, o corpus foi dividido em subgrupos organizados de acordo com a língua materna espanhol, alemão ou inglês. e então etiquetado com o USAS, um etiquetador semântico. Em seguida, foram extraídos e selecionados os n-gramas de classe semântica e calculada a sua chavicidade. Com essas variáveis foi feita uma análise fatorial, procedimento padrão da Análise Multidimensional (BIBER, 1988), e os fatores interpretados. Identificamos três dimensões: Dimensão 1. Cuidado, movimento, idade e interações sociais, Dimensão 2. Localização, deslocamento, idade, autoridade e emoção, Dimensão 3. Narrativa oral, marcadores de discurso, pausas preenchidas. Observamos que a tarefa e o modo desempenharam um papel importante na variação dos n-gramas de classe semântica utilizados pelos aprendizes.

Palavras-chave: Linguística de Corpus; Linguística de Corpus de Aprendiz; chavicidade; Análise Multidimensional; n-grama de classe semântica.

[33] Aluna de doutorado do Programa de Linguística Aplicada e Estudos da Linguagem PUC - SP, bolsista CAPES

CALIEMT: CORPUS DE APRENDIZES DA LÍNGUA INGLESA DO ENSINO MÉDIO TÉCNICO

Shirlene Bemfica de OLIVEIRA[34]
Lucas Renato dos Santos MURTA[35]
Jasper Vilan BRAGA[36]

O CALIEMT - Corpus de Aprendizes da Língua Inglesa do Ensino Médio Técnico é um corpus de amostras da produção oral e escrita de alunos do Ensino Médio Técnico que vem sendo coletado e compilado em aulas de inglês de um Instituto Federal no Estado de Minas Gerais. A proposta pedagógica traz uma abordagem heurística para o ensino de Inglês no âmbito da escola técnica integral onde os alunos discutem temas transversais em uma atmosfera de discussão e descoberta acadêmica e escrevem textos na língua inglesa, individualmente ou em coautoria com o uso de recursos tecnológicos (STORCH, 2005). A opção de propor a escrita em coautoria, especificamente em contextos de língua inglesa, baseia-se em aportes teóricos e pedagógicos focados em uma visão de aprendizagem socioconstrutivista e de letramentos críticos calcada nos trabalhos de Vygotsky (1978), Storch (2005) e Street (2014). Esta abordagem de Multiletramentos é promovida por meio de aulas que focam no desenvolvimento das habilidades integradas (reading, writing, listening, speaking), e na análise e produção de textos escritos de diversos registros multimodais que demonstram o posicionamento crítico dos alunos sobre temas transversais. A pesquisa visa tornar o ensino de idiomas mais orientado pelos dados (data-driven) e o processo de aprendizagem mais centrado no aluno que pode aprender por descoberta de padrões linguísticos, pelo uso do computador e suas ferramentas (Computer-Assisted Language Learning - CALL), enfatizando o pensamento crítico e a metacognição dos alunos. A compilação e organização do CALIEMT é embasada teórico e metodologicamente pela Linguística de Corpus (LC) que é a área do conhecimento que estuda a linguagem por meio da utilização do computador (BERBER-SARDINHA, 2000). Ela é definida como uma maneira de se chegar à linguagem por meio da análise dos padrões probabilísticos que se constroem nos contextos em que os falantes os empregam (BIBER et al., 1998; BERBER-SARDINHA, 2004). O projeto foi aprovado pelo Comitê Nacional de Ética na Pesquisa e todas as produções foram autorizadas pelos pais dos alunos por meio de termos de assentimento e de consentimento. A pesquisa tem grande impacto acadêmico, social, econômico e tecnológico na vida dos alunos participantes e o CALIEMT, resultado dessa pesquisa tem grande potencial inovador, uma vez que não existe no Brasil nenhum corpus eletrônico com amostras de alunos do Ensino Médio Técnico de escolas federais. Este vídeo tem como objetivo apresentar a organização do corpus, a interface de algumas

[34] Professora titular da Coordenadoria de Línguas Estrangeiras do IFMG - Campus Ouro Preto na área de Língua Inglesa, Pós-doutoranda no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais.

[35] Discente do Ensino Médio Técnico em administração do IFMG - Campus Ouro Preto, Ouro Preto, Minas Gerais. Bolsista PIBIF Jr. IFMG.

[36] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista FAPEMIG.

plataformas e os recursos online disponíveis para a compilação do CALIEMT, demonstrando a escolha dos metadados, as possibilidades de busca, os processos e as ferramentas envolvidas para a limpeza, preparação e anotação para tornar o corpus adequado para consultas linguísticas (ALUÍSIO, et. al., 2006; GONZÁLES, 2007). Os resultados podem contribuir para as pesquisas nas áreas de Educação, Linguística de Corpus e Linguística Aplicada, pois o corpus em um servidor próprio divulgado a comunidade científica externa tem grande potencial social e impacto tecnológico.

Palavras-chave: corpus de aprendizes; ensino médio técnico; língua inglesa; compilação; ensino de línguas.

NOMINALIZATION: CORPUS-BASED STUDY OF DISCUSSION SECTIONS IN FORESTRY RESEARCH ARTICLES

Shirlene Bemfica de OLIVEIRA[37]
Deise Prina DUTRA[38]
Jasper Vilan BRAGA[39]
Camila Alves RAMOS[40]
Ana Clara Taborda de Paula VICTOR[41]

The dissemination of scientific knowledge in Brazilian academic journals is predominantly conveyed through research articles (RAs) in Agricultural Sciences. However, there is a tendency for internal publishing which may influence the impact factor of national research. The low rate of international publishing may be related to the proficiency level of researchers when writing in English. Academic writing poses a challenge for Brazilian investigators and it is characterized by specific sections (abstract, introduction, methods, results, discussion, references) and linguistic features (lexical sophistication, syntactic complexity, text cohesion) which can be indicators of text quality and publishing acceptance (CROSSLEY, 2020). There has been extensive interest in RA investigations from different perspectives of form and function due to their value in generating and distributing new knowledge in the academic community (JALILIFAR, 2017). Nevertheless, few studies emphasize the separate sections of RAs even though they have shown important section variation. Such investigations generally focus on specialized corpora and terminologies, lexical combinations, and rhetorical strategies (CROSSLEY, 2020). Data discussion is the most demanding section for writers in academic writing, and variation in the noun phrase structures in this particular portion of RAs, across different disciplines, is productive. Furthermore, few corpus-based studies have examined the noun phrase structure of RAs' discussion sections, especially ones that describe the nominalization process. This study, therefore, focuses on the linguistic features of RAs, particularly the noun phrase structure in discussion sections, which are crucial for conveying complex information. It provides an overview of the noun phrase structure of discussion sections of Forestry RAs in high-impact journals and analyses the nominalization process in those sections. 90 RAs from nine high-impact Forestry journals were analysed and Sketch Engine was used for data exploration. Findings revealed that discussion sections

[37] Professora titular da Coordenadoria de Línguas Estrangeiras do IFMG - Campus Ouro Preto na área de Língua Inglesa, Pós-doutoranda no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais.

[38] Professora titular na Faculdade de Letras da UFMG na área de Língua Inglesa e no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais.

[39] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista FAPEMIG.

[40] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista CNPq.

[41] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista CNPq.

in Agrarian Science RAs are predominantly written with the use of embedded noun phrases centered by nominalization suffixes which convert an action expressed by a verb into a noun, typically to refer to general statements and processes as well as to explain them or to treat actions and processes as objects separated from human participants, increasing articles impersonality. Moreover, data shows that the high frequency of this pattern in the discussion sections indicates a high level of information density, abstraction, and grammatical complexity through a specialized pattern of information packaging (JALILIFAR, 2017). The latter suggests abstraction and grammatical complexity, which may be text quality and acceptance indicators for publication. This study is part of ongoing research that aims to improve academic writing produced in English by Brazilian researchers in Agrarian Sciences and assist in the international dissemination of Brazilian scientific production. The data gathered from this research will be instrumental in enhancing the authors' academic writing skills in Agrarian Sciences. By identifying effective noun phrase structures and the nominalization processes prevalent in successful RAs, we can develop targeted workshops and resources that guide researchers in improving their writing proficiency in English. These insights will make authors aware of their writing process, and promote greater acceptance and visibility of Brazilian research. Ultimately, this initiative aims to foster a stronger global presence for Brazilian scientific contributions.

Palavras-chave: English language; academic writing; research articles; nominalization; Agrarian Science corpus.

DICIPLINARY DIFFERENCES IN FORMULATION AND PRESENTATION OS RESERACH QUESTION, HYPOTHESES AND OBJECTIVES IN INTRODUCTIONS: INSIGHTS FROM SOCIAL SCIENCES AND HUMANITIES

Anna Clara TABORDA de Paula Victor[42]
Ana Eliza Pereira BOCORNY[43]
Deise Prina DUTRA[44]
Gustavo Leal TEIXEIRA[45]
Shirlene BEMFICA de Oliveira[46]

Biber (2006) demonstrates that there is language variation across different registers in academic settings, showing some distinct linguistic features that characterize the writing of various disciplines. On the genre analysis perspective, Swales' (1990) work explores the academic writing conventions. The differences in the formulation and presentation of research questions, hypotheses, and objectives across disciplines can be understood through the light of genre and register analysis based on corpus linguistics (Biber & Conrad, 2009). This study focuses on gathering information about how different disciplines present their own research in articles. This poster, specifically, presents the investigation of the linguistic and structural differences in the presentation of the terms “research questions”, “hypotheses”, and “objectives” in article introductions across disciplines within the fields of Social Sciences (SSci), Human Sciences (HSci), and Language, Linguistics, and Arts (LLArts). We used Antconc (ANTHONY, 2024) to explore our 1600 text corpus (1,573,549 tokens) with 100 texts sampled from each of the sixteen disciplines: Law and Legal Sciences, Communication, Demography, Economy (SSci), Archeology, Anthropology, Education, Geography, Public Policies, Psychology, Political Science, Religious Faiths, Sociology, Philosophy (HSci), and Linguistics and Languages (LLArts). The frequency counts of the targeted terms were normalized per thousand words for comparison across texts (BIBER, 1988). AntConc generated the concordance lines and identified typical lexical structures used to express research questions, hypotheses, and objectives. Our findings reveal distinct disciplinary conventions, indicating how research questions, hypotheses, and objectives are articulated in introductions. For example, in the Human Sciences (HSci), the term "Hypothesis" showed a high frequency, with ≈0.68 per thousand, especially in Political Science (≈0.47 in Religion and ≈0.34 per thousand in Psychology). On the other hand, in disciplines such as Education and Philosophy, the frequency of the term "Objective" stood out more with ≈0.24 and ≈0.21. In the Social Sciences (SSci), "Hypothesis" appeared at ≈0.41 per thousand in Economics. Meanwhile, in

[42] Graduanda – Bolsista CNPQ (420180/2022-2) - Universidade Federal de Minas Gerais, Belo Horizonte – MG

[43] Professora - Universidade Federal do Rio Grande do Sul, Porto Alegre - RS

[44] Professora - Universidade Federal de Minas Gerais, Belo Horizonte - MG

[45] Professor, Universidade Federal de Minas Gerais - Montes Claros, Minas Gerais – MG

[46] Professora titular do IFMG - Campus Ouro Preto, Pós-doutoranda no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais-MG.

Linguistics, Literature, and Arts (LLArts), the term "Hypothesis" had a significant presence in Linguistics, with ≈0.8 per thousand, indicating a different methodological emphasis compared to other fields. In conclusion, the empirical nature of SSci research favors more structured and repetitive phrasing while LLArts tend to employ a more implicit use of these expressions, reflecting the diverse nature of writing conventions in these fields. Our results have pedagogical implications and will inform MOOC activities as part of the CNPq project this paper is part of. For instance, in disciplines, such as Political Science, Religious Faiths, and Psychology, the pedagogical focus should be on how to formulate hypotheses. In disciplines such as Education and Philosophy, the teaching might focus on how to define clear and achievable objectives (e.g. the use of the word “objective”), while for LLArts, presenting the use of the word "hypothesis" in various articles would be more relevant. These variations between disciplines reflect distinct methodological traditions and emphasize the need for customized pedagogical approaches to support the formulation of research questions, hypotheses, and objectives in each discipline.

Palavras-chave: Research Questions, Hypotheses, Objectives, Social Sciences, Humanities.

EXPLORING MODAL VERB USAGE IN AGRARIAN SCIENCES RESEARCH ARTICLES: A CORPUS-BASED ANALYSIS

Camila Alves RAMOS[47]
Deise Prina DUTRA[48]
Gustavo Leal TEIXEIRA[49]
Shirlene Bemfica de OLIVEIRA[50]
Carolina Godoi de Faria MARQUES[51]

Modal verbs play an important role in academic writing, offering a way to convey stance and nuance (Biber et al., 2021). However, their specific usage in Agrarian Science Research Articles (RAs) remains underexplored. This study seeks to uncover how modals are employed by researchers in Agrarian Sciences to express stance in their papers, recognizing that modals are “by far the most common grammatical device used to mark stance in university registers” (Biber, 2006, p. 103). Following the Longman Grammar of Spoken and Written English (Biber et al., 2021), modal verbs are categorized by: possibility/permission/ability, necessity/obligation, and prediction/volition. These categories show logical possibility or predictions, rarely indicating personal agency and, from this perspective, this study compared abstract, introduction, method, results, discussion and conclusion RA sections. The CorAgrarian corpus, selected for analysis, consists of 447 academic articles from five sub-areas within Agrarian Sciences. The five areas are Agriculture Engineering, Agronomy, Animal Sciences, Food Engineering, and Forestry. These research articles were compiled and chosen from specialized journals to represent a wide range of research topics within the field. All sub-corpora were uploaded on Sketch Engine. The tag for modals (MD) was counted in each sub-corpus to determine their frequency and section prevalence. Due to large differences in word count among sections, all extracted data was normalized by 1000 words. The analysis revealed distinct patterns in modal use across RA sections. Abstract sections presented a low-frequency use of modals with a normalized frequency of 4.90. Introduction sections showed a normalized frequency of 7.58. Method sections exhibited the lowest frequency of modals with a normalized frequency of 2.02. Results sections also showed a lower normalized frequency of 3.32. In contrast, the Discussion sections displayed a prominent modal use, with a normalized frequency of 10,77. Finally, the conclusion section had the highest modal frequency, with a normalized frequency of 13,31. These findings show that modals are most frequent in the conclusion section (e.g. “... such policies should be continuously promoted and extended …”), followed by the discussion (e.g. “the comprehension rates could not be considered as sufficient”) and introduction sections (e.g. “...,

[47] Aluna de graduação, UFMG, Belo Horizonte - MG, bolsista institucional PIBIC/CNPq

[48] Professora - Universidade Federal de Minas Gerais, Belo Horizonte - MG

[49] Professor, Universidade Federal de Minas Gerais - Montes Claros, Minas Gerais - MG

[50] Professora - Pós-Doutoranda. Filiação: Instituto Federal de Minas Gerais Ouro Preto - MG e Universidade Federal de Minas Gerais, Belo Horizonte - MG

[51] Doutoranda, Universidade Federal de Minas Gerais, Belo Horizonte/MG. Bolsista CAPES (n. 88887.939578/2024-00)

the varied optimistic and pessimistic versions must be contrasted.”). Following, each modal was also examined and categorized according to its function, revealing that over 50% of the modals fell into the "possibility/permission/ability" category, making it the most prevalent. According to Liu and Xiao (2022, p. 47), conclusions enable authors to emphasize their research results, highlight contributions, and suggest future directions. These communicative purposes align with modal functions, explaining their high frequency. By exploring modal frequency and distribution across different RA sections, this study deepens the understanding of stance in Agrarian Sciences academic writing. This research aims to utilize these insights to develop teaching materials that enhance academic writing skills, particularly in using modals to convey stance and engage with research findings.

Palavras-chave: modal verbs; academic writing; Agrarian Sciences; research articles; Corpus Linguistics.

O USO DE *PRESENT SIMPLE, PRESENT PERFECT, PAST SIMPLE E PAST PERFECT* NAS INTRODUÇÕES DE ARTIGOS CIENTÍFICOS, TESES E DISSERTAÇÕES ESCRITOS EM INGLÊS NA ÁREA DE CIÊNCIAS AGRÁRIAS

Jasper Vilan BRAGA[52]
Carolina Godoi de Faria Marques[53]
Deise Prina Dutra[54]
Gustavo Leal Teixeira[55]
Shirlene Bemfica de Oliveira[56]

As Ciências Agrárias são uma área importante para o desenvolvimento nacional, com um grande volume de produção acadêmica. Entretanto, essas pesquisas apresentam pouco alcance internacional, como consequência de um déficit de publicações em inglês. Para auxiliar os pesquisadores brasileiros da área a dominarem a escrita acadêmica em inglês, aumentando suas chances de publicação internacional e obtenção de um maior alcance das suas pesquisas, são necessários estudos que descrevam a escrita acadêmica em inglês dessa área. No entanto, conforme foi descrito por Shi e Wannaruk (2014) são poucos os estudos a esse respeito. De forma a contribuir com esse cenário este trabalho se propõe a analisar as estruturas verbais de past e present tense utilizadas na introdução de produções acadêmicas dessa área. Segundo Swales e Feak (2012), a introdução de artigos acadêmicos visa estabelecer o espaço da pesquisa, apresentando e contextualizando o estudo realizado, assim como seu tema, objetivos e motivações. Na introdução essas funções são geralmente assistidas pelo present simple, o present perfect e o simple past (Swales e Feak, 2012; Biber et al. 2021). Ademais, pesquisas constataram que as formas verbais: present simple, present perfect, past simple e past perfect apresentam maior frequência de uso na introdução quando comparada com as demais seções dos artigos acadêmicos, quais sejam: resumo, metodologia, resultados e discussão e conclusão (Berber Sardinha et al., no prelo). Diante do exposto, hipotetiza-se que elas tenham uma frequência significativa nessa seção também nas produções acadêmicas de Ciências Agrárias. Para realização deste trabalho, utilizamos os corpora CorAgrarian e CorAgrSc. O primeiro é um corpus de artigos científicos publicados em inglês em revistas de alto fator de impacto (A1) com 447 textos , totalizando 2.532.420 palavras, representativo das seguintes subáreas das Ciências Agrárias: Engenharia Agrícola, Agronomia, Zootecnia, Engenharia de Alimentos e Engenharia Florestal. O segundo, por sua vez, é um corpus de teses e dissertações de programas de pós-graduação brasileiros escritos em inglês, com 26 textos, totalizando 89.036 palavras, contendo textos das mesmas subáreas que o primeiro. Para a realização deste estudo foi utilizado o Sketch Engine para anotar os corpora e, para realizar as análises

[52] Aluno da Graduação, UFMG, Bolsista FAPEMIG (APQ-01173-22).
[53] UFMG
[54] UFMG
[55] UFMG
[56] UFMG

linguísticas, sua ferramenta Colocate. Visando identificar quais formas verbais são mais frequentes e seu uso nas introduções das produções acadêmicas de Ciências Agrárias, foi realizada uma a busca, por CQL, das ocorrências de present perfect, present simple, simple past e past perfect em cada corpora. Os resultados preliminares indicam que nas introduções, seja tanto dos artigos científicos quanto das teses e dissertações, as formas verbais simples tanto no passado quanto no presente apresentam uma frequência elevada em relação às demais seções. O present perfect também ocorre significativamente na introdução em todos os corpora, entretanto de forma pontual, referenciando pesquisas anteriormente realizadas. Já, o past perfect apresenta poucas ocorrências na introdução quando comparado com as demais formas verbais analisadas. Trata-se de uma pesquisa em andamento com o objetivo de auxiliar a difusão internacional da pesquisa brasileira da área de Ciências Agrárias.

Palavras-chave: linguística de corpus; Ciências Agrárias; inglês acadêmico; formas verbais; introdução.

AUTOMATIZAÇÃO COM INTELIGÊNCIA ARTIFICIAL DA EXTRAÇÃO E CLASSIFICAÇÃO DE LEXICAL FRAMES E LEXICAL BUNDLES PARA ANÁLISE DE ARTIGOS ACADÊMICOS

Simone OLIVEIRA[57]
Ana Eliza Pereira BOCORNY[58]
Júlia TAMAGNO[59]
Pedro FERNANDES[60]
Tony Berber SARDINHA[61]

A pesquisa proposta se insere no contexto do projeto geral intitulado “A internacionalização da produção científica brasileira em Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes fomentada por recursos pedagógicos on-line baseados em corpus”, coordenado pelo Professor Dr Tony Berber Sardinha. A sua investigação tem o intuito analisar artigos científicos brasileiros do período de 2013 a 2023 de produções publicados na língua inglesa em revistas internacionais de alto impacto. Para colaborar com as investigações do projeto geral, essa pesquisa visa automatizar os processos de extrair, agrupar e categorizar dados linguísticos dos artigos utilizando técnicas de Processamento de Linguagem Natural (PLN) e Inteligência Generativa. O intuito é desenvolver um sistema automatizado para agilizar a análise de dados linguísticos a partir de corpora selecionados. Com isso, será necessário buscar técnicas de PLN adequadas, construir um modelo, criar um Produto Viável Mínimo (MVP) para automatizar os processos de extração, limpeza, categorização e armazenamento de dados em escala, com menor tempo e assertividade no processo por meio de padronizações. A fundamentação teórica está estruturada em Biber (2009) e Gray e Biber (2013) que propõe duas metodologias para extração de Estruturas Lexicais (ELs). O estudo de ELs permite identificar padrões na linguagem e compreender como os autores constroem seus textos. Esse processo com uso de PLN contribuirá significativamente para a análise de corpus a partir das classificações de textos e dos modelos de similaridade, oportunizando expandir a análise da Linguística de Corpus, mesmo com corpora extremamente grandes (DUNN, 2022). O método de investigação científica mais apropriado para esse contexto é o método de pesquisa aplicada, utilizando uma abordagem quantitativa com técnicas de análise computacional e experimentação. Será necessária a organização do corpora divididos por áreas, contendo subcorpora de 1 milhão de palavras cada. A primeira ação será a extração, depois o agrupamento de pacotes lexicais por similaridade e sentido. No processo de agrupamento de nGramas, utilizaremos o modelo Sentence-BERT (BIRD, 2024). Em seguida, aplicaremos a técnica de clusterização DBSCAN, que é capaz de agrupar frases com base em sua similaridade sem exigir um número predefinido de clusters.

[57] Bolsista CNPQ (Chamada CNPq/MCTI/FNDCT No 40/2022)
[58] UFRGS
[59] UFRGS
[60] UFRGS
[61] PUCSP

Utilizando a métrica de similaridade de cosseno para calcular a proximidade entre os embeddings, o algoritmo DBSCAN formará clusters de nGramas com significado semântico similar. Esta abordagem nos permite identificar grupos de expressões semelhantes, enquanto separa os outliers (ruído), que serão excluídos da análise principal. Alguns resultados preliminares podem ser observados, como um estudo comparativo entre processos manuais e automatizados de dados já coletados no ano anterior. A relevância do estudo reside na criação de um sistema inovador de Inteligência Artificial para agilizar o tempo e melhorar os resultados no trabalho em escala de grandes grupos de corpora para extrair, agrupar e categorizar. O projeto buscará não apenas avançar na análise linguística computacional, mas também contribuir significativamente para área, permitindo uma compreensão mais profunda da produção científica brasileira em menos tempo, oferecendo insights e ferramentas que possam apoiar na elaboração e publicação de futuros trabalhos acadêmicos.

Palavras-chave: corpora; quadros e pacotes Lexicais; automação com inteligência artificial; extração; agrupamentos.

ANÁLISE MULTIDIMENSIONAL LEXICAL EM CORPORA DE RESENHAS E VIDEORRESENHAS ONLINE: UMA ABORDAGEM DA LINGUÍSTICA DE CORPUS COMO ÁREA AUTÔNOMA DE PESQUISA CIENTÍFICA

Mauricio José Ferreira LOPES[62]

Este estudo investiga as variações léxico-discursivas em corpora de resenhas escritas e videorresenhas literárias, produzidas por influenciadores digitais literários (IDLs) nas plataformas Instagram e YouTube. A pesquisa posiciona a Linguística de Corpus (LC) como uma área de investigação científica autônoma, utilizando a Análise Multidimensional (AMD) para identificar padrões linguísticos em registros distintos, com base nos métodos de Biber (1988) e Berber Sardinha (2000). Para além da análise quantitativa oferecida pela AMD, a Análise do Discurso (AD) de Pêcheux é incorporada ao estudo, a fim de interpretar as práticas discursivas, observando-se suas formações ideológicas e sociais, conforme abordado por Pêcheux (2010). A intersecção entre LC e AD permite uma análise integrada dos registros, revelando como as práticas discursivas refletem contextos sociais e como as dimensões discursivas emergentes mostram formações ideológicas subjacentes aos discursos dos influenciadores. Além disso, o uso de técnicas de Inteligência Artificial (IA) possibilita uma análise mais sofisticada de grandes volumes de dados linguísticos, oferecendo novas oportunidades para investigar os discursos produzidos em plataformas digitais, como observado por Silva (2019). O estudo examina como influenciadores literários configuram suas práticas discursivas de acordo com o público-alvo e as características das diferentes plataformas. Resenhas publicadas no Instagram tendem a apresentar uma abordagem mais introspectiva e analítica, enquanto as videorresenhas no YouTube enfatizam a comunicação direta e a interação com o público. A combinação entre LC e IA permite uma análise mais precisa de gêneros e subgêneros literários, oferecendo insights valiosos sobre as dinâmicas discursivas em plataformas digitais. A pesquisa destaca o papel fundamental da LC como ciência autônoma e interdisciplinar, que fornece uma compreensão crítica das práticas discursivas contemporâneas e suas implicações sociais e ideológicas. O estudo, assim, contribui para a consolidação da LC como uma área científica que dialoga com outras disciplinas, ampliando o escopo da análise linguística em contextos digitais e colaborando para a formação de comunidades discursivas online e a disseminação do conhecimento literário.

Palavras-chave: Palavras-chave: Linguística de Corpus; análise multidimensional; práticas discursivas; redes sociais; resenhas literárias.

[62] Professor de Língua Estrangeira na rede pública municipal de São Paulo. Mestre e doutorando em Linguística Aplicada e Estudos da Linguagem pela PUC-SP, bolsista CAPES. Email: mauricio.lopes@sme.prefeitura.sp.gov.br

EAT THE FROG: USING GENERATIVE MODELS TO AID IN THE CORPUS-BASED IDENTIFICATION OF METAPHORS IN MULTILINGUAL TWEETS

Anna Beatriz Dimas FURTADO[63]
Anne O'CONNOR[64]

Recent technological advancements gave rise to new means of communication especially valuable and profitable in the Information Society: social media. First proposed to bridge the distance between people, social media became essential for many institutions as a means of bringing their followers close. The Catholic Church is not different; the Pope has been using Twitter since 2012 to discuss a wide range of topics, from climate change to daily religious practices. The @pontifex accounts are a case of tremendous success, reaching more than 1 million followers in daily basis. Such multilingual practice is underpinned by a large-scale translation endeavour in more than thirty languages. Indeed, Corpus Linguistics has revolutionized the study of language, especially translation, enabling the identification and description of patterns across multiple languages, textual features, and several domains (O'KEEFFE and MCCARTHY, 2022). An interesting feature of human language is the pervasive use of metaphorical language, especially on the religious domain (DORST, 2021). Notably, corpus-linguistics techniques have been efficiently employed in the identification and description of metaphors (STEFANOWITSCH 2006, 2020; TISSARI, 2017). However, even using corpus exploration tools, the identification of metaphors in a big-size corpus in multiple languages can be rather time-consuming and labour-intensive.The automatic treatment of metaphors with natural language processing is not a new task. It has been investigated through several subtasks: metaphor identification (MAO et al., 2019), metaphor interpretation (SHUTOVA, 2010), conceptual mappings (ROSEN, 2018). While several models have been tested as shown in the survey by Tong et al. (2019), results show that this task remains quite challenging. One of the reasons for this is the notion of metaphor itself. Therefore, we seek to investigate whether the use of recent generative chatbot models can aid in the identification of metaphors (defined as the result of the mapping between conceptual domains as in Lakoff and Johnson (1980)) on a ten-year corpus (2012-2022) comprising 35,684 tweets (619,984 words) in seven languages (Arabic, English, Italian, French, Spanish, Portuguese). For this ongoing case study, we subsampled the corpus into a 500-tweet sample to facilitate manual analysis. We employed ChatGPT (OPENAI, 2023) and Gemini (GOOGLE, 2023) to perform metaphor detection and source-and-target domain identification in English, Portuguese, Spanish, French, Italian and Arabic. We compare automatic results with corpus-based results by employing WMatrix

63 Research Assistant in the Institute for Creative Technologies, University of Galway, Ireland, funded by PIETRA Project Consolidator Grant No. 101001478, European Research Council

64 Full Professor in the School of Languages, Literatures, and Cultures, University of Galway, Ireland

(RAYSON, 2008) to extract key semantic domains and their corresponding keywords so that source-and-target domains can also be identified and recorded.

Our results show that metaphor detection is far from solved either by chatbots or corpus-based methods. While detecting key domains with corpus-based methods is more reliable, the task depends heavily on the quality of the reference corpus tagged for semantic fields. Although Gemini and ChatGPT can both be used to identify crystalised metaphors (65% in English and 40% in Portuguese), hallucinations are still pervasive in the source-and-target domain identification. The best approach is then, to combine both methods to facilitate the identification of metaphors.

Palavras-chave: metaphor identification; conceptual domain identification; corpus-based metaphor studies;

LINGUÍSTICA DE CORPUS E ACESSIBILIDADE: INTERFACES ENTRE CORPORA E SIMPLIFICAÇÃO TEXTUAL

Bruna Rodrigues da SILVA[65]

Este trabalho apresenta recorte, sob o viés da Linguística de Corpus, de pesquisa de Doutorado, que se insere nos estudos de Acessibilidade Textual e Terminológica (ATT). A pesquisa como um todo busca a união da experiência docente com a pesquisa acadêmica, por meio da investigação da leitura e da compreensão de materiais, em tese, adaptados para um público com doze anos ou mais, por jovens e adolescentes do Ensino Fundamental II de escola pública de Porto Alegre-RS. O objetivo principal do trabalho como um todo é descrever e analisar se um livro da área da saúde, disponível on-line, adaptado para um público leitor jovem, é compreendido por esse público e de que forma. Inicialmente, o foco será a publicação digital Aprendendo sobre vírus e vacinas, da Editora da UFCSPA. Essa editora lançou várias publicações, todas na área da saúde, adaptadas para diferentes públicos. A única dessas obras direcionada para público jovem, com doze anos ou mais, foi escolhida para análise neste estudo porque essa é a faixa etária que corresponde aos alunos da pesquisadora responsável, com os quais será possível dar continuidade à pesquisa, num próximo momento, por meio de testes de compreensão leitora. O recorte que se apresenta neste resumo faz parte do momento inicial do estudo, em que, com apoio da estatística linguística (BIDERMAN, 1978, 1998) e da Linguística de Corpus (BERBER SARDINHA, 2004), serão realizados contrastes do corpus de estudo com outros corpora. A fim de constatar possíveis diferenças de vocabulário escrito, o corpus selecionado para contraste foi a publicação digital Somos Heróis – os cuidados para o coronavírus ir embora, de Pedro Leite. Essa publicação foi seleciona porque tem vários pontos em comum com o corpus de estudo: o fato de ser adaptado; a faixa etária destinada; o acesso livre, disponível para download e a gratuidade; o formato digital; e a temática do COVID-19. A comparação será feita com o auxílio do software AntConc (ANTHONY, 2019). Esse é um software de acesso livre que contém ferramentas para gerar dados estatísticos, a partir de um texto em formato digital. Os contrastes iniciais indicam que cerca de 22% do vocabulário do corpus de estudo coincide com o vocabulário do corpus de contraste em questão. Porém, palavras como, por exemplo, analgésico, adsorver, ancorada, atenuado, papiloma, partículas, pneumocócica e proliferam, entre outras, fazem parte das diferenças entre os corpora. Assim como essas palavras, outros pontos de divergência surgirão dessas comparações, merecendo atenção, pois podem servir de base para os testes de compreensão leitora a serem realizados com os alunos nas etapas subsequentes da pesquisa. Além disso, tais contrastes também vão enriquecer a análise e a discussão sobre a acessibilidade desse material para esse público, servindo de base para o estudo como um todo.

---

[65] Doutoranda pelo PPG-Letras/UFRGS, professora da rede pública de ensino, Porto Alegre – RS

Palavras-chave: Acessibilidade; Linguística de Corpus; corpus; contraste; Linguística Computacional.

DESENVOLVIMENTO DE UMA METODOLOGIA E APRIMORAMENTOS DE RECURSOS LEXICOGRÁFICOS PARA UMA PLATAFORMA DE DICIONÁRIOS DE COLOCAÇÕES ACADÊMICAS EM PORTUGUÊS E INGLÊS

Adriane ORENHA-OTTAIANO[66]
Tanara Zingano KUHN[67]
Stella Esther Ortweiller TAGNIN[68]
Giseli Aparecida CECÍLIO[69]
Cristiane Krause KILIAN[70]

O objetivo do presente trabalho é apresentar a metodologia usada para identificar colocações acadêmicas em um corpus acadêmico de português no âmbito do projeto Dicionários Online de Colocações Acadêmicas. Embora as colocações acadêmicas tenham recebido considerável atenção nos últimos anos, revisão da literatura indica a existência de diferentes formas de entendimento acerca do que estas são (por exemplo, DURRANT, 2009; PAQUOT, 2010; ACKERMANN; CHEN, 2013). Com base nessas abordagens, e tendo em vista um posicionamento crítico, que considera também reflexões sobre vocabulário acadêmico, nosso entendimento sobre colocações acadêmicas considera dois níveis, quais seja, estatístico e fraseológico. Sob uma abordagem estatisticamente orientada, vemos as colocações acadêmicas como combinações frequentes de palavras em textos acadêmicos, cuja coocorrência é estatisticamente maior do que o esperado em comparação a quaisquer outras palavras combinadas aleatoriamente em uma língua específica e em um campo específico de conhecimento. Sob uma abordagem fraseológica, as colocações acadêmicas são combinações de palavras que são recorrentes e convencionalizadas em textos acadêmicos, que podem ter assumido um significado diferente ou novo daqueles usados na linguagem não acadêmica, podendo variar entre disciplinas. Uma vez definido nosso entendimento acerca das colocações acadêmicas, a metodologia adotada neste trabalho está estruturada nos seguintes passos. Primeiramente, criaremos uma lista de 50 palavras lexicais que ocorram com uma frequência mínima de 3em, no mínimo, 4 grandes áreas do subcorpus Brasil do CoPEP (KUHN; FERREIRA, 2020). O ponto de corte para a frequência mínima de ocorrência e o número mínimo de dispersão ainda serão definidos, uma vez que não há consenso nos estudos revisados. A seguir, serão extraídos automaticamente candidatos a colocações acadêmicas do corpus anotado com o UDPIPE. As colocações serão aquelas automaticamente calculadas pela função Word Sketch do Sketch Engine. Por fim, os candidatos serão importados para o Dictionary Writing System (ORENHA-

[66] Professora Associada do Programa de Pós-Graduação em Estudos Linguísticos, da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo

[67] CELGA-ILTEC, Universidade de Coimbra, Portugal

[68] Universidade de São Paulo

[69] Aluna de Doutorado do Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP)

[70] Instituto Superior de Educação Ivoti

OTTAIANO ET AL., 2021; ORENHA-OTTAIANO et al., 2023), para que as colocações que integrarão o dicionário sejam selecionadas. Para tanto, os lexicógrafos deverão seguir critérios discutidos e definidos pelos membros do projeto, sob uma perspectiva fraseológica, tendo em vista ainda o usuário final do dicionário.

Palavras-chave: colocações acadêmicas; dicionário de colocações; plataforma de dicionários; recursos lexicográficos

ANÁLISE COMPARATIVA DE FERRAMENTAS DE EXTRAÇÃO TERMINOLÓGICA AUTOMÁTICAS E SEMIAUTOMÁTICAS

Helena Cid TELES OLIVEIRA[71]
Elisa Duarte TEIXEIRA[72]

A prática da tradução especializada está diretamente associada ao uso de terminologia. Teixeira (2008) propõe que unidades linguísticas que são copiadas/repetidas/imitadas/mimetizadas no texto, sejam denominadas "Unidades de Tradução Especializada" (UTEs). Mas como identificar UTEs nos textos de partida e chegada, em grandes coletâneas de textos sobre o mesmo tema de forma rápida e minimamente confiável, principalmente, com a ampliação dos meios de tecnologia e o grande volume de informações? O avanço acelerado da tecnologia ocorrido nas últimas décadas impactou veementemente o trabalho terminológico (SARDINHA, 2004), o que desencadeou a oferta de ferramentas informatizadas que facilitem o processamento de tamanho volume de dados digitalizados e de corpora eletrônicos. Ferramentas de extração automática e semiautomática identificam candidatos a termo – trabalho antes feito manualmente – com busca de equivalentes em seus textos de partida e em corpora de apoio à tradução, processo que auxilia na criação de dicionários e glossários especializados (BOWKER, 2015). No entanto, há poucas pesquisas sobre quais ferramentas de extração terminológica estão disponíveis atualmente para uso por tradutores e terminógrafos, bem como sobre seus custos, a facilidade de uso, o tipo de input requerido, o nível de eficiência dos resultados, entre outros fatores que poderiam auxiliar na decisão de usá-las ou não, e qual delas escolher. O objetivo deste trabalho foi realizar o levantamento de algumas destas ferramentas e testá-las a fim de contribuir com a comunidade científica de terminologia e o trabalho tradutório. As duas que obtiveram os melhores resultados, de acordo com os critérios definidos, foram a Termostat (DROUIN, 2010) e a Sketch Engine (KILGARRIFF et al, 2003). A depender da demanda, as ferramentas podem vir a atender o tradutor, e mais ainda se o profissional dispuser de recursos para investir nessas ferramentas.

Palavras-chave: tradução especializada; terminologia para tradução; extratores automáticos e semiautomáticos de terminologia; ferramentas de auxílio à tradução

[71] Aluna de graduação em Letras Tradução Inglês da UnB

[72] Docente do Departamento de Línguas Estrangeiras e Tradução (LET) da UnB

ANÁLISE DE ATRIBUTOS-CHAVE FOR DUMMIES: O INÍCIO DE UM MANUAL

Carolina BOHORQUEZ[73]

A pesquisa em LC acerca da escrita acadêmica é capaz de produzir resultados que podem auxiliar no ensino dessa habilidade, principalmente através da comparação das características de diferentes registros (BIBER; CONRAD, 2009). Três principais metodologias podem cumprir esse objetivo: a Análise Multidimensional (AMD) (BIBER, 1988), a Análise de palavras-chave (SCOTT, 1997) e a Análise de atributos-chave (EGBERT; BIBER, 2023). Essa última, mais recente e menos complexa, envolve itens funcionalmente relevantes a um registro e conta com cálculos estatísticos refinados. Alunos das áreas de Ciências Humanas, ao se depararem com trabalhos ricos em cálculos e análises estatísticas, sentem-se receosos e muitas vezes optam por não utilizar aquela metodologia. Um estudo realizado em 2002 concluiu que esses alunos apresentam atitudes mais negativas em relação às disciplinas de matemática e estatística (SILVA et al., 2002). O estudo enfatiza também que quanto mais o aluno compreende os conceitos básicos dessas áreas, maior será a tentativa de aproximação das mesmas. Uma vez que a LC debruça-se sobre dados numéricos, a estatística é essencial para que se possa trabalhar eficazmente com informação quantitativa (BREZINA, 2018). Este trabalho pretende, portanto, desenvolver um manual que possa auxiliar alunos de Linguística a aprenderem a utilizar a metodologia de análise de atributos-chave em suas pesquisas. O manual tem o objetivo de detalhar o passo a passo da metodologia apresentada no trabalho de Biber & Egbert (2023); apresentar exemplos de programas que possam realizar as tarefas envolvidas; demonstrar uma aplicação manual da metodologia para que ela possa ser compreendida e para que, futuramente, um script possa ser desenvolvido com o intuito de automatizar o processo; correlacionar literatura relevante; descrever um exemplo de análise de atributos-chave aplicada no âmbito da escrita acadêmica contrastando introduções de artigos científicos e introduções de teses e, por fim; explicar os cálculos estatísticos presentes na metodologia. Neste estudo, um atributo escolhido no exemplo de análise foi o tamanho das palavras. Percebeu-se que ele se dá predominantemente nas introduções de artigos. Biber (1988) argumenta que quanto maior o tamanho da palavra, maior é o peso ou densidade informacional. O uso de palavras maiores são empregadas para expressar que o texto se caracteriza por ser um foco na informação (KITJAROENPAIBOON, W. et al., 2023). Notou-se que palavras como productivity, consolidation, agricultural, effectiveness, production e security estão presentes nas introduções de artigos, enquanto que as introduções de teses apresentaram menos palavras desta natureza . Esse resultado confirma a hipótese de que introduções de artigos, por serem menores, precisam condensar as informações de maneira eficaz para que sua função de incorporar o tema principal ao estudo realizado seja executada. Diante dos resultados, acredita-se que um manual que contivesse os detalhes da metodologia de análise de atributos-chave seria extremamente útil para

[73] Mestre em Linguistica Aplicada, Universidade Federal de Minas Gerais.

alunos que pretendem adotá-la. Aplicando-se a metodologia manualmente, detalhes importantes de busca foram revelados e pretende-se listar todas as soluções e casos problemáticos na versão definitiva. Os resultados do exemplo de análise deste trabalho podem ser desdobrados e auxiliar na produção de atividades didáticas para o aluno de escrita acadêmica.

Palavras-chave: análise de registro; escrita acadêmica; manual para análise de atributos-chave; estatística para alunos de humanas; seções de introdução

CONSTRUÇÃO DE CORPORA LINGUÍSTICOS COM PYTHON E IA: EXTRAÇÃO DE DADOS DE POSTS JORNALÍSTICOS, YOUTUBE E X (TWITTER) VIA WEB SCRAPING E APIs

Wagner da Cunha NUNES[74]

Este trabalho explora a metodologia para a obtenção de um corpus linguístico utilizando ferramentas de W*eb Sscraping* em *Python* (MITCHELL, 2018, p.47), com foco em dados provenientes do YouTube, X (anteriormente conhecido como Twitter) e posts jornalísticos de opinião e política. A construção de um *corpus* é essencial para diversas pesquisas em linguística, processamento de linguagem natural (PLN) e áreas afins. De acordo com Jurafsky e Martin (2021, p. 123), "o processamento de linguagem natural envolve a interação entre computadores e linguagem humana". As plataformas de redes sociais e sites de notícias são fontes ricas de dados textuais que refletem o uso cotidiano da linguagem, sendo, portanto, ideais para esse propósito. Um *corpus* linguístico é uma coleção estruturada de textos utilizados para conduzir análises e estudos linguísticos. A relevância de um *corpus* reside na sua capacidade de oferecer uma representação ampla e diversificada do uso da linguagem em contextos reais. Redes sociais como YouTube e X, bem como sites de notícias, fornecem uma abundância de dados textuais espontâneos e variados, fundamentais para análises aprofundadas em linguística e PLN. Foram escolhidas as plataformas YouTube, X e sites de notícias devido à diversidade e volume de comentários, postagens e artigos de opinião e política. Para a coleta de dados, utilizou-se a *API* do *YouTube Data* para acessar comentários de vídeos públicos e a *API* do *Twitter* para extrair *tweets* baseados em *hashtags* e palavras-chave. A extração de textos jornalísticos foi realizada por meio de *Web Scraping* em sites de notícias, focando em artigos de opinião e política. A implementação do *Web Scraping* foi realizada utilizando bibliotecas específicas do *Python*, como *BeautifulSoup*, *Selenium* e *Scrapy.* No caso da *API* do YouTube, foi empregada a biblioteca *google-api-python-client*, que facilita a interação com os serviços do Google. Para a *API* do *X*, utilizou-se a biblioteca *Tweepy*, amplamente utilizada para interagir com a *API* do Twitter usando *Python* (BROWN, 2017, p. 89). A integração dessas bibliotecas permitiu a construção de *scripts* automatizados para a extração de grandes volumes de dados textuais. Uma vez coletados, os dados passaram por um processo de limpeza, que envolveu a remoção de duplicatas, normalização de texto, eliminação de *emojis* e caracteres especiais. Foram utilizadas bibliotecas como *Pandas* e *Re* (expressões regulares) para a manipulação e limpeza dos dados. A criação de corpora por meio da linguagem *Python*, utilizando ferramentas de *Web Scraping* e *APIs*, juntamente com a integração de técnicas de inteligência artificial (*IA*) oferecidas pelo *ChatGPT*, desenvolvido pela *OpenAI* (BROWN et al., 2020, p. 30), tornou-se mais eficiente devido à facilidade de uso dessas ferramentas, mesmo sem a necessidade de conhecimento aprofundado em programação. O *corpus* linguístico, composto por comentários de vídeos do YouTube, *tweets* e textos de posts jornalísticos de opinião e política, proporciona uma rica base de dados para uma ampla gama de análises linguísticas e estudos de processamento de linguagem natural (PLN).

[74] Pesquisador Independente, Uberlândia – MG wagner.nunes@ufu.br

Palavras-chave: *IA*; Linguística de *Corpus*; *Web Scraping; Python* e *APIs.*

UM ETIQUETADOR PARA SINTAGMAS VERBAIS DA LÍNGUA ASURINÍ DO TOCANTIS

Luan Daniel dos Santos Sousa[75]
Thiago Blanch Pires[76]

É perceptível a necessidade de mais estudos e novas ferramentas que auxiliem nas pesquisas de línguas minorizadas, como grande parte das línguas indígenas brasileiras. Uma dessas línguas é o Asuriní do Tocantins, também conhecido como Asuriní do Trocará, do povo homônimo. Os Asurnís do Tocantins estão localizados no município de Tucuruí, no Pará, e são cerca de 500 habitantes da mesma etnia na região. Levando isso em consideração, como forma de contribuir para a revitalização linguística, este estudo visa criar um etiquetador morfossintático automático para sintagmas verbais na língua Asuriní do Tocantins a partir do corpus extraído do "Livro de Relatos Asuriní 2" utilizando-se de conhecimento do Processamento de Linguagem Natural (PLN) e diversas outras pesquisas que analisam a estrutura gramatical da língua. Manipulado computacionalmente por humanos, a linguagem de programação Python com o auxílio de três de suas bibliotecas, NLTK, spaCy e pandas, foi a ferramenta escolhida para criação do etiquetador morfossintático. Durante a criação do algoritmo para realizar a etiquetagem, houve uma tentativa de realizar o trabalho sem o uso da NLTK, usando apenas a spaCy para processamento de linguagem natural e a pandas para a análise de dados. Porém, o processo de criar as etiquetas customizadas usando a biblioteca spaCy se tornou inviável levando em consideração o tempo restante para realizar a pesquisa. Grande parte do trabalho feito com a spaCy foi aproveitado e a etiquetagem se resumiu usando NLTK e as ferramentas do pacote de Expressões Regulares do Python. Os resultados obtidos foram satisfatórios e possíveis de serem replicados e complementados por futuros pesquisadores.

Palavras-chave: Processamento de Linguagem Natural; Asuriní do Tocantins; etiquetador morfossintático; Python; sintagmas verbais.

---

75 Graduando de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação pela Universidade de Brasília. Artigo apresentado como Relatório Final do PIBIC Biênio 2022/2023.

76 Professor adjunto de Línguas Estrangeiras Aplicadas, Doutor em Gestão da Informação e Idealizador do GeLinC.

ARTIGOS CURTOS EBRALC-2024

REVISÃO E AMPLIAÇÃO DE ÁRVORES DE DOMÍNIO A PARTIR DA ANÁLISE DE CORPUS

Amanda Letícia Valadares dos SANTOS[77]
Flávia de Oliveira MAIA-PIRES[78]

RESUMO: Árvores de domínio são ferramentas de contextualização, servindo como estrutura norteadora de quais unidades linguísticas de fato constituem termos. Nesse sentido, não basta criar uma versão inicial da árvore, também se faz necessário revisá-la e ampliá-la conforme a pesquisa avança, de modo a incluir novos termos identificados. Para tanto, convém que o terminológo utilize a Linguística de Corpus (LC) para analisar relações de frequência e coocorrência nos textos da área.

Palavras-chave: Terminologia; Árvore de domínio; Linguística de Corpus; *Sketch Engine*; Lei Geral de Proteção de Dados Pessoais.

INTRODUÇÃO

A Terminologia é a área da Linguística responsável pelo estudo dos **termos**, isto é, das unidades linguísticas utilizadas em discursos especializados, cujo significado não é conhecido por leigos. Devido a esse recorte investigativo, entende-se que a Terminologia é interdisciplinar e envolve processos de familiarização com áreas distintas das Ciências da Linguagem.

Existem muitas formas de se aproximar e de compreender conceitualmente a estrutura de um novo tema de pesquisa. Nos estudos terminológicos, cabe destacar a elaboração de árvores de domínio, definidas como “a representação, em uma forma piramidal, dos conceitos-chave de um domínio e das relações que eles mantêm entre si” (ZAFIO, 1985, p. 161, tradução nossa).

Sendo assim, a Linguística de Corpus (LC) pode ser adotada como metodologia capaz de tornar essa elaboração mais precisa e eficiente. A partir do processamento dos textos, a LC produz um resultado numérico de quais termos são mais frequentes em textos da área de especialidade, bem como com quais substantivos, verbos e adjetivos esses vocábulos estabelecem relações significativas.

Portanto, este trabalho tem o objetivo geral de investigar como a LC pode auxiliar os terminólogos na elaboração das árvores de domínio. Notadamente, como objetivo específico, busca-se promover uma análise prática de aplicação da LC nos processos de revisão e ampliação das árvores de domínio inicialmente elaboradas pelo pesquisador — de modo a checar a pertinência dos termos

[77] Mestranda do Programa de Pós-Graduação em Linguística, da Universidade de Brasília (UnB), Brasília/DF. E-mail de contato: <linguista.amandavaladares@gmail.com>.

[78] Docente do Departamento de Linguística, Português e Línguas Clássicas (LIP), da Universidade de Brasília (UnB), Brasília/DF, e pesquisadora do grupo de pesquisa da UnB/CNPq: LexiC: Ciência, projetos e pesquisa sobre léxico: <http://lexic.com.br/>.

escolhidos para integrá-la, além de quais podem ser adicionados ao diagrama inicial.

Cabe destacar, ainda, que não se intencionou uma busca exaustiva de termos a partir da análise do *corpus*. Em vez disso, este estudo buscou mapear quais técnicas e recursos da ferramenta de LC *Sketch Engine* podem ser utilizados para os fins de revisão e ampliação das árvores de domínio — com destaque para a *Wordlist*, as *Keywords*, o *Word Sketch* e a *Concordance*.

FUNDAMENTAÇÃO TEÓRICA

Existem muitos modelos de árvores de domínio possíveis, todas com base epistemológica — algumas versões são mais ontológicas, outras taxionômicas e outras próximas de mapas mentais. Devido essa estrutura, também podem ser entendidas como uma forma de arquitetar informações. Nesse sentido, contribuem para o treinamento de Inteligências Artificiais (IA), ao constituir bases simplificadas e computacionalmente interpretáveis de conhecimento (KNIGHT, 2017).

Em geral, seguindo recomendações da própria ISO 704 (2000), as árvores de domínio, vistas como sistemas conceituais, fazem parte de uma etapa prévia de aproximação do terminológo com relação ao seu objeto de pesquisa (BARROS, 2004). Entretanto, cabe destacar que a árvore de domínio não é uma ferramenta estática. No decorrer dos estudos, o terminológo encontrará novos termos e novas relações epistemológicas, de modo que é prudente revisar e atualizar constantemente a primeira versão do diagrama.

> ...a árvore de domínio permite enquadrar cada termo em algum de seus ramos ou subáreas e desse modo garante tanto a existência do termo quanto seu pertencimento ao domínio (...) cada um dos termos deve se situar em algum lugar da estrutura básica que a árvore proporciona. Se isso não for possível, o candidato a termo deverá ser considerado um caso duvidoso ou inclusive ser excluído (BARITÉ, 2016, p. 97, tradução nossa).

Diante disso, ferramentas tecnológicas, utilizadas nos estudos terminólogicos que incluem a abordagem da LC, contribuem significativamente na identificação e na alocação de termos relevantes para a árvore de domínio da área estudada. Nesse sentido, recursos que identificam frequência e colocações auxiliam na construção de uma versão mais completa e verossímil do diagrama inicial, bem como garantem maior fidedignidade com os termos usados na prática. Com isso, é possível mitigar o que Aubert (2001) chama de “risco de ruído” e “risco de silêncio”. Isto é, respectivamente, o risco de incluir termos não relacionados ao campo estudado e o risco de não se incluírem termos importantes para a análise.

Do ponto de vista da estrutura, Barité (2016) explica que árvores de domínio são formadas por um anel nuclear e outro de termos afins. Por um lado, o anel nuclear é composto por termos que indicam conceitos diretamente e unicamente associados à área de especialidade em questão. No caso desse estudo, podem-se citar “titular”, “controlador”, “operador” e “dado pessoal” como exemplos. Por outro lado, os termos do anel afim seriam aqueles que se relacionam de forma mais circunstancial com aquela área do saber — no contexto da LGPD, é possível citar os termos “segurança da informação”, “termo de uso”, “terceiros”, “fornecedores” e “prestadores de serviço”.

METODOLOGIA

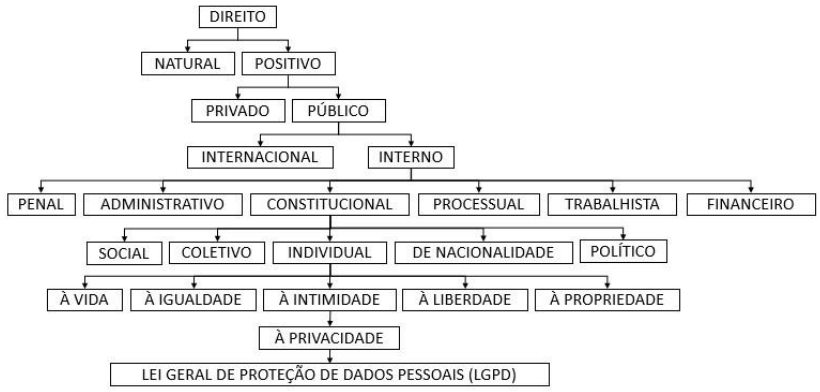
Para comprovar a necessidade de revisão e atualização de árvores conceituais em uma pesquisa terminológica, foram elaboradas árvores associadas a uma pesquisa em *corpus* com Políticas de Privacidade, leis, guias e normativos associados à Lei Geral de Proteção de Dados Pessoais – LGPD (BRASIL, 2018). Trata-se de uma lei que rege sobre a privacidade dos brasileiros, garantindo-lhes o direito de decidir o que fazer com suas informações, com quem as compartilha, com qual finalidade e por quanto tempo.

Para integrar o *corpus* dessa legislação, foram selecionados nove textos: (1) LGPD, com redação final de 2019; (2) Guia de Elaboração de Termo de Uso e Política de Privacidade para Serviços Públicos, na Versão 1.3; (3) Norma ABNT/NBR ISO 27701; (4) Política de Privacidade do BRB; (5) Política de Privacidade da Caixa; (6) Política de Privacidade do Banco do Brasil; (7) Política de Privacidade do Itaú; (8) Política de Privacidade do Bradesco; (9) Política de Privacidade do Santander. Esse *corpus* foi submetido ao *Sketch Engine* e, então, os recursos *Wordlist*, *Keywords*, *Word Sketch* e *Concordance* foram consultados para identificar possíveis candidatos a termos que se encaixariam na árvore de domínio, visando torná-la o mais completa possível.

DISCUSSÃO DOS DADOS

Em uma pesquisa da área do Direito, a primeira árvore de domínio elaborada buscou localizar como a LGPD se encaixa nessa área de especialidade, resultando em uma representação taxionômica da árvore disposta na Figura 1.

Figura 1 – Primeira árvore de domínio da LGPD

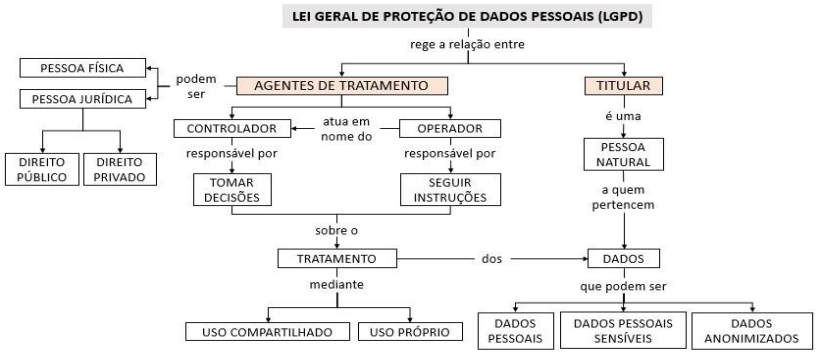


Fonte: Elaboração própria.

Uma segunda versão dessa árvore de domínio incluiu as relações específicas da própria LGPD, situando os termos da legislação em uma rede semântica. Dessa vez, a representação aproximou-se mais de estruturas ontológicas de organização das informações, conforme disposto, abaixo, na

Figura 2:

Figura 2 – Segunda árvore de domínio da LGPD



Fonte: Elaboração própria.

Após essa aproximação inicial, o *corpus* da LGPD foi submetido à ferramenta *Sketch Engine*. Em seguida, efetuou-se uma busca pela categoria "noun" no recurso *Wordlist*. Essa categoria gramatical foi a escolhida para investigação, pois a maioria dos termos são sintagmas nominais (KRIEGER, 2006). Em adição a isso, foi consultada a lista de *Keywords* da ferramenta. Trata-se de um recurso que identifica possíveis **candidatos a termo**, a partir de um cálculo comparativo de frequência — entre o *corpus* de estudo e um *corpus* de língua geral.

A Figura 3 ilustra os 15 primeiros substantivos e *keywords* (simples e complexos) identificados pelo *Sketch Engine*, de modo a demonstrar termos importantes ausentes nas árvores de domínio anteriores.

Figura 3 – Substantivos mais frequentes e keywords do *corpus* da LGPD

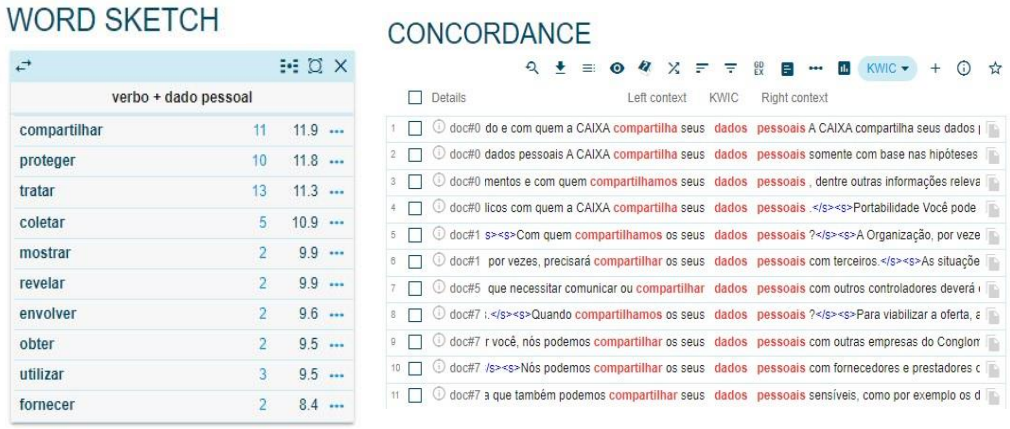
Fonte: Sketch Engine, 2024.

A partir dessa listagem, é possível notar que "dp", "informação", "organização", "serviço", "lei", "segurança", "direito", "controle", "anonimização", "subcontratado", "ANPD", "termo de uso" e "segurança da informação" são exemplos de termos que convém incluir na árvore de domínio.

Em seguida, executou-se uma busca pelos contextos de ocorrência de um dos termos relevantes da área: "dado pessoal", por meio do recurso *Word Sketch*. Essa busca revelou que "dado pessoal" estabelece relação com diversos verbos, sendo que

"compartilhar" é o mais frequente e se associa com o termo "uso compartilhado", presente na árvore da Figura 2.

Figura 4 – Verbos mais frequentes de “dado pessoal” e colocações de “compartilhar dado pessoal”



Fonte: Sketch Engine, 2024.

Uma verificação detalhada dessas ocorrências de “dado pessoal” associado a “compartilhar”, por meio do recurso *Concordance*, revelou a presença de entes de compartilhamento, como “terceiros”, “fornecedores” e “prestadores de serviço” — os quais também convém incluir na árvore de domínio elaborada.

CONCLUSÃO

Diante do exposto, comprova-se que a Linguística de *Corpus* pode ser uma abordagem metodológica muito importante na construção, revisão e ampliação de árvores de domínio durante a pesquisa terminológica. No *Sketch Engine*, ressaltam-se os recursos *Wordlist*, *Keywords*, *Word Sketch* e *Concordance*, como auxiliadores da identificação de possíveis candidatos a termos relevantes para o diagrama de representação conceitual da área.

Tais resultados contribuem para a elaboração de árvores de domínio mais completas e verossímeis com o uso real dos termos na comunicação especializada. Assim, o terminológo dispõe de mais ferramentas, além de sua leitura dos textos, para construir a esquematização conceitual de como os conceitos-chave se relacionam.

Além dos benefícios terminológicos de norteamento para a seleção e compreensão dos candidatos a termos, tal abordagem para revisão e atualização das árvores de domínio pode oferecer outras vantagens. No contexto das IAs, por exemplo, tais configurações cada vez mais densas de arquitetura da informação auxiliam no treinamento direcionado e completo de tecnologias relacionadas à linguagem em áreas de especialidade.

Agradecimentos: Agradecemos aos docentes e discentes da Universidade de Brasília (UnB) e do Grupo de pesquisa da UnB/CNPq: LexiC: Ciência, projetos e pesquisa sobre léxico, com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERÊNCIAS BIBLIOGRÁFICAS

AUBERT, Francis H. **Introdução à Metodologia da Pesquisa Terminológica Bilíngüe**. São Paulo, Humanitas Publicações-FFLCH/USP, 2001.

BARITÉ, Mario. Los árboles de dominio. *In*: CATALÁ, Sara A.; BARITÉ, Mario. (Org.). **Teoría y praxis en terminología**. 1. ed. Montevidéu: Ediciones Universitarias, Unidad de Comunicación de la Universidad de la República, 2016, v. 1, p. 91-102.

BARROS, Lídia A. **Curso Básico de Terminologia**. São Paulo: EdUSP, 2004.

BRASIL. Lei n° 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). **Diário Oficial da União**: seção 1, Brasília, DF, ano 155, n. 157, p. 59-64, 15 ago. 2018.

INTERNATIONAL STANDARD ORGANIZATION (ISO). **ISO 704**: Terminology work: principles and methods. Genebra: ISO, 2000.

KNIGHT, Michelle. Taxonomy vs Ontology: Machine Learning Breakthroughs. **Dataversity**, Los Angeles, 17 de out. de 2017. Disponível em: <https://www.dataversity.net/taxonomy-vs-ontology-machine-learningbreakthroughs/>. Acesso em: 07 jun. 2024.

KRIEGER, Maria G. Do ensino da terminologia para tradutores: diretrizes básicas. **Cadernos de tradução**, v. 1, n. 17, p. 189-206, 2006. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=4925478. Acesso em: 07 jun. 2024.

ZAFIO, Massiva N. L’arbre de domaine en terminologie. **Meta**: Journal des traducteurs, Montreal, vol. 20, n. 2, p. 161-168, jun. 1985. Disponível em: <https://www.erudit.org/fr/revues/meta/1985-v30-n2-meta308/004635ar.pdf>. Acesso em: 07 jun. 2024.

DISFLUÊNCIAS NA FALA ESPONTÂNEA DE PACIENTES COM ESQUIZOFRENIA: UMA ANÁLISE BASEADA NO CORPUS C-ORAL-ESQ

Átila Augusto Soares VITAL[79]
Bruno Neves Rati de Melo ROCHA[80]

ABSTRACT: This study compares disfluencies in the speech of patients with and without schizophrenia.The analysis uses the C-ORAL-ESQ corpus (ROCHA et al., 2020) for schizophrenic patients and CORAL-BRASIL I (RASO & MELLO, 2012). Disfluencies were examined in prosodic and pragmatic units. Results show higher occurrences in complex informational structures for both groups, with fewer reformulations by patients in sequences with more than one illocution.

Palavras-chave: disfluências; esquizofrenia; ilocuções; Prosódia; estrutura informacional.

INTRODUÇÃO

Este trabalho tem como objetivo apresentar os primeiros resultados de uma pesquisa em curso sobre disfluências em dois corpora de fala espontânea – o CORAL-ESQ (ROCHA et al., 2020), corpus que documenta a fala de pessoas com esquizofrenia durante consultas psiquiátricas, e o C-ORAL-BRASIL I (RASO & MELLO, 2012), corpus de referência do português brasileiro falado informal.

A esquizofrenia é uma doença mental com prevalência de cerca de 0,2% a 1% na população em geral, a depender de critérios diagnósticos e de grupo analisado. Os sintomas são subdivididos em positivos (como alucinações, delírios e baixo controle motor) e negativos (como anedonia, bloqueio afetivo e pobreza da fala). Muitas vezes, os sintomas linguísticos da esquizofrenia são descritos a partir da chamada desordem formal do pensamento (*formal thought disorder*), que inclui pobreza de conteúdo, falhas na expressão das informações, perda de objetivos, distração por sílabas e palavras e discurso incoerente. Nos últimos anos, há trabalhos que correlacionam os sintomas da patologia com a diminuição da variedade de unidades prosódicas produzidas pelos pacientes (disprosódia) (COVINGTON et al., 2004).

O corpus C-ORAL-ESQ tem sido compilado pelo Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL/UFMG) em cooperação com a equipe de psiquiatria do Instituto Raul Soares (IRS, Fundação Hospitalar do Estado de Minas Gerais - FHEMIG), em Belo Horizonte. Em momento oportuno, os dados serão disponibilizados eletronicamente para consultas e contarão com alinhamento texto-som (ROCHA et al, 2020).

LANGUAGE INTO ACT THEORY (L-ACT)

Assim como o C-ORAL-BRASIL I, o C-ORAL-ESQ tem sido segmentado e etiquetado em unidades informacionais segundo a *Language into Act Theory*

[79] Mestrando em Linguística Teórica e Descritiva pela Faculdade de Letras da Universidade Federal de Minas Gerais (FALE/UFMG), Belo Horizonte/MG. Contato: 4tilavital@gmail.com

[80] Professor da Faculdade de Letras da Universidade Federal de Minas Gerais (FALE/UFMG), Belo Horizonte/MG.

(L-AcT) (CRESTI, 2000; MONEGLIA & RASO, 2014; CAVALCANTE, 2016), teoria *corpusdriven* que considera que o enunciado e a *stanza* – sequências terminadas que veiculam duas ou mais ilocuções – são as unidades básicas da fala. Tanto o enunciado quanto a *stanza* veiculam atos de fala (AUSTIN, 1962) e representam unidades linguísticas com autonomia pragmático-prosódica. Nessa perspectiva, a fala espontânea implica uma constante execução de ações através da troca de ilocuções que compõem o contínuo do sinal sonoro. Segundo a L-AcT, a interpretação das ilocuções realizadas em enunciados e *stanzas* é guiada sobretudo por características do seu perfil prosódico.

As sequências terminadas podem ser de dois tipos: (i) simples, constituídas por uma única unidade tonal ou (ii) complexas, com duas ou mais unidades tonais, separadas por quebras prosódicas não terminais, que fazem com que a sequência linguística não seja autônoma pragmática e prosodicamente. Quando complexos, os enunciados são constituídos por estruturas internas que veiculam diferentes unidades informacionais, que também são distinguíveis através de formas prosódicas, função e posição em relação a outras unidades. Dentre elas, as unidades que constituem o texto da sequência e veiculam ilocuções são Comentário (COM), Comentário Ligado (COB) e Comentário Múltiplo (CMM). Há aquelas que constituem o texto, fornecendo informações sobre como interpretá-lo, sem veicular ilocuções: Apêndice de Comentário (APC), Tópico (TOP), Apêndice de Tópico (APT), Parentético (PAR) e Introdutor locutivo (INT). Por fim, há também as unidades dialógicas, que não participam da constituição do texto, mas que possuem a função de regular a interação: Alocutivo (ALL), Conector Discursivo (DCT), Incipitário (INP) e Expressivo (EXP).

As *stanzas,* que representam sequências prosodicamente terminadas de diferentes níveis de complexidade informacional, contêm mais de uma ilocução (CRESTI, 2009). Entre uma unidade ilocucionária e outra, podem haver unidades textuais e dialógicas que enriquecem os padrões informacionais.

Os fenômenos de disfluência – foco deste trabalho – são distinguidos, a princípio, em três tipos: enunciados interrompidos (marcados com “+”), quando há quebra prosódica e reformulação completa do planejamento da sequência; retractings ([/n], em que “n” é o número de palavras reformuladas), definidos como uma espécie de borracha prosódica e utilizados para reformulação de trechos do enunciado, sem que haja a interrupção completa (CAVALCANTE, 2020); e escansões (SCA), quando há a quebra do isomorfismo, e uma unidade informacional passa a ser realizada em duas ou mais unidades tonais. No exemplo 1, há um enunciado interrompido, já que, após a quebra não terminal anotada como “+”, há a reformulação completa do texto do enunciado. A barra simples (“/”) corresponde à quebra prosódica não terminal, enquanto que a barra dupla (“//”) corresponde à quebra terminal.

Exemplo 1 (bpubmn01)

*SHE: **então eu já levo meu som / eu já levo +** graças a Deus / né //

Nos exemplos 2 e 3, a seguir, há dois fenômenos de retractings, anotados com “[/2]” e “[/4]”, respectivamente. Em (2), há o cancelamento das duas palavras que constituem a primeira unidade tonal, o que faz com que elas não integrem o texto do enunciado. Nesse caso, a unidade é anotada com uma etiqueta vazia

(EMP). Em (3), por outro lado, há o cancelamento de 4 das 8 palavras que constituem a primeira unidade tonal, já que algumas delas são utilizadas para a constituição do sentido no restante do enunciado. Nesse caso, a quebra prosódica produz o retracting e uma unidade de escansão (SCA).

Exemplo 2 (bfamcv02)

*RUT: **e a [/2]=EMP=** e o meninim /=COM= né //=AUX=

Exemplo 3 (bfamcv08)

*REN: então quanto que dá **os que nũ &so [/4]=SCA=** os que sobraram //=COM=

METODOLOGIA

Através dos corpora etiquetados informacionalmente, foram realizadas buscas automáticas pelos fenômenos de disfluência descritos na seção anterior. Para a comparação entre a fala patológica e a fala não patológica, foi utilizada a metodologia apresentada por Raso et al. (2023), a partir dos estudos com os corpora C-ORALBRASIL I e C-ORAL-ESQ.

Os enunciados passaram por um pré-processamento para retirada de elementos que seriam desconsiderados para a análise. No caso do C-ORAL-ESQ, por exemplo, foram retirados os turnos de fala dos psiquiatras e dos acompanhantes, mantendo apenas a fala dos pacientes.

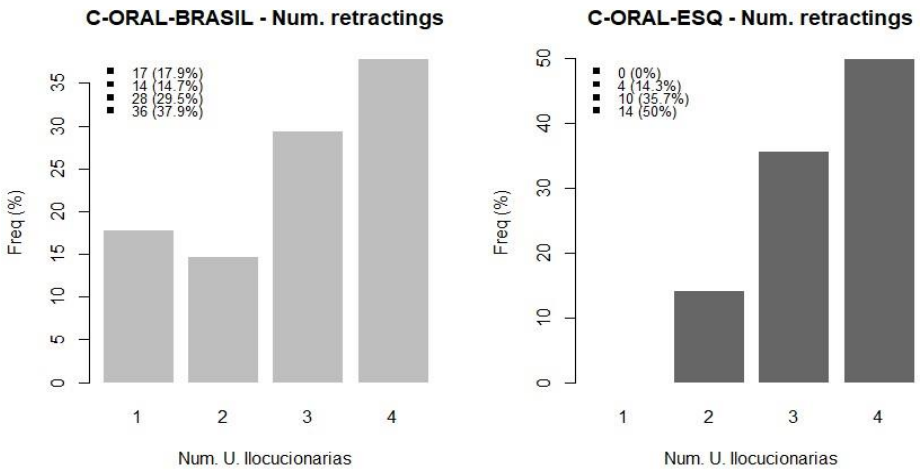
Os textos foram segmentados em sequências terminadas, que foram agrupadas em listas a partir do número de ilocuções. Tais listas comportam as stanzas de diferentes complexidades (de uma a quatro unidades ilocucionárias). Para cada um dos níveis de complexidade, foi realizada amostragem aleatória de quarenta e uma sequências terminadas, dentro das quais foram realizadas as medidas de quantidade e tipo de disfluências, posição na sequência e quantidade de palavras envolvidas.

A comparação entre os corpora se dá através da comparação da distribuição de disfluências entre os diferentes níveis de complexidade de *stanzas*, de modo que as amostras que contenham uma unidade ilocucionária no corpus de fala patológica sejam comparadas com aquelas que também contenham uma unidade ilocucionária no de fala não patológica. O mesmo procedimento é realizado para cada um dos quatro níveis de complexidade.

RESULTADOS E DISCUSSÕES

Após a coleta das amostras, foram realizadas as medidas das disfluências. Nas 41 amostras do C-ORAL-ESQ, foram encontradas 28 ocorrências de retractings, sendo que 82,9% cancelam de uma a duas palavras. No caso do C-ORAL-BRASIL, para o mesmo número de enunciados, foram encontrados 95 retractings, sendo 94,7% de uma ou duas palavras. Conforme a figura 1, ambos os corpora concentram maior quantidade de retractings em enunciados com maiores níveis de complexidade (com três e quatro unidades ilocucionárias). Aparentemente, a fala patológica conta com menor número de disfluências em enunciados menos complexos.

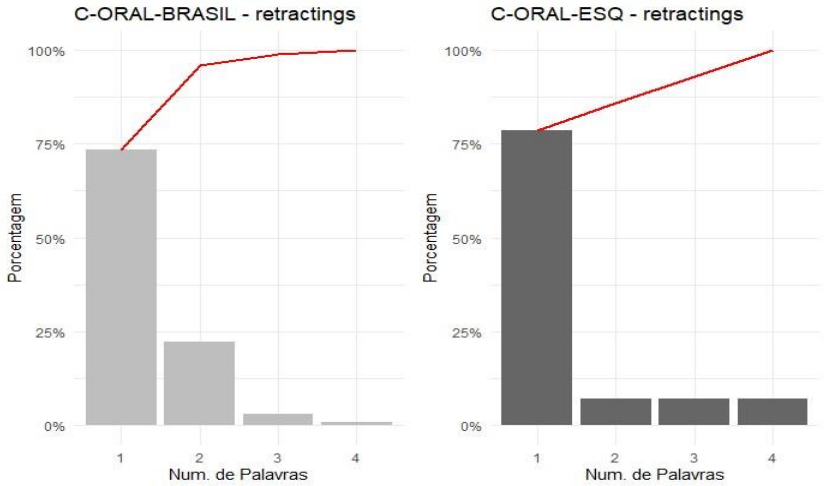
Figura 1: Gráficos sobre a quantidade de retractings presentes em cada nível de complexidade das amostras dos corpora.



Fonte: elaboração própria.

No caso do número de palavras canceladas, há diferenças entre a fala patológica e a não patológica. Na figura 2, é possível perceber que, no discurso dos pacientes, são menos frequentes retractings que cancelem mais de uma palavra.

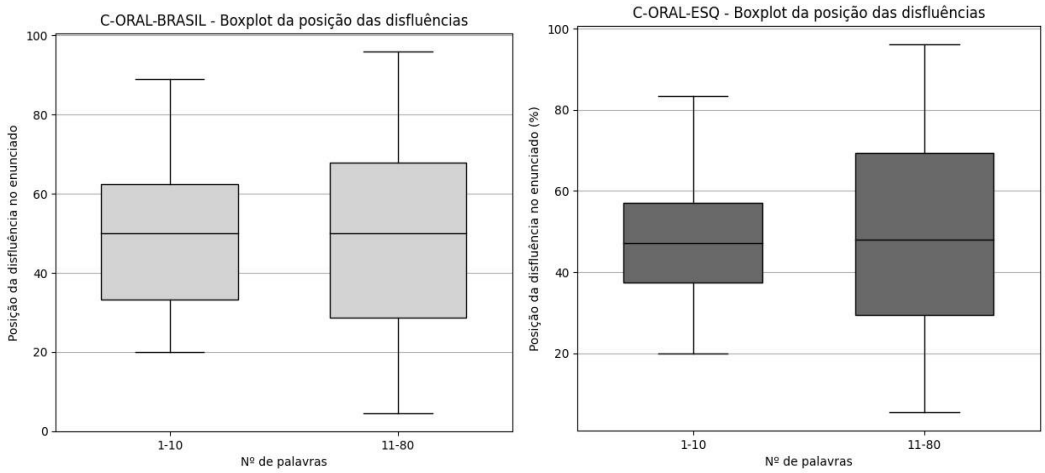
Figura 2: Gráficos sobre o número de palavras canceladas nas amostras dos corpora.



Fonte: elaboração própria.

Em relação à medida das posições das disfluências, para os dois corpora, quanto maior é o número de palavras no enunciado, mais variadas tendem a ser as posições para a ocorrência de reformulações, interrupções e escansões, conforme a figura 3.

Figura 3: Boxplots da posição das disfluências nos enunciados dos corpora. A posição é medida em termos da porcentagem do enunciado (ex.: uma escansão ocorreu em 30% de um enunciado de 10 palavras; a escansão ocorreu entre a segunda e a terceira palavra).



Fonte: elaboração própria.

Quanto ao tipo de retractings, no C-ORAL-BRASIL, são mais frequentes aqueles em que o falante cancela todas as palavras da unidade tonal. No C-ORALESQ, por outro lado, são mais frequentes os cancelamentos parciais.

Os resultados apresentados são preliminares e serão revisados à medida em que mais consultas psiquiátricas forem gravadas, transcritas e revisadas para constituição do C-ORAL-ESQ. Os próximos passos da pesquisa são a análise das unidades informacionais em que mais ocorrem as disfluências e dos enunciados em que retractings e escansões são realizados em sequências. Com a finalização do corpus e a disponibilização de mais ferramentas para o estudo da fala de pacientes com esquizofrenia, novos caminhos serão abertos para a descrição da linguagem.

REFERÊNCIAS

AUSTIN, J. L. *How to do things with words*. Oxford University Press, Oxford 1962.

CAVALCANTE, F. A., *The informational unit of topic: a crosslinguistic, statistical study based on spontaneous speech corpora*. PhD dissertation in Theoretical and Descriptive Linguistics, Universidade Federal de Minas Gerais, 2020.

CAVALCANTE, F. A., *The topic unit in spontaneous american English: a corpus-based study*. Master 's dissertation in Linguistics, Universidade Federal de Minas Gerais, 2016.

COVINGTON, M.A. et al. Schizophrenia and the structure of language: the linguist's view. *Schizophrenia research*, v. 77, n. 1, p. 85-98, 2005.

CRESTI, E., *Corpus di Italiano parlato*, Accademia della Crusca, Firenze 2000.

CRESTI, E. La Stanza: un'unità di costruzione testuale del parlato. In: *Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana*, SILFI 2008. Basilea, 30.06-03.07 2008. 2009, p. 1-25.

MONEGLIA, M., RASO, T., Notes on Language into Act Theory (L– AcT), in T. Raso, H. Mello (eds), *Spoken Corpora and Linguistic Studies*, John Benjamins Publishing Company, Amsterdam – Philadelphia 2014, pp. 468–495.

RASO, T., DE MELO ROCHA, B.N.R., SALGADO, J.V. et al. The C-ORAL-ESQ project: a corpus for the study of spontaneous speech of individuals with schizophrenia. *Language Resources and Evaluation*, p. 1-21, 2023.

RASO, T. MELLO, H. (eds), C–ORAL–BRASIL I: *Corpus de referência do Português Brasileiro falado informal*, Editora UFMG, Belo Horizonte 2012.

ROCHA, B., FERRARI, L. A., MANTOVANI, L. M., RASO, T., SALGADO, J. V., A corpus of Brazilian Portuguese speech by schizophrenic patients: preliminary observations. *Lingua e patologia: i sistemi instabili*, 2020, pp. 307-333.

RODA VIVA: UM CORPUS ORAL E A UNIVERSAL DEPENDENCIES

Cláudia Dias de BARROS[81]
Oto Araújo VALE[82]

Resumo: Neste artigo é apresentado o trabalho sobre a construção de um subcorpus composto por quatro entrevistas extraídas do Corpus Roda Viva (MIRANDA JR. et al., 2024), o qual é composto por 713 entrevistas do programa Roda Viva da TV Cultura. As quatro entrevistas trabalhadas foram transcritas automaticamente por meio da ferramenta Whisper (RADFORD et al., 2023), anotadas e revisadas com as etiquetas de Universal Dependencies.

Palavras-chave: Universal Dependencies; sintaxe; Linguística de Corpus; PLN; corpus oral.

INTRODUÇÃO

Os trabalhos na área de Processamento de Línguas Naturais (PLN) se utilizam muito de corpora a fim de comprovarem hipóteses linguísticas sobre algum fenômeno ou pesquisar novos fenômenos, por exemplo.

Neste artigo é apresentada a pesquisa realizada sobre a construção de um subcorpus do Corpus Roda Viva (MIRANDA JR. et al., 2024), que é formado por 713 entrevistas de vários anos do programa Roda Viva da TV Cultura, transcritas por jornalistas de forma textualizada, nas quais há complementações das falas, por meio de inserções textuais, informações enciclopédicas, entre outros, o que faz com que percam o status de língua oral, passando a língua escrita.

Dessa forma, a pesquisa aqui retratada tomou a decisão de construir o subcorpus com quatro entrevistas, totalizando 4024 sentenças, e, a fim de manter o status de língua oral, decidiu-se realizar a transcrição automática das entrevistas trabalhadas por meio de um ASR (Sistema de Reconhecimento Automático de Fala) chamado Whisper (RADFORD et al., 2023). Os textos transcritos apresentaram alguns problemas como transcrição equivocada de algumas palavras e erro de segmentação das sentenças, que precisaram ser corrigidos manualmente posteriormente.

A escolha das quatro entrevistas se deu baseada na possível diversidade sintática apresentada pelos quatro entrevistados, sendo eles: uma governadora, um desenhista de história em quadrinhos, um jogador de futebol e um rapper.

A partir dos textos transcritos revisados foi realizada a anotação automática com o formalismo da Universal Dependencies (DE MARNEFFE et al., 2021). Essa anotação foi realizada pelo parser PortParser (LOPES et al., 2024).

---

[81] Docente do Curso de Licenciatura em Letras, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Câmpus Sertãozinho, claudias84@gmail.com.

[82] Docente do Curso de Licenciatura em Letras e Bacharelado em Linguística, Universidade Federal de São Carlos – UFSCar.

Após a anotação automática, foi feita uma revisão manual por meio da ferramenta Arborator-Grew ElizIA (GUIBON et al., 2020) e foram identificados alguns fenômenos característicos da língua falada, como a presença de vocativos e marcas discursivas.

O objetivo dessa anotação é fornecer um corpus de língua oral ao projeto Porttinari (PARDO et al., 2021), um grande corpus multigênero do Português do Brasil, composto por textos escritos, como artigos de jornal, tweets do mercado financeiro brasileiro, revisões de consumidores de ecommerce e revisões de livros.

Inicialmente, serão adicionadas ao Porttinari as quatro entrevistas trabalhadas, porém, posteriormente, após o treinamento do parser com essas entrevistas, serão também anotadas automaticamente e adicionadas ao projeto as outras 709 entrevistas do Corpus Roda Viva (MIRANDA JR. et al., 2024).

FUNDAMENTAÇÃO TEÓRICA

A pesquisa apresentada neste artigo teve como arcabouço teórico a Universal Dependencies[83] (UD) (DE MARNEFFE et al., 2021), um projeto que tem como objetivo uma anotação gramatical consistente (etiquetas morfossintáticas, características morfológicas e dependência sintática), entre línguas humanas diferentes. A UD é um esforço colaborativo de cerca de 500 colaboradores que produziram quase 200 treebanks para aproximadamente 100 línguas.

Atualmente, a UD possui dezessete etiquetas morfossintáticas ou Partof-Speech (PoS) tags, como: ADJ: adjetivo, VERB: verbo e ADV: advérbio. Ela também possui 37 etiquetas de relações de dependência – deprel (de dependency relation). Uma deprel é uma relação que liga dois a dois os elementos (tokens) de uma sentença. Um deles é chamado de *head* (núcleo), que é sempre uma palavra de conteúdo (verbo, substantivo, adjetivo, pronome, numeral e advérbio) – exceções são símbolos que podem ser expressos por palavras, como R$ (reais), % (por cento) e § (parágrafo); e o outro é chamado de dependente.

Toda sentença tem uma raiz (normalmente o predicado da oração principal), marcada como dependente da deprel *root*. A atribuição de relações de dependência deve observar o princípio da projetividade, ou seja, os arcos das relações não devem se cruzar.

Alguns exemplos de etiquetas *deprel* são: amod: modificador adjetival; appos: modificador aposicional e aux: auxiliar.

Na subseção a seguir são apresentados os passos metodológicos seguidos na pesquisa.

METODOLOGIA

Nesta subseção são apresentados os passos metodológicos seguidos até o momento para a construção do subcorpus com as quatro entrevistas retiradas do Corpus Roda Viva (MIRANDA JR, et al., 2024):

[83] Disponível em: https://universaldependencies.org/

1. Escolha das quatro entrevistas que compõem o subcorpus;
2. Transcrição automática dos vídeos presentes no Youtube das quatro entrevistas, realizada pelo ASR Whisper (RADFORD et al., 2023), a qual gera um arquivo .txt;
3. Correção manual de problemas apresentados nos textos transcritos, como transcrição equivocada de palavras (principalmente estrangeirismos e dicção ruim dos entrevistados) e segmentação errada de sentenças;
4. Anotação automática com as etiquetas da Universal Dependencies (DE MARNEFFE et al., 2021), por meio do parser PortParser (LOPES et al., 2024);
5. Revisão manual da anotação automática, por meio da ferramenta Arborator-Grew ElizIA (GUIBON et al., 2020);
6. Verificação automática da anotação e revisão, por meio da ferramenta Verifica UD (LOPES et al., 2023);
7. Observação de fenômenos linguísticos característicos de um corpus de fala.

DISCUSSÃO DOS DADOS

Nesta subseção serão apresentados e discutidos os fenômenos linguísticos observados na pesquisa.

Por meio das análises das sentenças trabalhadas, puderam ser notados alguns fenômenos típicos da fala, como o uso de vocativos, como mostra a Figura 1 e de palavras discursivas, como ‘né’, apresentado na Figura 2:

Figura 1: Exemplo do uso de um vocativo. Fonte: elaborado pela autora

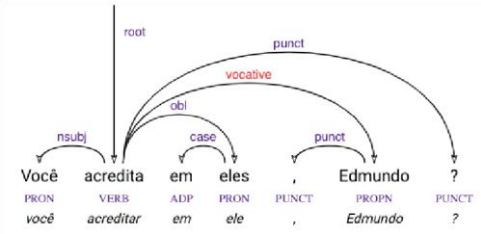
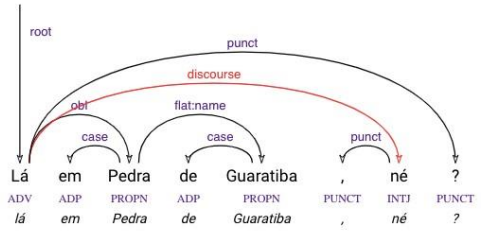
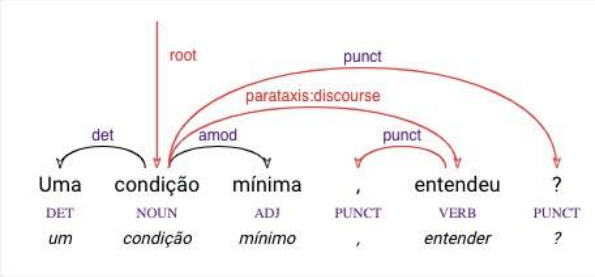


Figura 2: Exemplo do uso de uma palavra discursiva. Fonte elaborado pela autora



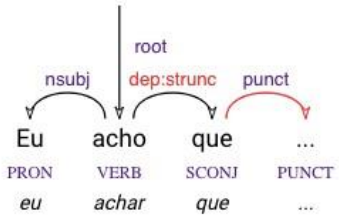
Com relação às expressões discursivas que contêm um verbo, decidiuse que seriam etiquetadas com *parataxis:discourse*, como é mostrado na Figura 3.

Figura 3: Exemplo de uso da etiqueta *parataxis:discourse.* Fonte: elaborado pela autora



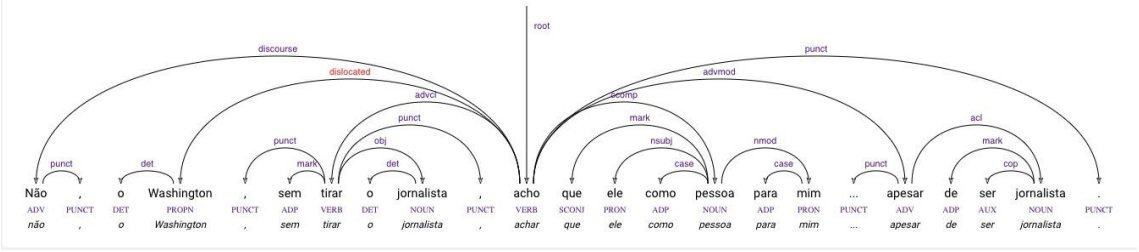
Outro fenômeno bastante recorrente observado no corpus foram as sentenças quebradas, ou seja, que apresentam um truncamento, nas quais o falante não termina sua linha de raciocínio. Elas são sempre marcadas pela presença das reticências. Para anotar esse fenômeno, decidiu-se criar a subrelação *strunc* (truncamento de sentença) e adotou-se a relação *dep* (dependência não especificada), formando-se a etiqueta *dep:strunc*. Um exemplo desse tipo de anotação pode ser observado na Figura 4.

Figura 4: Exemplo de uso da etiqueta *dep:strunc.* Fonte: elaborado pela autora



Foi observada, também, a presença frequente de um outro fenômeno, o qual é relacionado à posição de sujeito das sentenças: os sujeitos deslocados, anotados com a etiqueta dislocated. Nesses casos, o falante utiliza dois sujeitos para um mesmo verbo, sendo que um deles é topificado, no início da sentença, e o outro fica próximo ao verbo. Um exemplo desse fenômeno pode ser notado na Figura 5.

Figura 5: Exemplo de uso da etiqueta *dislocated.* Fonte: elaborado pela autora



Como se havia previsto, a entrevista com o rapper foi a que mais apresentou menor formalidade e se mostrou desafiadora para o parser anotar corretamente as relações sintáticas.

Na entrevista com a governadora do estado, observou-se uma grande quantidade de orações subordinadas e coordenadas, fruto de um discurso mais prolixo, característico de um político.

A entrevista do jogador de futebol apresentou a ocorrência de muitas etiquetas *dislocated.*

O objetivo final do trabalho é realizar o mesmo processo retratado aqui com as outras entrevistas do corpus Roda Viva, a fim de se aumentar a porção de corpus oral do projeto Porttinari (PARDO et al., 2021).

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.

REFERÊNCIAS BIBLIOGRÁFICAS

DE MARNEFFE, M. C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal dependencies. *Computational linguistics*, 47(2), pp. 255– 308, 2021.

GUIBON, G.; COURTIN, M.; GERDES, K.; GUILLAUME, B. When collaborative treebank curation meets graph grammars: arborator with a grew back-end. *Proceedings of the 12th Language Resources and Evaluation Conference,* Marseille, France, European Language Resources Association, pp. 5293- 5302, mai. 2020.

LOPES, L.; DURAN, M. S.; PARDO, T. A. S. Verifica UD - A verifier for Universal Dependencies annotation in Portuguese. In: *Proc. of the UDFest-BR 2023*, 2023. DOI: https://doi.org/10.5753/stil.2023.25485

MIRANDA Jr., Isaac; PEDRO, Gabriela Wick; BARROS, Cláudia Dias de; VALE, Oto Araújo. Roda Viva Boundaries: an overview of an audio-transcription corpus. *Proceedings of the 16th International Conference on Computational Processing of Portuguese,* vol, 2, pp. 165-169, mar. 2024.

PARDO, Thiago Alexandre Salgueiro; DURAN, Magali Sanches;, LOPES, Lucelene; DI FELIPPO, Ariani; ROMAN, Norton T.; NUNES, Maria das Graças Volpe. Porttinari - a large multi-genre treebank for brazilian portuguese. *Proceedings of the XIII Symposium in Information and Human Language (STIL)*, pp. 1-10. nov, 29 a dez, 3. 2021.

RADFORD, A; KIM, J.W.; XU, T.; BROCKMAN, G.; MCLEAVEY, C.; SUTSKEVER, I. Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning,* PMLR 202, pp. 28492-28518, 2023.

OS (DES)ENCONTROS DA LINGUÍSTICA DE CORPUS COM A TRADUÇÃO FEMINISTA

Luciana Carvalho FONSECA[84]

Resumo: Este estudo, publicado em FONSECA (2024), explora as interseções entre a Linguística de Corpus (LC) e a Tradução Feminista, destacando a lacuna existente em pesquisas que combinem LC, tradução *e* gênero. Trata-se de uma revisão bibliográfica de 23 coletâneas e 17 números especiais de periódicos sobre gênero/feminismo/mulheres e tradução, publicados até 2022. A análise revelou uma marcada ausência de pesquisas nos Estudos Feministas da Tradução que se valem de LC. **Palavraschave:** Estudos Feministas da Tradução; Linguística de Corpus; gênero; feminismo; tradução.

Introdução

A Linguística de Corpus (LC) tem sido consistentemente empregada para analisar linguagem e gênero. Há estudos sobre gênero e diferença, gênero e linguagem, gênero e discurso, gênero e representação, gênero e terminologia etc. No que diz respeito aos estudos da tradução (ET) e gênero, a amplitude de campos e objetos de estudo é ainda maior e abrange os estudos teóricos, descritivos e aplicados da tradução. No entanto, estudos que reúnam especificamente a LC, tradução *e* gênero não têm atraído muito interesse de pesquisa.

Onde os estudos da tradução e os estudos de gênero se encontram, situa-se a tradução feminista (TF), identificada como um subcampo específico dos ET na década de 1970 no Quebec. Mais de 50 anos depois, a TF – um campo que se ocupa de gênero e tradução, mulheres e tradução etc. – tem sido praticada e teorizada por pesquisadores e pesquisadoras da tradução em todo o mundo a partir de múltiplos pontos de vista. No entanto, metodologicamente, o campo ainda se baseia predominantemente em leituras atentas (*close reading*) e métodos manuais (*hand and eye*).

Assim, para identificar, investigar e discutir como as abordagens de LC têm sido empregadas na TF, começo com um breve relato da inter-relação entre "corpus e tradução", "corpus e gênero" e "tradução e gênero". A partir de uma revisão da literatura baseada em 23 coletâneas e 17 números especiais de periódicos sobre gênero/feminismo/mulheres e tradução publicados até 2022, apresento e discuto como – e se – os ETF (Estudos da Tradução Feminista) têm se valido de métodos e ferramentas da LC.

Fundamentação teórica

O uso de corpora nos ET permite que pesquisadoras e pesquisadores testem hipóteses de tradução, identifiquem padrões e lacunas de tradução, construam e alimentem memórias de tradução e empreguem tradução

[84] Professora Doutora do Departamento de Letras Modernas, Programas de Pós-Graduação em Letras Estrangeiras e Tradução (LETRA) e Estudos Linguísticos e Literários em Inglês (ELLI), da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH/USP). E-mail: lucianacarvalhof@usp.br

automática, para citar apenas alguns pontos de entrada. Além dessa aplicação abrangente nos ET, os corpora também desempenharam um papel importante nas pesquisas sobre gênero na LC, em combinação com disciplinas como a sociolinguística e a análise do discurso. Contudo, nos estudos sobre gênero, a utilização de métodos de LC está longe de ser generalizada. Isso foi ilustrado por Paul Baker, em sua introdução ao número especial sobre LC do periódico *Gender and Language* (BAKER, 2013, p.1). O autor reuniu cinco artigos publicados anteriormente pelo mesmo periódico que empregam métodos de LC a temas de gênero. Entretanto, a tradução não foi abordada em nenhum deles.

Se pesquisas que reúnem gênero e LC têm negligenciado a tradução, as pesquisas reunindo tradução e gênero também têm negligenciado a LC. Pouco antes do número organizado por Baker, Olga Castro editou um número especial do mesmo *Gender and Language* intitulado “Gender, Language and Translation at the Crossroads of Disciplines” (CASTRO, 2013). Dos seis artigos que compõem o número de Castro, nenhum faz referência explícita aos métodos de LC, embora um deles mencione um 'corpus' de várias centenas de páginas (SANTAEMILIA, 2013).

Em suma, os números especiais organizados por Baker e Castro indicam a existência de pesquisas em gênero e LC e em gênero e ET; no entanto, sugerem uma ausência de pesquisas combinem LC, tradução e gênero. Seria de se esperar que tal combinação fosse encontrada nos ETF. Foi com o intuito de aprofundar a discussão sobre as metodologias de LC nos ETF que esta pesquisa teve início.

Metodologia

Embora apenas dois números especiais de *Gender and Language* sejam meramente indicativos de uma possível lacuna na literatura sobre o emprego de LC nos estudos sobre tradução e gênero, a revisão da literatura realizada neste estudo não só confirma esta lacuna, mas também revela as poucas instâncias em que LC, tradução e gênero foram reunidos em estudos sobre tradução e gênero/feminismo/mulheres.

A revisão da literatura baseou-se em 23 coletâneas e 17 números especiais de periódicos sobre gênero/feminismo/mulheres e tradução publicados até 2022. As 40 publicações contêm estudos que abordam o emprego de LC, em alguma medida, para estudar tradução e gênero/feminismo. Classifiquei os estudos em três grupos. O Grupo 1 é composto por 10 artigos/capítulos que adotaram uma abordagem manifestamente baseada em ou orientada por corpus. O Grupo 2 contém artigos/capítulos cuja metodologia aparentava ser informada por corpus, mas que não explicitam seus métodos. E o Grupo 3 é composto por artigos que mencionaram brevemente pesquisas de corpus em suas referências, mas não se engajam profundamente com a LC.

Discussão

Em termos de língua e edição, é interessante notar que das 10 publicações – 9 capítulos e 1 artigo – do Grupo 1, todas foram todas escritas em inglês (CAIAZZO, 2011; CALABRESE, 2011; FEDERICI; LEONARDI, 2012; FIANO; GRIMALDI, 2021; LEONARDI, 2017; PETTINI, 2020; PIZARRO SEIJAS, 2018; SALDANHA, 2003; SANTAEMILIA, 2017a; WILLIAMS CAMUS, 2018) e

publicadas em obras cujas organizadoras e organizador são principalmente da Espanha (CAMUS; CASTRO;

WILLIAMS CAMUS, 2017; CASTRO; ERGUN, 2017; FEDERICI; MACI, 2021;

GODAYOL, 2012; SANTAEMILIA, 2003; WILLIAMS CAMUS et al., 2018; ZARAGOZA

NINET et al., 2018). A distribuição por idioma e nacionalidade das organizadoras convida a discussões sob a atual tendência transnacional nos ETF (CASTRO et al., 2024), os quais têm sido produzidos predominantemente em inglês. Tal constatação nos permite questionar o papel - ou a falta do papel - do multilinguismo nas publicações e/ou colaborações transnacionais na TF.

Embora os estudos feministas da tradução tenham tradicionalmente se preocupado com a tradução literária, três textos revisados no Grupo 1 abordam a tradução especializada. Como afirmam Federici e Maci “estudos feministas recentes demonstram que a tradução especializada está se tornando um campo produtivo de estudo na tradução feminista” (FEDERICI; MACI, 2021, p. 9), e as organizadoras mencionam duas coleções de tradução feminista que incluíram estudos sobre linguagem especializada (CAMUS; CASTRO; WILLIAMS CAMUS, 2017; SANTAEMILIA, 2017b), ambas levantadas durante a presente pesquisa de revisão bibliográfica.

Um aspecto constatado sobre os estudos sobre gênero e tradução baseados em corpus que merece ser destacado é ausência de explicitação do emprego de métodos baseados em corpus e uma hesitação em nomear a LC. Apenas dois dos estudos mencionam corpus no título (PIZARRO SEIJAS, 2018; SANTAEMILIA, 2017a); dois artigos não mencionam tradução *e* gênero em seus títulos (CALABRESE, 2011; PETTINI, 2020); e, o que continua a surpreender – embora seja esperado – apenas um dos textos reivindica abertamente o título político de feminista (CASTRO; ERGUN, 2017, p. 2). O capítulo que faz essa menção é o de Santaemília (2017), que aborda especificamente o tema da terminologia no campo da TF. A falta geral de informação em títulos e palavras-chave – embora nove dos dez textos fossem capítulos e não contivessem palavras-chave individuais – explica a dificuldade inicial que tive ao reunir publicações que tratassem de corpus, tradução e gênero/feminismo/mulheres consultando bancos de dados indexados.

Assim, a partir do levantamento e análise realizados, foi constatado que os métodos e/ou ferramentas de LC têm sido empregados de forma bastante incipiente, mas que os estudos sobre gênero e tradução envolvendo linguagens de especialidade são uma promissora porta de entrada para abordagens e métodos em LC.

. Por fim, a presente pesquisa, a qual se encontra publicada na íntegra em (FONSECA, 2024) também ofereceu subsídios para se discutir as possíveis dificuldades na integração da TF com a LC, as quais podem ser resumidas da seguinte forma: a natureza quantitativa da LC e a natureza qualitativa da TF; a suposta natureza objetiva da LC e a natureza subjetiva da TF; o dilema da abordagem ‘baseada em corpus; a dependência excessiva na forma por parte da LC e a instabilidade das categorias na teoria feminista; a prevalência da

tradução literária nos ETF e a prevalência da tradução especializada na LC; e a institucionalização da teoria feminista ao lado da literatura e dos ET ao lado da linguagem especializada, onde a LC também costuma ser encontrada.

Referências

BAKER, Paul. Introduction: Virtual Special Issue of Gender and Language on corpus approaches. **Gender and Language**, *[S. l.]*, v. V1, n. Virtual Special Issue, p. 1–5, 2013. DOI: 10.1558/8psxqda5wh3d.

CAIAZZO, Luisa. Female text(ure) and science: Ada Byron's Notes and translation.
Em: PALUSCI, Oriana (org.). **Traduttrici: female voices across languages**. Trento: Tangram, 2011. p. 59–72.

CALABRESE, Rita. Living on the edge of two languages: le costruzioni possessive in In the Second Person di Smaro Kamboureli. *Em*: PALUSCI, Oriana (org.).
Traduttrici: female voices across languages. Trento: Tangram, 2011. p. 175–188.

CAMUS, Carmen Camus; CASTRO, Cristina Gómez; WILLIAMS CAMUS, Julia T. **Translation, Ideology and Gender**. Newcastle upon Tyne: Cambridge Scholars Publishing, 2017.

CASTRO, Olga (ORG.). Gender, language and translation at the crossroads of disciplines. **Gender and Language**, *[S. l.]*, v. 7, n. 1, p. 5–12, 2013. DOI: 10.1558/genl.v7il.5.

CASTRO, Olga; ERGUN, Emek (ORG.). **Feminist Translation Studies: Local and Transnational Perspectives**. New York & London: Routledge, 2017.

CASTRO, Olga; ERGUN, Emek; SPURLIN, William J.; BRACKE, Maud Anne; FONSECA, Luciana Carvalho. Transnationalizing Feminist Translation Studies? Insights from the Warwick School of Feminist Translation. **Journal of Feminist Scholarship, Special issue: Translating Transnational Feminisms**, *[S. l.]*, v. 24, n. 24, 2024.

FEDERICI, Eleonora; LEONARDI, Vanessa. Using and Abusing Gender in Translation: The Case of Virginia Woolf's A Room of One's Own Translated into Italian. **Dossier. La traducció i els estudis de gènere. Quaderns**, *[S. l.]*, v. 19, p. 183–198, 2012.

FEDERICI, Eleonora; MACI, Stefania (ORG.). **Gender Issues: Translating and** Mediating Languages, Cultures and Societies. Bern: Peter Lang, 2021. v. 281
FIANO, Carmen; GRIMALDI, Agnese Daniela. Gender Advisor, a New Role to Ensure Gender Equality Within NATO. To Translate or Not to Translate? *Em*: FEDERICI, Eleonora; MACI, Stefania (org.). **Gender Issues: Translating and Mediating Languages, Cultures and Societies**. Linguistic Insights: Studies in Language and CommunicationBern: Peter Lang, 2021. v. 281p. 199–220.

FONSECA, Luciana Carvalho. Corpora, Translation and Gender: Feminist Translation and Corpus Linguistics at the Crossroads. *Em*: LI, Defeng; CORBETT, John (org.). **The Routledge Handbook of Corpus Translation Studies**. London and New York: Routledge, 2024. p. 544–563.

GODAYOL, Pilar (ORG.). Dossier. La traducció i els estudis de gènere. **Quaderns**, *[S. l.]*, v. 19, 2012. Disponível em: https://raco.cat/index.php/QuadernsTraduccio/article/view/105017.

LEONARDI, Vanessa. Gender, Language and Translation in the Health Sciences: gender biases in medical textbooks. *Em*: CAMUS, Carmen Camus; CASTRO, Cristina Gómez; WILLIAMS CAMUS, Julia T. Williams (org.). **Translation, Ideology and Gender**. Newcastle upon Tyne: Cambridge Scholars Publishing, 2017. p. 8–31.

PETTINI, Silvia. Gender in war video games: The linguacultural representation and localization of female roles between reality and fictionality. *Em*: FLOTOW, Luise Von; KAMAL, Hala (org.). **The Routledge Handbook of Translation, Feminism and Gender**. (Eds.)London /New York: Routledge, 2020. p. 444–456.

PIZARRO SEIJAS, Paloma. Using Corpus Tools to Analyse the Rendering of Joseph Conrad's Women in the Heart of Darkness into Four Spanish Translations. *Em*: WILLIAMS CAMUS, Julia T.; GÓMEZ CASTRO, Cristina; ASSIS ROSA, Alexandra; CAMUS CAMUS, Carmen (org.). **Translation and gender. Discourse strategies to shape gender**. Santander: Cantabria University Press, 2018. p. 135–152.

SALDANHA, Gabriela. Studying Gender-Related Linguistic Features in Translated Language. *Em*: SANTAEMILIA, José (org.). **Género, lenguaje y traducción**. Valencia: Universitat de València, 2003. p. 420–432.

SANTAEMILIA, José (ORG.). **Género, Lenguaje y Traducción: actas del Primer Seminario Internacional sobre Género y Lenguaje**. Valencia: Guada Impresores, 2003.

SANTAEMILIA, José. Translating international gender-equality institutional/legal texts: The example of 'gender' in Spanish. **Gender and Language**, *[S. l.]*, v. 7, n. 1, p. 75–94, 2013. DOI: 10.1558/genl.v7i1.75.

SANTAEMILIA, José. A Corpus-Based Analysis of Terminology in Gender and Translation Research: The case of Feminist Translation. *Em*: CASTRO, Olga; ERGUN, Emek (org.). **Feminist Translation Studies: Local and Transnational Perspectives**. [s.l.] : Routledge, 2017. a. p. 15–28.

SANTAEMILIA, José (ORG.). **Traducir para la igualdad sexual: hacia una ética activa y responsable**. Granada: Editorial Comares, 2017. b.

WILLIAMS CAMUS, Julia T.; GÓMEZ CASTRO, Cristina; ASSIS ROSA, Alexandra; CAMUS CAMUS, Carmen (ORG.). **Translation and gender.**

Discourse strategies to shape gender. Santander: Cantabria University Press, 2018.

WILLIAMS CAMUS, Julia Teresa. Translation and Gender: Franco, my dear, might give damn. *Em*: ZARAGOZA NINET, Gora; MARTINEZ SIERRA, Juan José; CEREZO MERCHÁN, Beatriz; RICHART MARSET, Mabel (org.). **Traducción, género y censura en la literatura y en los medios de comunicación**. InterlinguaGranada: Editorial Comares, 2018. p. 191–204.

ZARAGOZA NINET, Gora; MARTINEZ SIERRA, Juan José; CEREZO MERCHÁN, Beatriz; RICHART MARSET, Mabel (ORG.). **Traducción, género y censura en la literatura y en los medios de comunicación**. Granada: Editorial Comares, 2018.

QUÃO CONFIÁVEIS SÃO AS FERRAMENTAS DE IA PARA A TRADUÇÃO DE RECEITAS CULINÁRIAS? ALGUMAS SURPRESAS

Stella E. O. TAGNIN[85]
Rozane R. REBECHI[86]

Introdução

Desde a proposta inovadora de Warren Weaver em 1949 (Weaver 1955 (1949)) de usar computadores para realizar traduções de forma automática, os tradutores temeram perder sua função. No entanto, ao longo dos tempos, se deram conta de como essas ferramentas podem agilizar suas tarefas, embora o produto necessite de correções e ajustes, a chamada pós-edição. A área evoluiu empregando vários sistemas até chegar à Inteligência Artificial. Em 2018 a OpenAI lançou o modelo GPT, que apresentou avanços significativos em termos de qualidade, tanto na tradução automática quanto em outras tarefas de Processamento de Linguagem Natural (PLN). Nosso objetivo é avaliar o desempenho de algumas dessas ferramentas na tradução de uma receita culinária baseando-nos em dois corpora comparáveis em inglês e português, um de Culinária Geral e outro de Culinária Brasileira.

Metodologia e aporte teórico

As ferramentas a serem analisadas são o Google Tradutor, o Microsoft Bing, o Deep-L da Microsoft e o Chat GPT da OpenAI. O objeto de investigação serão as traduções produzidas por essas ferramentas para o português e o inglês de uma das 790 receitas do livro *La Scienza in Cucina e l'Arte di Mangiar Bene* de Pellegrino Artusi (1891), considerado revolucionário na época e que pretendia popularizar as tradições culinárias das várias regiões da Itália.

Na análise proposta serão especialmente considerados termos da culinária sob a ótica das noções de adequação e aceitabilidade, conforme Toury (1995). Por adequação, entende-se uma tradução que mais se aproxima da língua de partida, enquanto a aceitabilidade privilegia uma tradução que melhor se insere na língua de chegada.

O original juntamente com as traduções publicadas em inglês (ARTUSI, 2004) e em português (ARTUSI, 2009) formam um corpus paralelo, que servirá de base para nossos comentários.

Análise

A análise das traduções será feita primeiramente para o português e em seguida para o inglês. Foi selecionada uma receita curta no formato usual do autor, neste caso com uma anedota introdutória, sem uma lista de ingredientes e incluindo um verso.

---

[85] Professora Associada, Universidade de São Paulo, São Paulo, SP
[86] Professora Adjunta, Universidade Federal do Rio Grande do Sul

A receita original, em italiano, é a seguinte:

276. PICCIONI IN UMIDO
A proposito di piccioni sentite questa che vi do per vera, benché sembri incredibile, e valga come riprova di ciò che vi dicevo sulle bizzarrie dello stomaco
Una signora prega un uomo, che le capita per caso, di ucciderle un paio di piccioni, ed egli, lei presente, li annega in un catino d'acqua. La signora ne ricevé una tale impressione che d'allora in poi non ha più potuto mangiar la carne di quel volatile.
Guarnite i piccioni con foglie di salvia intere, poneteli in un tegame o in una cazzaruola sopra a fettine di prosciutto grasso e magro e conditeli con olio, sale e pepe. Quando essi avranno preso colore, aggiungete un pezzo di burro e tirateli a cottura con brodo. Prima di ritirarli dal fuoco spremeteci sopra un limone e adoperate il loro sugo per servirli con fette di pane arrostito postevi sotto. Avvertite di salarli pochissimo a motivo del prosciutto e del brodo. Al tempo dell'agresto,potete usare quest'ultimo invece del limone, seguendo il dettato:
Quando Sol est in leone
Bonum vinum cum popone
Et agrestum cum pipione

Apesar de a introdução à receita ser bastante ilustrativa do estilo do autor, por limitação de espaço, e pelo fato de não estar diretamente ligada ao gênero ‘receita culinária’, não será aqui analisada.

Vejamos as traduções dos títulos em português:

GoogleTranslator	Microsoft Bing	Deep-L	ChatGPT
POMBOS NO MOLHADO	POMBOS COSTURADOS	POMBOS ESTUFADOS	POMBOS EM MOLHO

Todas as ferramentas traduziram o ingrediente principal de forma correta, contudo não houve concordância em relação ao modo de preparo da ave. O Google Translator propôs uma tradução inadequada para uma receita culinária, mas fazendo referência à suculência do prato; o Microsoft Bing ofereceu uma tradução equivocada do termo *in umido*; o Deep-L forneceu um equivalente adequado na variante portuguesa, ‘estufado’, apesar de ter sido selecionada a variante brasileira para a pesquisa; já o ChatGPT apresentou uma tradução adequada, do ponto de vista da culinária. Contudo, de acordo com pesquisa nos corpora citados, observamos que a fraseologia convencional se dá com ‘ao’ e ‘com’, em geral acompanhada da preposição ‘de’, para fazer referência ao ingrediente principal desse molho, como, por exemplo, ‘ao/no molho de laranja’. Quando não se faz referência ao tipo de molho, costuma-se adotar o termo ‘ensopado’, estratégia utilizada na tradução publicada – ‘pombos ensopados’. Contudo, os corpora nos mostraram que é comum utilizar a forma singular, ainda que se refira a mais de uma unidade. Assim, julgamos que uma tradução adequada para o título da receita seria ‘pombo ensopado’.

A receita em análise não apresenta uma lista de ingredientes, como as receitas contemporâneas (TAGNIN, REBECHI e TEIXEIRA, 2022). Os ingredientes são incluídos diretamente no preparo do prato, sem menção às quantidades necessárias.

As ferramentas utilizadas mantiveram o formato original de apresentação dos ingredientes na tradução para o português.

Quanto à tradução dos recipientes adequados para a cocção – *tegame* ou *cazzaruola* –, as sugestões foram:

Google Translator	Microsoft Bing	Deep-L	ChatGPT
tacho ou tacho	panela ou lixo	panela ou caçarola	panela ou caçarola

Observamos, portanto, que a primeira simplesmente repete o utensílio, possivelmente por não ter conseguido identificar uma diferença entre eles. A segunda oferece para o termo *cazzaruola* uma tradução absolutamente equivocada e inadequada para o gênero, enquanto as duas últimas, além do termo genérico 'panela', propõem também 'caçarola', tipo de panela com duas alças laterais. Vale ressaltar que essas duas traduções, apesar de fiéis ao texto de partida, desconsideram que as receitas em português brasileiro costumam ser menos detalhadas (REBECHI, 2015), apoiando-se no senso comum do leitor para decidir qual o tipo apropriado de recipiente usar. Na versão publicada, os termos *tegame* e *cazzaruola* foram traduzidos por 'frigideira' e 'caçarola', respectivamente.

Passemos às traduções dos títulos para o inglês:

Google Translator	Microsoft Bing	Deep-L
WET PIGEONS	STEWED PIGEONS	STEWED PIGEONS

Nota-se que o Google Translator fez uma tradução literal, totalmente inadequada do ponto de vista culinário. As duas outras traduções são perfeitamente adequadas.

A primeira tradução produzida pelo Chat GPT (aqui numerada como 0) causou tanta surpresa que foram pedidas outras, mas essa parte será discutida mais adiante. É apenas mencionada aqui para explicar porque há, na realidade, cinco traduções do Chat GPT.

Chat GPT 0	Chat GPT 1	Chat GPT 2	Chat GPT 3	Chat GPT 4
PICCIONI IN UMIDO (Stewed Pigeons)	276.PICCIONI IN UMIDO (STEWED SQUABS)	STEWED PIGEONS	STEWED PIGEONS	STEWED PIGEONS

Quanto ao título, observa-se que nas versões Chat GPT 0 e Chat GPT 1 ele foi mantido em italiano e a tradução dada entre parênteses, num claro esforço de adequação, por manter o título no original, mas também de aceitabilidade, para se aproximar do público leitor.

As outras versões apresentam apenas a tradução *Stewed Pigeons*. Cabe ressaltar, entretanto, que a tradução norte-americana publicada traz como título *Stewed Squabs*. Enquanto *pigeons* são mesmo 'pombos', *squabs* são pombos jovens que ainda não desenvolveram penas. No português, salvo engano, essa diferença não é lexicalizada.

O que realmente chamou nossa atenção foi a primeira tradução do Chat GPT, que apresentou os ingredientes em formato de lista:

Ingredients:
- Whole sage leaves
- Pigeons
- Slices of fatty and lean ham (prosciutto)
- Olive oil
- Salt and pepper
- Butter
- Broth
- Lemon

Na receita em italiano, como se observa acima, os ingredientes são mencionados no decorrer da explicação de como elaborar o prato. Pareceu-nos surpreendente a ferramenta 'extrair' os ingredientes do texto e apresentá-los no formato usual de uma receita contemporânea. O mesmo ocorreu com o modo de preparo, que também foi apresentado em formato de lista.

Foi essa tradução que nos levou a solicitar mais algumas ao Chat GPT para verificar se o fenômeno se repetia, mas as outras quatro versões mantiveram o texto corrido. Por outro lado, na comparação dessas versões com a versão publicada em livro, a ChatGPT 1 mostrou-se idêntica à publicada, sem qualquer indicação – nem referência – de que essa tradução seja a publicada.

Na parte referente ao modo de preparo há algumas 'traduções' que nos chamaram a atenção. Em italiano, Artusi indica uma *tegame o [...] una cazzaruola* para cozinhar a ave. As opções das respectivas ferramentas foram:

Google Tradutor	Microsoft Bing	DeepL	ChatGPT 0	ChatGPT 1	ChatGPT 2	ChatGPT 3	ChatGPT 4
saucepan or saucepan	a pan or a garbage	a pan or trough	pot or casserole dish	pot or saucepan	pot or casserole	pan or casserole dish	pan or casserole dish

Desnecessário salientar a inadequação tanto das traduções do Google Tradutor, que simplesmente repete *saucepan*, quanto do Microsoft Bing, que traduz *cazzaruola* por *garbage* ('lixo'). A tradução do Deep-L também é inaceitável, uma vez que *trough* designa um 'cocho', ou seja, onde é colocado o alimento para os animais.

Dentre as opções oferecidas pelo ChatGPT, as mais adequadas nos parecem ser as das versões 3 e 4, *pan or casserole dish*, porém todas são aceitáveis. Como a ave, depois de dourada, deve ser cozida com caldo, justifica-se *pot* por ser um recipiente mais fundo, e mesmo *saucepan*, uma panela comum.

Considerações Finais

A análise salientou a inadequação das traduções do Google Tradutor em todos os itens analisados nas duas línguas. O Microsoft Bing apresentou resultado similar, com exceção do título em inglês, que foi adequado. O Deep-L teve desempenho satisfatório, exceto no título em português, quando o produziu na variante portuguesa. As surpresas ficaram a cargo do ChatGPT que, na

versão 0, extraiu a lista de ingredientes de um texto corrido, assim como apresentou o modo de fazer em tópicos. Outra surpresa foram os títulos traduzidos para o inglês. A versão 0 manteve o título original e acrescentou a tradução Stewed Pigeons, a versão 1 trouxe o título original com a tradução Stewed Squabs, exatamente como a tradução publicada. Outras ferramentas, como o Gemini e o Copilot, serão discutidas na apresentação oral.

Referências

ARTUSI, Pellegrino. *A Ciência na Cozinha e a Arte de Comer Bem*. Tradução de Marusca Oliva Bertolozzi e Anabela Cristina Costa da Silva Ferreira. Salto e Itu, SP:
Associação Emiliano Romagnoli Bandeirante, 2009.

----- *La Scienza in Cucina e l'Arte di Mangiar Bene*. Firenze: Salvadore Landi, 1891. ----- *Science in the Kitchen and the Art of Eating Well*. Toronto, Buffalo, London: University of Toronto Press, 2004.

REBECHI, Rozane Rodrigues. *A Tradução da Culinária Típica Brasileira para o Inglês: um estudo sob o Enfoque da Linguística de Corpus*. 2015. 393 p. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.

TAGNIN, Stella E. O.; REBECHI, Rozane R.; TEIXEIRA, Elisa D. A fraseologia das receitas culinárias – com destaque para as brasileiras. In NOVODVORSKI, A.; BEVILACQUA, C. (org.) *Fraseologia: enfoques especializados e contrastivos* Uberlândia: Universidade Federal de Uberlândia, p. 441-473, 2022.

TOURY, Gideon. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins, 1995.

WEAVER, W. Translation. In LOCKE, W. N. and BOOTH, A. D. (eds.) *Machine translation of languages: fourteen essays.* Cambridge, Mass.: Technology Press of The Massachusetts Institute, p. 15-23. 1955 (1949).

UD_NHEENGATU-COMPLIN: O CORPUS SINTATICAMENTE ANOTADO DO NHEENGATU DA COLEÇÃO *UNIVERSAL DEPENDENCIES*

Leonel Figueiredo de ALENCAR[87]

ABSTRACT: This paper introduces the latest version of UD_Nheengatu-CompLin, the treebank of the Universal Dependencies collection for Nheengatu, a Brazilian Indigenous language threatened with extinction. It is the largest and has the highest evaluation grade among the 21 Amerindian languages in this collection.

Palavras-chave: linguística computacional; nheengatu; *treebank*; tupinologia; *parsing* sintático.

A linguística de corpus e o processamento de linguagem natural (PLN) desenvolveram-se muito nos últimos 20 anos no Brasil. Essas disciplinas, contudo, têm focado entre nós quase que exclusivamente o português e outras línguas majoritárias, ignorando a enorme diversidade de línguas indígenas, salvo iniciativas mais recentes, como Galves *et al.* (2017) e Rodríguez *et al.* (2022). Este trabalho relata sobre o atual estágio de um esforço iniciado há três anos de inclusão digital do nheengatu ou Língua Geral Amazônica (LGA), por meio da construção do UD_Nheengatu-CompLin, o primeiro *corpus* sintaticamente anotado (*treebank*) dessa língua, visando tanto investigações linguísticas computacionalmente embasadas quanto a implementação de um *parser* sintático neural.

A LGA foi língua oficial do Maranhão e Grão-Pará de 1689 a 1727 e, até meados do século XIX, sobrepujava o português na região norte (RODRIGUES, 1996; CRUZ, 2011, 2015; FREIRE, 2011; MOORE, 2014; NAVARRO, 2016). Atualmente, encontra-se ameaçada de extinção, não obstante 6000 falantes no município brasileiro de São Gabriel da Cachoeira e 8000 na Colômbia (EBERHARD; SIMONS; FENNIG, 2024). Várias características a distinguem no quadro das línguas indígenas brasileiras. Em primeiro lugar, não se restringe a uma única etnia. Em São Gabriel da Cachoeira (AM) é a língua materna dos barés, uarequenas e baníuas, originalmente de línguas aruaques. Nunca foi língua tribal, tendo emergido do tupinambá, uma das variedades do tupi, pela sua utilização como língua geral por portugueses e seus descendentes mestiços e membros de inúmeras etnias incorporadas ao sistema colonial. É a principal língua adotada em iniciativas de (re)vitalização em diversas regiões do país como meio de afirmação de identidade étnica, contando ainda com um crescente número de traduções de clássicos da literatura universal realizadas por não indígenas (NAVARRO; AVILA; TREVISAN, 2017).

Todos esses fatores concorreram para tornar o nheengatu a língua indígena brasileira, segundo parece, com o maior volume de registros escritos, que permitem acompanhar seu desenvolvimento histórico desde o século XVII,

[87] Professor Titular do Departamento de Letras Estrangeiras e do Programa de Pós-graduação em Linguística da Universidade Federal do Ceará, Fortaleza, CE.
E-mail: leonel.de.alencar@ufc.br

com uma produção notável na segunda metade do século XIX e início do século XX. Não obstante isso, antes da iniciativa objeto deste trabalho, não havia nenhum *corpus* sintaticamente anotado do nheengatu. Além de um certo descaso da área de PLN pelas línguas indígenas brasileiras, desprovidas de apelo comercial, vários outros fatores contribuíram para esse estado de coisas. Em primeiro lugar, constitui um empecilho ao desenvolvimento de recursos e ferramentas de PLN a grande disparidade de ortografias com que a LGA tem sido registrada ao longo dos séculos. Além disso, a maior parte das publicações só está disponível em papel ou em arquivos que demandam transcrição manual.

Desse quadro resultou uma situação de estagnação que perdurou até recentemente: sem ferramentas de PLN, a língua não dispunha de *corpora* anotados, o que, por sua vez, impedia o treinamento de modelos por meio de aprendizado de máquina supervisionado. Para romper esse ciclo, iniciamos há cerca de quatro anos um projeto de construção de recursos computacionais para o nheengatu, que culminou na implementação, em 2022, do Yauti, um analisador sintático baseado em regras (ALENCAR, 2023), utilizado na anotação do UD_Nheengatu-CompLin.

O UD_Nheengatu-CompLin conforma-se ao modelo Dependências Universais (doravante UD) (MARNEFFE *ET AL.*, 2021), aparentemente o mais difundido para anotação sintática de *corpora*. De fato, a coleção UD cresceu de 10 *treebanks* de 10 línguas na versão 1.0, de 15.01.2015, para 283 *treebanks* de 161 línguas na versão 2.14, de 15.05.2024, que inclui 21 *treebanks* de línguas ameríndias, 14 das quais do Brasil. Uma razão da popularidade de UD é seu foco tanto no processamento computacional quanto na tipologia linguística. Outra vantagem decisiva é a disponibilidade de uma variada gama de ferramentas gratuitas para edição, visualização e manipulação de *treebanks*, treinamento de analisadores sintáticos neurais (STRAKA; STRAKOVÁ, 2017) etc.

Entre os 21 *treebanks* de línguas ameríndias da versão mais recente da coleção UD, o UD_Nheengatu-CompLin sobressai em diversos parâmetros quantitativos. Possui, por exemplo, 27,74% mais palavras do que o UD_Mbya_Guarani-Dooley, o segundo maior com base nesse critério. A versão de desenvolvimento do UD_Nheengatu-CompLin supera as versões correspondentes dos dois outros maiores *treebanks* de línguas tupis em várias estatísticas computadas pela ferramenta `conllu-stats.pl` do projeto UD (Tabela 1). O *treebank* do nheengatu é o único de língua ameríndia que tem crescido significativamente a cada versão semestral da coleção UD, desde quando estreou na versão 2.11, de 15.11.2022, com apenas 196 sentenças, perfazendo 2146 palavras (ALENCAR, 2024). Da versão 2.14 da coleção UD para a atual versão de desenvolvimento, o número de palavras e sentenças aumentou em 27,17% e 24,10%, respectivamente.

Tabela 1: Dados quantitativos das versões de desenvolvimento do UD_Nheengatu-CompLin (UNC), UD_Guajajara-TuDeT (UGT) e UD_Mbya_Guarani-Dooley (UMD) em 27.09.2024.

Treebank	**Sentenças**	**Palavras**	***POS-tags***	**Lemas**	**Formas**	**Relações de dependência**	***Features***
UNC	1824	19122	16	1447	2109	37	82

UGT	1182	9160	15	593	1314	29	72
UMD	1046	11771	16	103	114	34	44

Fonte: Elaboração própria.

Construímos o UD_Nheengatu-CompLin incrementalmente, começando com sentenças de estrutura mais simples, incorporando progressivamente fenômenos mais complexos. A ferramenta inicial disponível era apenas um analisador morfológico baseado num léxico computacional derivado do glossário de Navarro (2016). Implementamos regras em Python para projetar as árvores dependenciais a partir da análise morfológica, levando em conta características sintáticas gerais da língua, como a ordem básica SVO e a posição final de adposições, subordinadores e do núcleo nominal da construção genitiva (ALENCAR, 2023). Em seguida, expandimos progressivamente o glossário e o analisador morfológico com dados extraídos de Avila (2021), aprimorando o analisador sintático por meio da sua aplicação a sentenças representativas de um espectro cada vez mais amplo de fenômenos gramaticais.

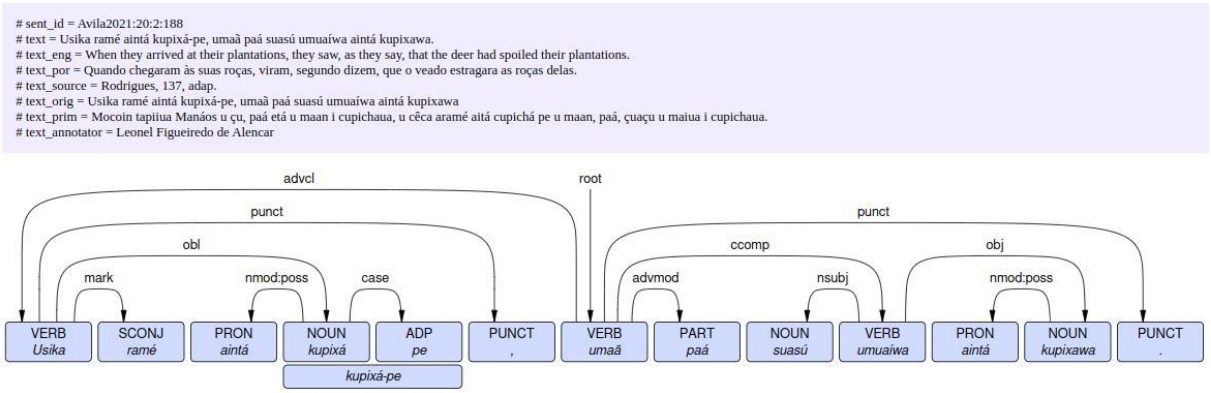
Figura 1: Aplicação do analisador Yauti a exemplo extraído de Avila (2021).

```
Python 3.8.10 (default, Sep 11 2024, 16:02:53)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
>>> import Yauti
>>> s=''' Yautí uyana piri/advg se suí? (Muniz, 84, adap.) -  O jabuti corre mais do que eu?'''
>>> Yauti.parseExample(s,'Avila2021',0,0,190,check=False)
# sent_id = Avila2021:0:0:190
# text = Yautí uyana piri se suí?
# text_eng = Does the tortoise run faster than me?
# text_por = O jabuti corre mais do que eu?
# text_source = Muniz, 84, adap.
# text_annotator = Leonel Figueiredo de Alencar
# inputline = Yautí uyana piri/advg se suí?
1       Yautí   yautí   NOUN    N       Number=Sing     2       nsubj   _       TokenRange=0:5
2       uyana   yana    VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       TokenRange=6:11
3       piri    piri    ADV     ADVG    AdvType=Deg     2       advmod  _       TokenRange=12:16
4       se      se      PRON    PRON2   Case=Gen|Number=Sing|Person=1|PronType=Prs      2       obl     _       Tok
enRange=17:19
5       suí     suí     ADP     ADP     AdpType=Post    4       case    _       SpaceAfter=No|TokenRange=20:23
6       ?       ?       PUNCT   PUNCT   _       2       punct   _       SpaceAfter=No|TokenRange=23:24
```

Fonte: Elaboração própria.

Um total de 58,9% das sentenças do *treebank* são exemplos isolados, a maioria das restantes integram blocos de duas, três ou mais sentenças que constituem trechos contínuos dos textos dos quais foram extraídas, incluindo uma lenda inteira de Magalhães (1876), as 12 lendas de Casasnovas (2006) em sua quase totalidade, a transcrição de uma conversa entre dois falantes, extraída de Moore, Facundes e Pires (1994), e quatro textos de Navarro (2016). Quase 40% dos exemplos provêm de Avila (2021). Este dicionário contém mais de 4000 abonações em ortografia normalizada, que basta copiar e colar no IDLE, ambiente de desenvolvimento de Python, para realizar a análise sintática dependencial por meio do Yauti (Figura 1). Navarro (2016) e Casanovas (2006) contribuem cada um com 11,8% e Cruz, com 6,6%.

Figura 2: Análise do UD_Nheengatu-CompLin para exemplo extraído de Avila (2021).



Fonte: Elaboração própria.

O restante das sentenças do *treebank* distribui-se entre 18 outras publicações, a maior parte do século XIX e início do século XX. Metadados informam a procedência de cada sentença e se integra um bloco (e, nesse caso, qual a sua posição relativa dentro do bloco) ou constitui exemplo isolado (Figura 2). Conquanto limite o potencial de utilização do *treebank* para investigações quantitativas ou de cunho discursivo, a atual composição do *treebank* não impede que venha a ser utilizado com proveito para investigações qualitativas de caráter lexical, morfológico ou sintático ou para treinamento de um *parser* neural (ALENCAR, 2024).

O UD_Nheengatu-CompLin atende a todos os critérios do validador `validate.py`, a que são submetidos os *treebanks* da coleção UD a cada *release*. Esta ferramenta verifica não apenas a obediência às especificações de formato, mas também aspectos da consistência com o esquema de anotação da teoria UD. Na versão 2.14 da coleção UD, o UD_Nheengatu-CompLin, o UD_Mbya_Guarani-Thomas e os dois de variedades regionais do nahuatl são os únicos de línguas ameríndias com a avaliação de duas estrelas, numa escala de zero a cinco (ALENCAR, 2024). Essa classificação, computada pela ferramenta `evaluate_treebank.pl`, leva em conta uma série de fatores, como disponibilidade, diversidade de gêneros textuais e quantidade de erros computados pela ferramenta `udapy`. Atualmente, a versão de desenvolvimento do UD_Nheengatu-CompLin é a única a obter 3,5 estrelas, representando uma melhora de 75%. As notas das versões de desenvolvimento dos demais *treebanks* de línguas ameríndias variam entre 0 e 2 estrelas. O alto desempenho na validação e avaliação automáticas, porém, não assegura que cada sentença do UD_Nheengatu-CompLin tenha sido analisada da melhor maneira conforme o modelo UD. Por enquanto, apenas 30,54% das análises foi revisada por um anotador adicional. Procuraremos sanar essa lacuna numa próxima versão do *treebank*.

Agradecimentos: FAPESP (Processo 22/09158-5).

Referências

ALENCAR, L. F. de. Yauti: A tool for morphosyntactic analysis of Nheengatu within the Universal Dependencies framework. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14, 2023, Belo Hori zonte/MG. *Anais* … Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 135- 145.

ALENCAR, L. F. de. Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo Dependências Universais. *Texto Livre* , Belo Horizonte, v. 17, p. e52653, 2024.

AVILA, M. T. *Proposta de dicionário nheengatu-português*. 2021. Tese (Doutorado em Estudos da Tradução) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2021.

CASASNOVAS, A. Noções de língua geral ou nheengatú: gramática, lendas e vocabulário. 2. ed. Manaus: Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, 2006.

CRUZ, A. *Fonologia e gramática do nheengatú*: a língua falada pelos povos Baré, Warekena e Baniwa. Utrecht: LOT, 2011.

CRUZ, A. The rise of number agreement in Nheengatu. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, Belém, v. 10, n. 2, p. 419-439, 2015.

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (Org.). *Ethnologue*: languages of the world. 27. ed. Dallas: SIL International, 2024. Disponível em: http://www.ethnologue.com. Acesso em: 28 set. 2024.

FREIRE, J. R. B. *Rio Babel*: a história das línguas na Amazônia. 2. ed. Rio de Janeiro: EdUERJ, 2011.

GALVES, C. *et al.* Annotating a polysyntheticlanguage: from Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, v. 59, n. 3, p. 631-648, 2017.

MAGALHÃES, J. V. C. de. *O selvagem*. Rio de Janeiro: Typographia da Reforma, 1876.

MARNEFFE, M.-C. de et al. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255–308, 2021.

MOORE, D.; FACUNDES, S.; PIRES, N. *Nheengatu (Língua Geral Amazônica), its history, and the effects of language contact*. UC Berkeley: Department of Linguistics, 1994. Disponível em: https://escholarship.org/uc/item/7tb981s1 Acesso em: 31 mai. 2023.

MOORE, D. Historical development of Nheengatu (Língua Geral Amazônica). In: MUFWENE, S. S. (Org.). *Iberian imperialism and language evolution in Latin America*. Chicago: University of Chicago Press, 2014. p. 108-142.

NAVARRO, E. A. *Curso de Língua Geral (nheengatu ou tupi moderno)*: a língua das origens da civilização amazônica. 2. ed. São Paulo: Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2016.

NAVARRO, E. A.; AVILA, M. T.; TREVISAN, R. G. O Nheengatu, entre a vida e a morte: a tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, Campina Grande, v. 6, n. 2, p. 9-29, 2017.

RODRIGUES, A. D. As línguas gerais sul-americanas. *Papia*, São Paulo, v. 4, n. 2, p. 6-18, 1996.

RODRÍGUEZ, L. M. et al. Tupian Language Resources: data, tools, analyses. In: MELERO, M.; SAKTI, S.; SORIA, C. (Org.). *Proceedings of the LREC 2022 Workshop of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*. Paris: European Language Resources Association, 2022. p. 48-58.

STRAKA, M.; STRAKOVÁ, J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task*: Multilingual parsing from raw text to Universal Dependencies. Vancouver: Association for Computational Linguistics, 2017. p. 88–99.

LEVANTAMENTO DE COLOCAÇÕES EM BLOGS DE COWORKING: UM COTEJO PRELIMINAR DE TEXTOS AUTÊNTICOS E TRADUZIDOS

Patrícia Helena FREITAG[88]

RESUMO: Este artigo apresenta os resultados preliminares de um estudo sobre colocações (Tagnin, 2013) que se insere nos Estudos da Tradução e utiliza Linguística de Corpus como metodologia. O objetivo é comparar artigos de blogs de coworking em português autêntico e traduzido, verificar se há diferenças nas colocações e investigar os motivos. Partiu-se de listas de n-gramas para identificar as colocações e de linhas de concordância para realizar a análise. Espera-se que os resultados contribuam para a conscientização acerca do uso de colocações consagradas em traduções.

Palavras-chave: blog de coworking; colocações; tradução; convencionalidade; Linguística de Corpus.

Introdução

No setor da tradução especializada, o controle de qualidade é uma etapa importante no fluxo de trabalho das agências de tradução. Existem diversos modelos de avaliação de tradução profissional, como o LISA e o DQF-MQM, em que um avaliador encontra e classifica erros de acordo com categorias e gravidades predefinidos (Portilho, 2019). Uma das categorias costuma ser reservada para inadequações de estilo, englobando traduções literais, estruturas que não soam naturais e combinações de palavras pouco convencionais. Por isso, é importante que os tradutores produzam textos não apenas corretos, mas que soem naturais e convencionais.

A convencionalidade pode ser entendida como o uso rotineiro da linguagem (López-Rodríguez, 2016) e ocorre em diferentes níveis, como o sintático, o semântico e o pragmático (Tagnin, 2013). No nível sintático, atua a combinabilidade, ou seja, a atração de determinadas palavras entre si por uma questão de convenção de uso.

Esse é justamente o nível de interesse neste trabalho, que investiga um tipo específico de combinação de palavras — as colocações — em artigos de blogs de coworking, cotejando as ocorrências em um corpus de português traduzido e um corpus de português autêntico. Como exemplo de colocação neste gênero, temos *trabalho remoto*. Nos corpora de estudo, não ocorrem alternativas com significado semelhante, como *trabalho distante* ou *trabalho afastado*, pois simplesmente não se convencionou usar essas combinações.

Colocações e tradução

As colocações consistem em palavras que coocorrem com maior frequência que o acaso, e Tagnin (2013) observa que pode haver hífen, artigo e/ou preposição entre elas. Existem colocações adjetivas (Adj. + S ou S + Adj.), adverbiais (Adv. + Adj. ou V + Adv.), nominais (S + S) e verbais (V + S ou V +

[88] Doutoranda na linha de pesquisa Estudos do Léxico e da Tradução do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre/RS. patriciafreitag@gmail.com

Adj.). Tagnin (2013) apresenta ainda dois outros tipos de colocações: expressões especificadoras de unidade e coletivos, que não serão abordados aqui.

Como mostram alguns estudos sobre processamento da linguagem, há evidências de que as traduções podem sofrer influência do texto-fonte (HansenSchirra; Nitzke; Oster, 2017). Partimos do pressuposto que essa influência pode acabar afetando os padrões colocacionais na língua-alvo, caso o tradutor considere as palavras individualmente em vez de tratar as combinações como unidades linguísticas convencionais. Caso isso se confirme, é provável que ocorra quebra de convencionalidade na tradução no que diz respeito a colocações. Alguns estudos já mostram evidências de que as traduções contêm padrões colocacionais diferentes daqueles de textos autênticos (Mauranen, 2007), e buscamos investigar isso no gênero blog de coworking.

Corpora do estudo

Este trabalho com base em corpus visa comparar as colocações encontradas em artigos de blogs de coworking escritos originalmente em português e artigos do mesmo gênero escritos originalmente em inglês e traduzidos para o português. Com esse fim, foi compilado um corpus bilíngue paralelo de textos autênticos em inglês e traduzidos para o português e um corpus monolíngue de textos autênticos em português do mesmo gênero. Esse gênero textual foi escolhido por ser relevante nos dias de hoje, em que o trabalho remoto é uma realidade: segundo a Woba, houve um aumento de 63% no número de espaços de coworking no Brasil (Woba, 2023). Além disso, existem empresas internacionais que oferecem espaços de escritório no país e precisam ter seus materiais traduzidos.

O corpus paralelo é formado por artigos de três empresas de coworking internacionais: Spaces, Regus e WeWork. Não foram encontradas outras empresas com publicações de blog em inglês traduzidas para português. Portanto, o corpus paralelo contém 100% do universo possível desse gênero traduzido. Esse corpus contém 335 artigos, sendo que a porção em português contém 12.795 types e 398.039 tokens.

Havia uma disponibilidade maior de blogs de empresas nacionais. Para fins de seleção para o corpus monolíngue, optou-se por desconsiderar blogs com menos de 10 publicações. Foram necessários 535 artigos de seis empresas nacionais (CWK, nex., Trust Coworking, Vip Office, Club Coworking e Coworking Town) para chegar a um número balanceado em termos de tokens (8.437 types e 396.296 tokens) em relação ao corpus de textos traduzidos.

Todos os blogs eram de livre acesso. Os artigos foram consultados manualmente e salvos em arquivos .txt individuais com um nome exclusivo. Os metadados (URL, data de coleta e data de publicação, quando disponível) foram registrados em uma planilha de Excel. Os arquivos .txt dos corpora em português foram carregados na ferramenta Sketch Engine (Kilgarriff *et al*. 2014). Além disso, os arquivos do corpus paralelo foram alinhados com o LF Aligner (Farkas, 2017) e carregados no AntPConc (Anthony, 2017) para servir de apoio na análise.

Extração e organização de colocações

Foram geradas as listas de n-gramas (2 a 4 palavras) para os dois corpora de português. Em seguida, as listas foram exportadas para Excel e reorganizadas em três planilhas: 1) n-gramas que apareciam nos dois corpora; 2) n-gramas que apareciam exclusivamente no corpus de português autêntico; e 3) n-gramas que apareciam exclusivamente no corpus de português traduzido. Como ponto de corte para este trabalho, foram excluídos os itens que apareciam em menos de 5% dos textos dos corpora. Em seguida, foi feita a leitura atenta, linha a linha, identificando as colocações e excluindo os demais itens.

Durante a análise, para entender as semelhanças e diferenças entre o corpus de textos traduzidos e o de textos autênticos, consultamos linhas de concordância no Sketch Engine (para ver o contexto de uso em português) e no AntPConc (para ver o texto em inglês e investigar uma possível influência da língua-fonte); também usamos o recurso Word Sketch do Sketch Engine para visualizar outros possíveis colocados para uma palavra de busca.

Resultados e discussão

Muitas das colocações que ocorrem tanto no corpus de textos traduzidos quanto no de textos autênticos destacam os pontos em comum abordados por empresas nacionais e internacionais de coworking. Estes são alguns exemplos: a) colocações adjetivas: *grandes empresas*, *home office*, *trabalho híbrido*, *trabalho remoto*; b) colocações nominais: *ambiente de trabalho*, *espaço de coworking*, *espaço de trabalho*, *local de trabalho*, *modelo de trabalho*, *sala de reunião*; e c) colocações verbais: *trabalhar em casa*. Nota-se que as empresas de coworking, tanto as nacionais quanto as internacionais, abordam bastante os espaços de trabalho e os modos de trabalhar.

Desses dados, é interessante observar que o estrangeirismo *home office* aparece nos dois corpora. Com base em nosso conhecimento de mundo e de língua, pensamos em outras opções que poderiam expressar esse conceito e as buscamos no corpus de textos autênticos: encontramos 5 ocorrências de *escritório em casa* e nenhuma de *escritório doméstico* ou *escritório residencial*. Os números foram parecidos no corpus de textos traduzidos: 1 ocorrência apenas de *escritório em casa* e nenhuma das outras duas opções. Ou seja, parece que apenas *home office* é amplamente aceito para esse conceito. É importante para o tradutor ter esse conhecimento para que aposte no uso do estrangeirismo, em vez produzir uma tradução pouco convencional simplesmente com a finalidade de evitá-lo.

Quanto às diferenças, um exemplo interessante é *sala privativa*, colocação adjetiva encontrada apenas no corpus de textos autênticos. Para identificar se e como esse conceito aparece no corpus de textos traduzidos, buscamos o lema *sala* no recurso Word Sketch, que mostra palavras que coocorrem com a palavra de busca. Os resultados não foram reveladores, uma vez que as colocações de *sala* + adjetivo foram: *sala comercial*, *sala espaçosa*, *sala pequena*, *sala grande* e *sala interna*. Ou seja, nenhuma dessas colocações se concentra no caráter privativo do espaço. Partimos para mais uma busca com o Word Sketch, dessa vez no corpus de textos traduzidos. Utilizamos a palavra de busca *privativo* e encontramos 50 ocorrências de *escritório privativo*. Isso sugere que, enquanto em textos autênticos se fala em *sala privativa*, nos textos

traduzidos se usa *escritório privativo*. Com auxílio do AntPConc, buscamos *escritório privativo* no corpus traduzido e constatamos que parece haver influência do texto em inglês na tradução, visto que o texto-fonte usava *private office* nesses casos, e é de amplo conhecimento que uma tradução direta de *office* costuma ser *escritório*. Para garantir uma tradução que soe natural, o tradutor poderia optar por uma tradução menos direta de *office* no contexto de *private office*, resultando na colocação *sala privativa*, visto que é a combinação consagrada em textos autênticos em português.

Mais um exemplo de diferença é a colocação adverbial *bem localizado*, com 34 ocorrências no corpus de textos autênticos contra apenas 1 ocorrência no de textos traduzidos. Isso poderia indicar que os coworkings brasileiros estão preocupados com a localização de seus escritórios e que as empresas internacionais não têm essa mesma preocupação. No entanto, isso não parece plausível, pois é de conhecimento geral que as empresas devem levar a localização em conta para serem bemsucedidas. Então, usamos o AntPConc para consultar o texto-fonte em inglês dessa única ocorrência de *bem localizado* no corpus traduzido e investigar as possíveis formas de expressar uma boa localização em inglês. O texto em inglês para *bem localizado* era *well-positioned*. Imaginamos que poderia haver outras ocorrências de *well-positioned* na porção em inglês do corpus paralelo que poderiam ter sido traduzidas de outra forma. Procuramos, então, *well-positioned* na porção em inglês do corpus paralelo, mas não foram encontradas ocorrências. Em mais uma tentativa, buscamos parte dessa colocação: *position**. Nessa última busca, encontramos diferentes estruturas para falar sobre a localização dos escritórios, como *occupies a prime position*, *enjoys a premium position*, *is perfectly positioned*, *it's ideally positioned*, entre outras. Em todas elas, a tradução ficou com o verbo *posicionar*, conjugado conforme adequado gramaticalmente. Voltando para a ideia original de *bem localizado*, continuamos as buscas na porção em inglês do corpus bilíngue, dessa vez buscando por *located*, imaginando que *localizado* poderia ser uma opção que ocorre nas traduções devido à semelhança das palavras em inglês e português. De fato, encontramos três ocorrências de *conveniently located* traduzidas como *convenientemente localizado*. Essas buscas mostram que, quando existem os termos *located* ou *positioned*, a tradução parece ser influenciada pelo texto-fonte e utiliza os cognatos *localizado* e *posicionado*. Porém, no primeiro caso (*localizado*), a influência é positiva, já que gera uma forma convencional no gênero na língua de chegada (como apresentamos no início do parágrafo, *bem localizado* é uma colocação frequente no corpus de textos autênticos). Por outro lado, o segundo caso (*posicionado*), consiste em uma influência negativa, pois trata-se de uma forma atípica no gênero em português. Dessa forma, convém que o tradutor considere a opção *bem localizado*, mesmo que o texto-fonte faça menção a *position*, como em *occupies a prime position* e *is perfectly positioned*.

Considerações finais

O estudo ainda está em andamento. A análise das colocações continuará até o final das listas de n-gramas, sempre buscando as semelhanças e diferenças e procurando entender se o texto-fonte em inglês parece influenciar as traduções.

Com esta análise preliminar, observa-se que este estudo descritivo pode ajudar na conscientização de tradutores profissionais e em formação sobre a importância de se produzir combinações de palavras convencionais na língua-alvo e no gênero trabalhado.

As reflexões aqui levantadas e os processos de extração e análise de colocações podem ser aplicados a outros gêneros, evidenciando a utilidade de corpora para a identificação de combinações convencionais para fins de tradução.

REFERÊNCIAS

ANTHONY, Laurence. **AntPConc**. Versão 1.2.1. Tóquio, 2017. Programa. Download disponível em https://www.laurenceanthony.net/software. Acesso em: 05 maio 2022.

FARKAS, András. **LF Aligner**. Versão 4.21. 2019. Disponível em: https://sourceforge.net/projects/aligner/. Acesso em: 20 jun 2022.

KILGARRIFF, Adam *et al*. The Sketch Engine: ten years on. **Lexicography**, [S.l.], v. 1, n. 1, p. 7-36, jul. 2014. Equinox Publishing. DOI: http://dx.doi.org/10.1007/s40607-014-0009-9. Acesso em: 05 maio 2022

LÓPEZ-RODRÍGUEZ, Clara Inés. Using corpora in scientific and technical translation training: resources to identify conventionality and promote creativity. **Cadernos de Tradução**, [S.l.], v. 36, n. 1, p. 88-120, 26 abr. 2016. Universidade Federal de Santa Catarina (UFSC). DOI: http://dx.doi.org/10.5007/2175-7968.2016v36nesp1p88. Acesso em: 07 set. 2020.

MAURANEN, Anna. Universal Tendencies in Translation. *In*: ANDERMAN, Gunilla; ROGERS, Margaret (ed.). **Incorporating Corpora**: the linguist and the translator. Clevedon: Multilingual Matters Ltd, 2007. Cap. 3, p. 32-48.

PORTILHO, Talita. **Avaliação de tradução nos contextos profissional e pedagógico:** proposta de unidade didática para revisão e avaliação por pares. 2019. Dissertação (Mestrado em Estudos da Tradução) — Universidade Federal de Santa Catarina, Florianópolis, 2019.

TAGNIN, Stella E. O. **O jeito que a gente diz:** combinações consagradas em inglês e português. São Paulo: Disal, 2013.

WOBA. **Censo Coworking**: uma análise Woba do mercado brasileiro. S.L, 2023. Disponível em: https://21669165.fs1.hubspotusercontentna1.net/hubfs/21669165/censo-coworking-woba-

2023%20(1).pdf?utm_medium=email&_hsmi=254376693&_hsenc=p2ANqtz8O5nFxZ6EV_WhB9qQbTIegWh6pWfwYnn61mBIdIpkFoCLVvuAvcTYcVnob07OItjeoptberYS6ibz_8F2RHJuWkLcUYg_sfYbfpA0ufEuKTUHovo&utm_content=254376693&ut m_source=hs_automation. Acesso em: 31 jan. 2024.

ANOTAÇÃO DE CÓRPUS, UM LUGAR PRIVILEGIADO DE OBSERVAÇÃO LINGUÍSTICA: UM ESTUDO DAS APOSIÇÕES DO PORTUGUÊS BRASILEIRO SEGUNDO O MODELO *UNIVERSAL DEPENDENCIES*

Magali Sanches DURAN[89]
Thiago Alexandre Salgueiro PARDO[90]

Resumo: Para a Linguística de Córpus e para o Processamento de Línguas Naturais (PLN), o córpus é uma fonte de conhecimento. A partir dessa constatação, argumenta-se que a anotação de córpus, uma das atividades linguísticas essenciais para o PLN, constitui também uma atividade interessante para outros linguistas, pois permite registrar as análises linguísticas no próprio córpus. A fim de exemplificar como a anotação de córpus pode ser inspiradora para as reflexões linguísticas, discute-se o caso das aposições predicativas à esquerda do sujeito, utilizando a abordagem *Universal Dependencies* para o português.

Palavras-chave: Anotação de Córpus; Aposições Predicativas; PLN; UD; Português.

Os fazeres humanos se modificam em função das tecnologias disponíveis e os estudos linguísticos são um bom exemplo disso. Antes dos anos 90, todo projeto linguístico que não quisesse ser chamado de “linguística de poltrona” tinha que conter um item que se chamava “córpus de estudo”. Era a descrição do conjunto de textos (em papel) sobre o qual o linguista se debruçava para fazer suas análises. Consistia em um esforço de olhar para as realizações da língua, evitando utilizar apenas seu modelo mental. Com a popularização dos computadores, essa acepção da palavra “córpus” incorporou os textos digitais e a linguística de córpus encarregou-se de desenvolver novos métodos para explorá-los e deles extrair conhecimento. Paralelamente, o Processamento de Línguas Naturais (PLN) também se desenvolvia. O PLN adotou os córpus como fonte de conhecimento e passou a utilizar métodos computacionais para “aprender” tarefas envolvendo a língua. Isso só foi possível porque se desenvolveu o que hoje conhecemos como “anotação de córpus”, ou seja, um processo pelo qual atribuem-se etiquetas a partes do córpus de modo a explicitar uma análise humana sobre essas partes.

Há uma dupla responsabilidade para que um córpus anotado propicie um bom aprendizado automático da tarefa. Da parte da linguística, é importante que a anotação seja consistente ao longo de todo o córpus, ou seja, etiquetas iguais sejam atribuídas a fenômenos semelhantes e etiquetas diferentes a fenômenos cuja distinção seja relevante para o projeto. Da parte do PLN, é importante que os métodos utilizados sejam capazes de “capturar” a lógica expressa na anotação, gerando algoritmos que reproduzam automaticamente a anotação humana.

[89] Pesquisadora - Núcleo Interinstitucional de Linguística Computacional (ICMC-USP) - pós-doc C4AI

[90] Professor - ICMC-USP São Carlos-SP (taspardo@usp.icmc.br)

A atividade de anotação de córpus e os métodos de aprendizado de máquina evoluíram muito ao longo das últimas décadas. O aumento da capacidade de processamento dos computadores possibilitou o desenvolvimento de novas abordagens de aprendizado automático, inclusive sem o uso de córpus anotados, chegando-se aos atuais modelos gerativos (como o famoso ChatGPT). Entretanto, a anotação de córpus se mantém relevante para várias tarefas de PLN (por exemplo, o próprio ChatGPT necessitou de córpus anotado para aprender a detectar discursos de ódio) e também para estudos linguísticos, já tendo muitas metodologias testadas (Hovy & Lavid, 2010; Ide & Pustejovsky, 2017).

Neste artigo, apresenta-se um caso de anotação de córpus. O cenário é o de anotação de relações de dependência sintática em um córpus de português brasileiro, o Porttinari-base (Duran et al. 2023). O conjunto de etiquetas escolhido é o do projeto *Universal Dependencies* (UD) (de Marneffe et al., 2021). O fenômeno em foco é o conjunto das aposições que predicam sujeitos, em especial as antepostas.

A proposta da UD, aqui adotada, é fornecer um conjunto de etiquetas aplicável a diversas línguas, de forma a constituir uma base de comparação entre córpus anotados. Na data de escrita deste artigo, a UD tinha 233 córpus anotados em 141 línguas. São previstas 17 etiquetas morfossintáticas e 37 relações sintáticas. O uso dessas etiquetas em português está descrito em dois manuais (Duran, 2021; Duran, 2022) que serviram de base para este artigo. A anotação de dependências sintáticas é feita ligando dois a dois os tokens de uma sentença, de modo que um token seja o *head* da relação e o outro seja o dependente. Cada token pode ser dependente de uma única relação, mas pode ser *head* de várias relações. A anotação das relações de dependência resulta no que chamamos de "árvore sintática de dependências" de uma sentença. Essa árvore (Figura 1) tem como raiz (*root*) o núcleo do predicado da oração principal.

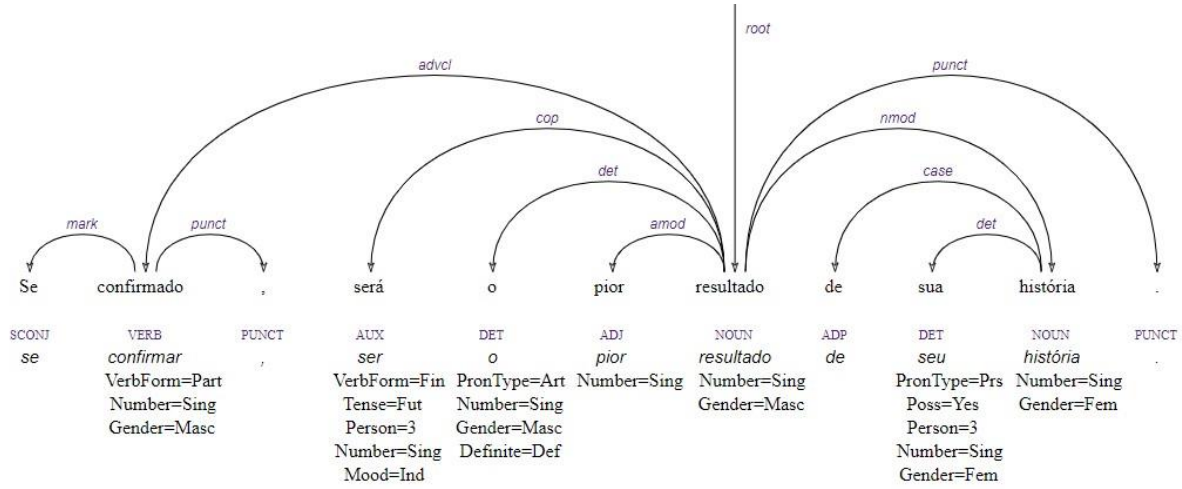


Figura 1 - Exemplo de sentença anotada com etiquetas e diretrizes da abordagem UD

Delineado esse contexto, passamos à discussão das aposições à esquerda do sujeito. Na tradição gramatical do português, as aposições (porções de texto separadas por vírgulas) que predicam o sujeito podem constituir apostos ou orações adjetivas; porém, nos exemplos prototípicos, essas aposições são pospostas aos nominais modificados. Na UD, só são reconhecidas como apostos

relações da esquerda para a direita e entre nominais que sejam intercambiáveis, como nas sentenças 1 e 2 a seguir (aposições em negrito, sujeito sublinhado):

1. O presidente do sindicato dos guardas, **Clovis Pereira**, defende a iniciativa do prefeito.
2. Clovis Pereira, **o presidente do sindicato dos guardas**, defende a iniciativa do prefeito.

Já o exemplo 3 não é considerado aposto na UD:

3. Rodrigo Rocha Loures, **filho de família rica**, achou que valia carregar R$500 mil de Joesley.

A justificativa para isso, na UD, é que, na inversão, o termo à direita, "Rodrigo Rocha Loures", continua sendo sujeito, mostrando que sua função não é intercambiável com a função de "filho de família rica", como fica explícito em (4):

4. **Filho de família rica**, Rodrigo Rocha Loures achou que valia carregar R$500 mil de Joesley.

Poderíamos simplesmente dizer que a restrição de anotação de apostos da UD está equivocada e que temos, sim, apostos à esquerda do sujeito. Contudo, nos deparamos com casos em que o aposto à esquerda está presente e o sujeito está elíptico, como em (5):

5. **Leitor voraz desde garoto,** aos 20 anos começou a vender contos para revistas.

Embora saibamos que "leitor voraz desde garoto" é um predicativo do sujeito elíptico, não haveria um token para ser *head* desta aposição, mesmo que a UD permitisse apostos à esquerda do sujeito. Vamos observar outras aposições à esquerda do sujeito, na forma de adjetivos, como em (6):

6. **Educativas,** brincadeiras clássicas atravessam gerações [...]

Nesse exemplo, o adjetivo "educativas" não tem somente a função de qualificar "brincadeiras", ou seja, não é o mesmo que dizer "brincadeiras clássicas educativas". Poderia se tratar de uma oração adjetiva reduzida de predicativo (sem pronome relativo e sem verbo de cópula), porém orações adjetivas não ocorrem antes de seus antecedentes, o que pode ser observado ao fazermos, para a sentença (6), versões de oração adjetiva desenvolvida posposta ao sujeito (7) e anteposta ao sujeito (8):

7. Brincadeiras clássicas, que são **educativas**, atravessam gerações [...]
8. *Que são **educativas**, brincadeiras clássicas atravessam gerações [...]

A hipótese mais aceitável é a de que a aposição à esquerda do sujeito seja uma oração adverbial, como em (9):

9. [Por serem] **educativas**, brincadeiras clássicas atravessam gerações [...]

Assim, embora as aposições à esquerda do sujeito parecessem, à primeira vista, casos que deveriam ter o sujeito como *head* da relação de dependência, a leitura mais adequada parece ser a de uma oração adverbial.

Essa leitura, por ter o predicado da oração matriz como *head* da relação de dependência, elimina o problema representado pelo sujeito elíptico em (5). Na verdade, essas construções parecem amalgamar dois predicados com um sujeito compartilhado semelhante ao que a tradição gramatical denomina de predicado verbo-nominal. Desdobrados esses dois predicados, teríamos (sujeitos em negrito):

10. **Rodrigo Rocha Loures** é filho de família rica. **Rodrigo Rocha Loures** achou que valia carregar R$500 mil de Joesley.
11. **Ele** é um leitor voraz desde garoto. Aos 20 anos, **ele** começou a vender contos para revistas.
12. **Brincadeiras clássicas** são educativas. **Brincadeiras clássicas** atravessam gerações.

Alguns adjetivos predicativos são mais facilmente identificáveis como orações adverbiais, posto que imprimem uma qualificação circunstancial do sujeito no momento do evento descrito na oração matriz, como no exemplo 13 (aposição em negrito):

13. **Preocupada**, a psicóloga Júlia Prado, 25, pensa em cancelar o pedido.

Essa hipótese ganha força quando testamos o deslocamento da aposição em 14, 15 e 16:

14. A psicóloga Júlia Prado, 25, [por estar] **preocupada**, pensa em cancelar o pedido.
15. A psicóloga Júlia Prado, 25, pensa, [por estar] **preocupada**, em cancelar o pedido.
16. A psicóloga Júlia Prado, 25, pensa em cancelar o pedido, [por estar] **preocupada**.

Em todas essas três versões (14, 15, 16), a possibilidade de uma leitura de “preocupada” como uma oração adverbial reduzida de predicativo[91] (sem conjunção subordinativa e sem verbo de cópula) é factível. Quando a aposição é um sintagma nominal, contudo, como nos exemplos 4 e 5, a versão posposta não é possível, talvez porque um sintagma nominal possa ter outras funções após o verbo e isso pudesse gerar ambiguidade. Contudo, em ambos os casos, a análise seria a mesma, de oração adverbial (**advcl** na UD), como ilustrado nas Figuras 2 e 3:

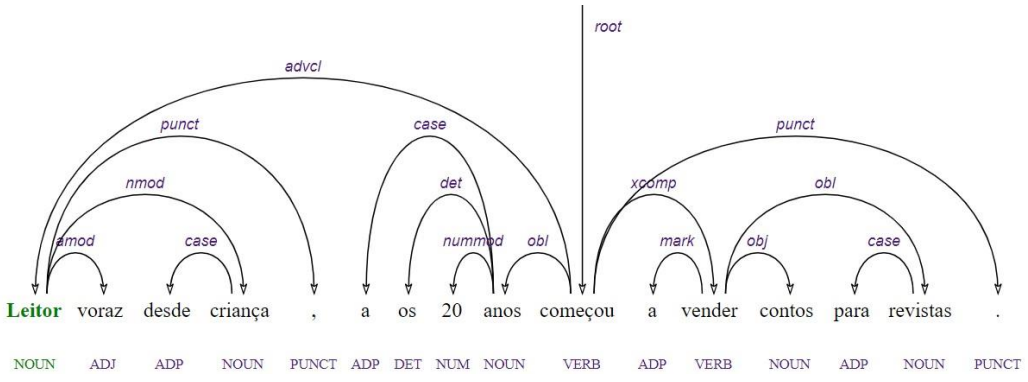


Figura 2 - Aposição (NOUN) à esquerda do sujeito anotada como oração adverbial

[91] Embora as gramáticas só mencionem orações reduzidas de infinitivo, gerúndio e particípio, temos encontrado muitos casos de orações em que o verbo de cópula está elíptico, o que nos levou a empregar o termo “oração reduzida de predicativo”.

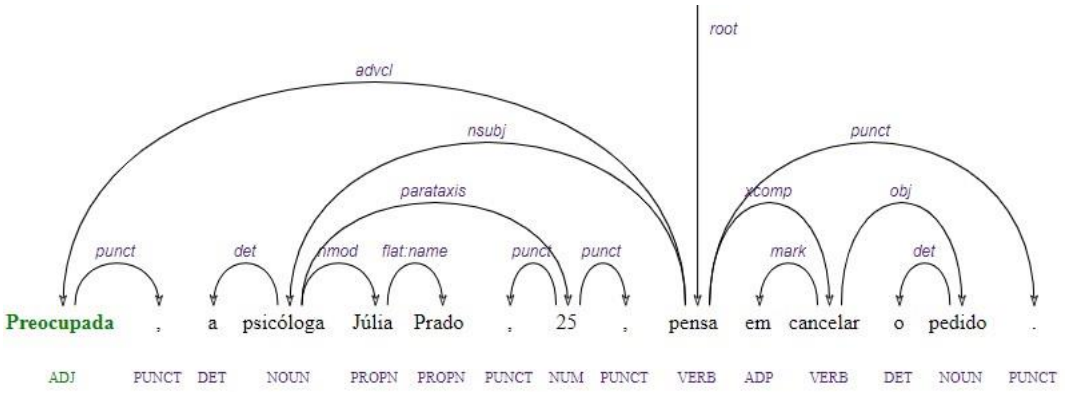


Figura 3 - Aposição (ADJ) à esquerda do sujeito anotada como oração adverbial

A proposta de anotação das aposições à esquerda do sujeito como oração adverbial que modifica outra oração foi implementada recentemente no córpus Porttinari-base. Como esse fenômeno é pouco frequente no córpus e o predicado nominal ora é um adjetivo, ora é um sintagma nominal, há um problema de esparsidade de dados, o que pode prejudicar o aprendizado automático da classificação proposta. O parser treinado na versão anterior desse córpus (Lopes & Pardo, 2024), disponível on-line[92], quando ainda não havíamos adotado um padrão para anotar aposições de sujeito, classifica casos semelhantes ora como oração adverbial (advcl, na UD), ora como adjunto adnominal (amod ou nmod na UD). Em trabalhos futuros, pretendemos anotar mais sentenças que contenham este fenômeno e disponibilizá-las, juntamente com a nova versão do córpus, para retreinamento do parser.

Pelo que pode ser observado, há problemas que emergem durante a tarefa de anotação e que exigem reflexões para que sejam tomadas decisões de anotação. O córpus deixa de ser apenas um lugar para testar hipóteses pré-definidas e passa a ser o próprio *locus* de observação, de levantamento de hipóteses, de reflexão e de armazenamento das decisões tomadas.

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.

Referências

BIBER, DOUGLAS. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In: **The Oxford Handbook of Linguistic Analysis**. Oxford Academic: 2015.

DE MARNEFFE, MARIE-CATHERINE; MANNING, CHRISTOPHER D.; NIVRE, JOAKIM; ZEMAN,

DANIEL. Universal Dependencies. **Computational Linguistics**, 47(2), p.255-308, 2021.

[92] http://portparser.icmc.usp.br:8082/

DURAN, MAGALI SANCHES. (2021). Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). **Relatório Técnico do ICMC n. 434**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

DURAN, MAGALI SANCHES. (2022). Manual de Anotação de Relações de Dependência –Versão Revisada e Estendida. **Relatório Técnico do ICMC n. 440**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

DURAN, MAGALI SANCHES; LOPES, LOPES; NUNES, MARIA DAS GRAÇAS VOLPE; PARDO,

THIAGO ALEXANDRE SALGUEIRO. The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. *In*: **Proceedings of the 14th Symposium in Information and Human Language Technology** (STIL), p. 115-124, 25-29 setembro, 2023.

HOVY, EDUARD; LAVID, JULIA. Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. **International Journal of Translation**, 22(1), p. 13-36, 2010.

IDE, NANCY; PUSTEJOVSKY, JAMES. **The Handbook of Linguistic Annotation**. Springer: 2017.

LOPES, LUCELENE; PARDO, THIAGO ALEXANDRE SALGUEIRO. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. *In* **Proceedings of the 16th International Conference on Computationa Processing of Portuguese** (PROPOR), Vol. 1, p. 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics (ACL): 2024.

DESAFIOS DA LINGUÍSTICA DE *CORPUS* IMPOSTOS PELA INTELIGÊNCIA ARTIFICIAL: REDISCUTINDO ALGUNS CONCEITOS

Jackson Wilke da Cruz SOUZA[93]

RESUMO: Neste trabalho busco discutir impactos e desafios ocasionados pela Inteligência Artificial à Linguística de *corpus*. Tal discussão parece ser emergente frente a diferentes metodologias supervisionadas e não supervisionadas por humanos quanto à mineração e obtenção de dados textuais. Assim, busco recuperar conceitos sobre autenticidade, processamento e representatividade de informações linguísticas em *corpus*, além de indicações para trabalho com dados de redes sociais.

Palavras-chave: Linguística de *corpus*. Inteligência Artificial. Conceitos. *User Generated Content*.

ABSTRACT: In this paper, I aim to discuss the impacts and challenges posed by Artificial Intelligence to Corpus Linguistics. Such a discussion is timely given the emergence of various human-supervised and unsupervised methodologies for mining and obtaining textual data. Therefore, I intend revisit concepts related to the authenticity, processing, and representativeness of linguistic information in *corpora*, as well as providing insights for working with data from social media.

Keywords: *Corpus* Linguistics. Artificial Intelligence. Concepts. User Generated Content.

INTRODUÇÃO

É notório que as redes sociais têm sido fonte dos processos de produção, circulação e recepção de conteúdos que interessam diferentes segmentos sociais. Tais processos foram potencializados com as últimas ondas da Web (Passarelli; Gomes, 2020), fazendo com que o usuário não apenas consumisse, mas também gerasse conteúdo na internet, trazendo à tona o conceito de *User Generated Content* (UGC).

Santos (2022), ao estudar a proeminência de estudos que observam UGC, aponta que há quatro grandes temas de estudo: (1) conteúdo criado pelo usuário, (2) práticas comunicativas, como o jornalismo, (3) estudos sobre usos linguísticos que ocorrem na interlocução entre usuário e audiência e (4) plataforma Web 2.0. O autor ainda destaca que UGC, na literatura mais recente, tem sido foco nas disciplinas de Comunicação social, mídia e jornalismo, Consumo e negócios, Ciência da informação, Ciências humanas e Ciência política.

Nas áreas de Processamento de Línguas Naturais (PLN) e Linguística de *corpus* (LC), UGC tem sido de interesse especialmente em pesquisas que visam ao processamento de diferentes níveis da língua (como o morfológico, sintático e semântico), além de compreender perfis ideológicos e comportamentais de

[93] Docente da Universidade Federal da Bahia no Instituto de Ciência, Tecnologia e Inovação (ICTI) e no Programa de Pós-Graduação em Língua e Cultura (PPGLinC). E-mail: jackcruzsouza@gmail.com

usuários a partir de suas construções textuais. É importante destacar que, nesse contexto, há diferentes vertentes metodológicas sobre as pesquisas em LC. Há trabalhos que desenvolvem pesquisas com *corpus* assíncrono, em que os textos[94] são compilados e processados fora de um ambiente on-line; já outros, com *corpus* síncrono, em que a compilação e o processamento se dão de maneira on-line e contínua; por fim, outros que utilizam a web como corpus, em que há o material linguístico está sendo produzido e pensado de maneira síncrona ou assíncrona.

O que destaco é que em todas essas abordagens há conceitos propostos pela LC que nos serviram de base até aqui. Porém, por mudanças sócio-históricas, precisamos repensar esses conceitos, especialmente porque as fronteiras entre o digital e o não digital estão cada vez mais borradas (Burnham, 2014). Ainda que discussões sobre IA em PLN não sejam novidade, elas se tornaram populares devido ao destaque midiático que se teve por conta dos *Large Language Models* (LLM). Esses modelos podem ser de língua geral, como o GPT (do inglês, *Generative Pretrained Transformer*), ou de domínio, como o BloombergGPT (Wu *et al.*, 2023). Em ambos os tipos, os modelos são treinados a partir de grande quantidade de dados para que processem aspectos das línguas naturais.

Assim, meu objetivo neste trabalho é discutir como conceitos caros à LC, como autenticidade, naturalidade e extensão, requerem novas discussões frente ao cenário de *Big data* e IA. Além disso, destaco o cuidado que os pesquisadores precisam ter em seus estudos para lidar com (meta)dados linguísticos que envolvem informações advindas de UGC.

REVISITANDO CONCEITOS

Podemos dizer que as áreas de PLN e LC são próximas graças ao fator computacional comum a elas. O entendimento que temos hoje sobre objeto e metodologia da LC apresentou diversos avanços com abordagens do PLN com relação ao processamento de dados linguísticos. Os textos que passaram a integrar os corpora deveriam, então, estar em formatos legíveis por máquina (Sardinha, 2000), possibilitando pré-processamento, processamento e análise em larga escala de dados. Como resultado, espera-se que quanto maior o volume de dados, maior seja o conhecimento da língua e sobre os usos que os falantes fazem dela e se constituem a partir dela.

No início da área, os *corpora* mais extensos tinham em torno de 1 milhão de palavras, como o projeto Brown *corpus* que era composto de 500 textos de 2000 palavras em 15 gêneros. Mais tarde, foram publicados outros *corpora* que tinham uma quantidade de palavras muito mais elevada, como o *corpus* multilíngue *News on the Web*, que soma mais de 5 bilhões de palavras (Tagnin, 2018). Esse crescimento se deu por conta de contribuições de distintas áreas da Linguística, como a Lexicografia, Terminologia e a Tradução (Viana; Tagnin, 2015) a partir do desenvolvimento de *corpora* especializados.

---

[94] Neste trabalho tratarei “texto” como material que integra o *corpus*, mesmo admitindo que há outros materiais que podem integrar os *corpora*, como áudio e imagens.

Aqui cabe, então, discutir um primeiro conceito caro à LC: a *representatividade*. Sinclair (1991) defende que um *corpus* representativo deve ser o maior possível, já que a linguagem é um sistema probabilístico (Halliday, 1991) e o *corpus* é "uma amostra de uma população cuja dimensão não se conhece (a linguagem como um todo)" (Sardinha, 2000, p. 342). Tendo esse pressuposto, no âmbito da língua geral, o *corpus* deve ser extenso, para que a amostra possa ser, de fato, representativa.

Nesse sentido, deparamo-nos com ao menos dois desafios. O primeiro é de ordem metodológica. Sabe-se que as áreas de LC e PLN dispõem de métodos bastante eficazes e robustos para a coleta e processamento de textos, como *Web Scraping* (Zhao, 2022). Nesse método, os dados são extraídos da Web e salvos em sistemas de arquivos ou banco de dados para que possam ser analisados ou recuperados posteriormente. Esse método pode ser feito manualmente ou a partir de técnicas computacionais. Neste último caso, a partir da disponibilização de um endereço eletrônico sem que haja as especificações paramétricas corretas, o sistema computacional pode trazer indiscriminadamente todo o conteúdo do site. Neste último caso, não é trivial dizer que os textos são selecionados para os *corpora* devam ser submetidos a uma curadoria em função do objetivo da pesquisa e sobre o que ele representa.

O segundo desafio diz respeito ao processamento, pois a partir da coleta dos dados, será necessário processar as informações linguísticas que foram coletadas. Caso a coleta dos dados linguísticos seja feita por métodos puramente automáticos e não haja nenhum tratamento para classificar os textos entre produzidos por humanos ou por inteligência artificial, as informações que serão extraídas dos dados poderão ser artificiais. Atualmente, há esforços para classificar textos gerados por IA e por humanos, como o trabalho de Ayapova e Skripnikova (2022), que classifica textos jornalísticos. Entretanto, esse tipo de processamento dos dados ainda parece distante dos estudos em LC, pois deveria estar aliado aos métodos utilizados nas pesquisas, como apontado anteriormente.

Outro conceito da LC que destaco aqui: a *autenticidade* dos conjuntos de textos. A literatura sobre LC (Sinclair, 1991; Sardinha, 2000; Freitas, 2024) aponta para a necessidade de os textos que compõem o *corpus* serem autênticos. Isso significa dizer que o material compilado deva ser produzido por humanos. Reunir textos autênticos é garantir que os resultados descritos e analisados refletem, de fato, a língua e como seus falantes a utilizam. Ao extrair informações linguísticas de seus contextos naturais de uso, a depender do objetivo, deve-se prever a compilação de textos produzidos por usuários da Web. Isso se justifica pelo fato de estarmos lidando com "produtores de conteúdo" e não apenas com "usuários de redes sociais".

Porém, diante da possibilidade de existirem produtos disponíveis que podem ser compilados que não foram produzidos intelectualmente por humanos, a autenticidade tornou a ter relevância. Atualmente, sabe-se sobre a existência de obras literárias completas que foram geradas a partir de IA, como discutido por Dalte (2020). Mais recentemente, algumas discussões giram em torno de questões éticas, como os direitos autorais da obra (Garcia, 2020).

O que levanto aqui não é o fato de se os textos artificiais não devam figurar os *corpora*, mas sim a maneira como eles estão agregados à coletânea de textos. Cabe, então, ao pesquisador tomar dois cuidados: (i) caso opte por métodos automáticos, deve-se aplicar métodos suplementares de identificação e classificação de textos (não) naturais que farão parte do conjunto de dados; (ii) caso opte por incluir textos artificiais em seu conjunto de dados, deverá identificá-los e alertar ao consulente sobre essa característica. Textos artificialmente produzidos podem fazer parte do conjunto de dados linguísticos, desde que o propósito sobre esse tipo de inclusão não prejudique descrições linguísticas ou ainda tome algo como verdadeiro.

Por fim, destaco sobre a disponibilização dos dados do corpus. Ao trabalharmos com corpus, especialmente os de UGC, algumas informações serão necessárias para compreender questões sociais intrínsecas à linguagem, como gênero/sexo, idade e localidade, por exemplo. Essas informações são importantes pois auxiliam no monitoramento de perfis acerca de uma temática, observando como dado grupo de usuários se manifestam linguisticamente sobre ela. Porém, ao tratar os dados, é necessário que estratégias de anonimização (como Supressão, Generalização e/ou Pertubação, como trabalhado em Brito e Machado (2017)) sejam aplicadas, sobretudo por estarmos sob a égide da Lei 13.709, conhecida como a Lei Geral de Proteção dos Dados (LGPD). A LGPD atribui obrigações específicas a quem trata os dados quando a base legal é a pesquisa, entendendo tratamento como compilação, análise, estudo e disponibilização dos dados. Assim, é importante que nossas pesquisas ponderem se o corpus completo pode ser disponibilizado (com ou sem anonimização), se contém nele informações sensíveis (como questões relacionadas a raça, gênero, sexualidade, posições políticas, por exemplo) e por quanto tempo os dados podem ficar sob domínio do pesquisador (Almeida, 2021).

CONSIDERAÇÕES FINAIS

Neste artigo meu propósito foi discutir alguns conceitos da LC sob a luz e os desafios impostos pela Inteligência Artificial aos trabalhos atuais com corpora. É quase impossível voltarmos atrás: estamos presenciando um momento em que passamos a produzir conteúdo sobretudo na Web, e esse conteúdo autêntico disputa espaço, muitas vezes, com um conteúdo artificial. O que propus aqui, então, foi uma breve reflexão sobre como esses conteúdos podem e devem figurar em nossos estudos de descrição e de aplicação sobre a linguagem frente aos métodos e ferramentas computacionais.

Há muitos outros desafios que devem ser debatidos nesse campo e muitos outros conceitos da própria LC que merecem ser discutidos, que culminará em uma nova tipologia e organização e compreensão do objeto corpus. Alguns desses desafios já foram enfrentados em outras disciplinas próximas, como a Linguística Aplicada, quanto à anonimização dos dados, por exemplo; quanto a isso, podemos aprender e utilizar com mais facilidade. Porém, há outros que ainda estão surgindo por conta do avanço da IA; para esses, as soluções ainda estão sendo pensadas conforme os desafios surgem.

Por fim, pondero que as pesquisas em LC não se findaram por estarmos desafiados. Muito pelo contrário: temos novos caminhos a trilhar, o que traz novo

fôlego à área e outras perspectivas de pesquisa. Minha contribuição aqui é chamar a atenção para nossas metodologias de pesquisa e as concepções que estamos utilizando sobre determinados conceitos cunhados há algum tempo, antes da evolução da IA. Importa dizer, então, que o mundo que conhecemos pela literatura clássica em LC mudou por conta do computador; nossos conceitos sobre ela também devem ser repensados.

Agradecimentos: Agradeço ao Programa de Pós-Graduação em Língua e Cultura (PPGLinC) da Universidade Federal da Bahia (UFBA)pelo suporte e apoio.

REFERÊNCIAS

ALMEIDA, F.F. **Guia de proteção de dados pessoais**: pesquisa. CEPI FGV Direito SP, 2021.

AYAPOVA, S. M.; SKRIPNIKOVA, A. I. Ai and human created media texts: experiment results. **Herald of journalism**, v. 64, n. 2, 2022. DOI: https://doi.org/10.26577/hj.2022.v64.i2.08

BRITO, F.T.; MACHADO, J.C. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. **Jornadas de atualização em informática**, [*s.l.*], p. 91-130, 2017.

BURNHAM, T.F. Reconstrução-síntese de contribuições ao XI CINFORM, à guisa de apresentação. In BORGES, J; BARREIRA, M.I.J.S.; CUNHA, F.J.A.P. (Org.). **Mundo digital**: uma sociedade sem fronteiras? 1ed.João Pessoa: Ideia, 2014, v. 1, p. 7-20.

DALTE, P. Inteligência artificial e poesia. **Revista 2i: Estudos de Identidade e Intermedialidade**, v. 2, n. 2, p. 165–177, 2020. DOI: https://doi.org/10.21814/2i.2505

FREITAS, C. Dataset e corpus. *In*: CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (Orgs.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. [s.l.]: BPLN - Brasileiras em PLN, 2023, p. 1–37. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-dadoshttps://brasileiraspln.com/livro-pln/2a-edicao/parte-dados-avaliacao/cap-dataset-corpus/cap-dataset-corpus.pdfavaliacao/cap-dataset-corpus/cap-dataset-corpus.pdf>.

GARCIA, A.C. Ética e inteligência artificial. **Computação Brasil**, n. 43, p. 14-22, 2020.

HALLIDAY, M.A.K. Corpus studies and probabilistic grammar. In AIJMER, K.; ALTENBERG, B. (Orgs). **English corpus linguistics**: Studies in honour of Jan Svartvik, London: Longman, 2014. p. 42-55.

PASSARELLI, B.; GOMES, A.C.F. Transliteracias: A Terceira Onda Informacional nas Humanidades Digitais. **Revista Ibero-Americana de Ciência da Informação**, v. 13, n. 1, p. 253–275, 2020. DOI: https://doi.org/10.26512/rici.v13.n1.2020.29527

SANTOS, M.L.B. The “so-called” UGC: an updated definition of user-generated content in the age of social media. **Online Information Review**, v. 46, n. 1, p. 95– 113, 2021. DOI: https://doi.org/10.1108/oir-06-2020-0258

SARDINHA, T.B. Linguística de *corpus*: histórico e problemática. **DELTA: Documentação de Estudos em Linguística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000. DOI: https://doi.org/10.1590/s0102-44502000000200005 SINCLAIR, John. **Corpus, Concordance, Collocation**. [s.l.]: Oxford University Press, USA, 1991.

TAGNIN, S. E. A Linguística de Corpus vai desbravando novos horizontes. In

FINATTO, MJB; REBECHI, T.; SARMENTO, S; BOCORNY, A. EP (Org). Linguística de corpus: perspectivas. Porto Alegre: Instituto de Letras da UFGRS, p. 11-15, 2018.

VIANA, V.; TAGNIN, S. E. *Corpora* na tradução. São Paulo: HUB Editorial, 2015.

WU, S.; IRSOY, O.; LU, S.; *et al*. **BloombergGPT: A Large Language Model for Finance**. arXiv.org. Disponível em: <https://arxiv.org/abs/2303.17564>.

ZHAO, Bo. Web Scraping. *In*: **Encyclopedia of Big Data**. Cham: Springer International Publishing, 2022, p. 951–953. DOI: http://dx.doi.org/10.1007/978-3-319http://dx.doi.org/10.1007/978-3-319-32010-6_48332010-6_483.

SATIRICORPUS.BR: A *CORPUS* OF SATIRICAL NEWS FOR BRAZILIAN PORTUGUESE

Gabriela WICK-PEDRO[95]
Oto Araújo VALE[96]

ABSTRACT: This paper presents **SatiriCorpus.Br**, a corpus of satirical news in Brazilian Portuguese aimed at investigating linguistic differences between satirical and real news. A subcorpus was created to compare satirical news with their real counterparts. This comparison enhances the understanding of the linguistic features that distinguish satirical humor from factual reporting. The findings offer valuable insights for corpus linguistics and natural language processing (NLP).

Keywords: satirical news; corpus linguistics; satire; linguistic features; natural language processing.

1. Introduction

This work presents the SatiriCorpus, a *corpus* of satirical news for Brazilian Portuguese, and a *subcorpus* composed of satirical news and their respective real news versions, with the purpose of investigating the main differences between these two types of content. Additionally, morphosyntactic aspects are analyzed and described, as well as the differences in verbal occurrences between satirical and real news.

Satirical news have a fictional nature and function as parodies of real events and news, usually using humor, irony, exaggeration, and ridicule to criticize social, political, and cultural issues. However, unlike deceptive content, or popularly known as "fake news," which intentionally disseminates false information to deceive, manipulate, and harm or favor certain agendas, satirical news seek to provoke laughter or amusement in their audience (RUBIN; CHEN; CONROY, 2015; WARDLE; DERAKHSHAN, 2018; TANDOC; LIM; LING, 2018).

Although satirical news are often created to be humorous, there is a risk of some people confusing satirical content with factual information. Moreover, such news can intentionally mislead less attentive readers or those without contextual and cultural knowledge into believing what they are reading (RUBIN *et al.*, 2016). Therefore, it is essential for people to be alert and capable of distinguishing between satirical and non-satirical content to avoid the spread of misinformation.

2. Theoretical Framework

Satire is a literary or artistic genre that uses elements such as humor, irony, exaggeration, and ridicule to criticize social, political, cultural, or individual

[95] Pesquisadora em Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília/DF, bolsista FINATEC. E-mail: gabrielawick@ibict.br.

[96] Docente do Departamento de Letras (DL) na Universidade Federal de São Carlos (UFSCAR), São Carlos/SP.

aspects, aiming to sensitize the public and stimulate reflection on important societal issues (KREUZ; ROBERTS, 1993; SIMPSON, 2003; ATTARDO, 2014). This form of expression can be found in various forms and media, such as literary works, theatrical plays, cartoons, comic strips, television programs, and news (LAMARRE; LANDREVILLE; BEAM, 2009; SINGH, 2012).

Simpson (2003) explores satire as a discursive genre and provides a detailed analysis of the elements that compose this type of humor, examining how satire uses linguistic, stylistic, and rhetorical resources to create political and social criticisms in a satirical manner. For this purpose, he also presents a triad for the configuration of satire based on three positions of the discursive subject: the satirist (the producer), the "satiratee" (the recipient), and the satirized (the target).

According to Simpson (2003), the satirist and the satiratee are two "legitimate" participants, while the satirized is unwelcome in satirical discourse. This target can be an individual, an event, an experience, or even a discourse. The author also highlights that the positions of the subjects in the triad are subject to constant shifts and (re)organizations, which may change during the satirical discourse, depending on changes in the focus of criticism or the humor represented. Thus, it is possible to observe the dynamic and fluid nature of satire, where the positions of the participants are constantly transforming to convey their satirical messages.

3. Methodology

3.1. Construction of SatiriCorpus.Br

SatiriCorpus is a *corpus* of satirical news automatically extracted[97] from the Sensacionalista website[98], a Brazilian electronic news program that satirizes various topics, such as politics and entertainment in Brazil. For its construction, a crawler automatically collected news from the Sensacionalista website, extracting the text of the news body from the page and excluding noise such as tags, HTML, and images.

Following the thematic classification established by the website, the *corpus* was divided into five categories: i) behavior (human behavior; daily life in society) with 56 news articles, 8,014 tokens, 6,947 types, and 267 sentences; ii) entertainment (celebrities; topics from Brazilian and international TV) with 1,406 news articles, 255,972 tokens, 218,531 types, and 10,720 sentences; iii) sports (Brazilian and international sports) with 738 news articles, 120,332 tokens, 103,568 types, and 4,931 sentences; iv) world (international politics) with 1,001 news articles, 178,185 tokens, 158,555 types, and 8,250 sentences; v) country (Brazilian politics) with 1,847 news articles, 316,588 tokens, 271,692 types, and

[97] The corpus extraction period was in January 2019, thus considering the beginning of news postings on the Sensacionalista website from 2016 until the end of 2018.

[98] Available at: https://www.sensacionalista.com.br/. Accessed on: January 22, 2023.

12,347 sentences. In total, the *corpus* contains 5,048 news articles, 879,091 tokens, 759,293 types, and 36,515 sentences.

It is important to note that although there are other portals, such as Piauí Herald[99] and O Bairrista[100], which are also dedicated to satirical journalism, the preference for Sensacionalista, founded in 2009, is justified by its status as the main representative of this type of content in Brazil.

3.2. Construction of the subcorpus

In accordance with the criteria established by Rubin, Chen, and Conroy (2015) for the construction of a fake news corpus, particular emphasis is placed on the alignment between fake and real news, with the objective of verifying positive and negative instances and validating linguistic patterns. Thus, the SatiriCorpus, described in the previous section, was divided into a *subcorpus* composed of 300 news articles, with 150 satirical news articles randomly selected from the "country" category and 150 real news articles related to the satirical news. For the real news, the collection was done manually, first delimiting keywords identified in the satirical news and then manually searching for each real news equivalent to the satirical news.

The *subcorpus* contains 22,993 tokens, 4,843 types, and 1,212 sentences for satirical news, and 107,133 tokens, 11,304 types, and 5,721 sentences for real news. In total, there are 130,096 tokens, 16,147 types, and 6,933 sentences.

Additionally, the morphosyntactic information comes from the parser PALAVRAS (BICK, 2000). Besides syntactic annotations, the tool marks the grammatical class for each word. There are 15 classes in total: adjective, adverb, determinant, compound element, interjection, coordinating conjunction, subordinating conjunction, noun, numeral, personal pronouns, proper nouns, preposition, specifiers, and verbs.

4. Discussion

Based on the data extracted by PALAVRAS (BICK, 2000), the average between grammatical class and the total number of words was calculated. There is a balance of grammatical classes between satirical and factual news. The use of adverbs (5.69% in satirical news and 4.24% in real news), determinants (10.30% in satirical news and 9.55% in real news), and verbs (17.98% in satirical news and 15.06% in real news) occurs proportionally more in satirical news, while prepositions (18.35% in real news and 17.50% in satirical news) and punctuation (15.42% in real news and 13.12% in satirical news) are more relevant in real news.

The analysis of verb tenses was also conducted to find specific characteristics between the news types. However, there is no verb tense with a higher predominance in relation to the news; there is only a higher percentage of infinitive in satirical news (21.35%) compared to real news (16.45%). One

---

[99] Available at: https://piaui.folha.uol.com.br/. Accessed on: January 22, 2023.

[100] Available at: https://obairrista.com/. Accessed on: January 22, 2023

possibility is that real news tend to use more auxiliary verbs compared to satirical news, but the PALAVRAS parser does not have a specific tag for auxiliary verbs. The percentage ratio was calculated between the frequency of each verb tense and the total number of verbs annotated by the parser.

Regarding the analysis of verbal persons, it was expected that real news would have a higher incidence of verbs in the third-person singular and plural because, as Tavares (1997, p. 130–131) indicates, "journalistic text is characterized by the impersonality of the subject," meaning that "the verbal person that refers to the referent (the one being talked about – 'he,' 'they') allows the text to be more objective." The author also points out that the use of the first and second person is not expected in journalistic texts because they make the text more subjective and personal. Thus, based on the data obtained by PALAVRAS, it is noted that satirical news, although not based solely on reality, have a higher incidence of verbs in the first and third-person singular, while real news have more verbs in the first and third-person plural.

5. Conclusions

This study presented the SatiriCorpus, a *corpus* of news for Brazilian Portuguese, and investigated morphosyntactic patterns present in satirical and real news to compare linguistically how they behave.

It is understood that the characteristics extracted by the PALAVRAS parser (BICK, 2000) did not present very significant results, but they can still be seen as indications of a satirical news, such as the higher verbal and adverbial incidence.

As future work, it is hoped to create a parallel *corpus* of real news for the remaining 4,898 satirical news.

References

ATTARDO, Salvatore. Encyclopedia of humor studies. Los Angeles: SAGE Reference, 2014.

BICK, Eckhard. The Parsing System Palavras: Automatic Grammatical Analysis. Aarhus Denmark; Oakville, Conn: Aarhus University Press, 2000.

KREUZ, Roger J.; ROBERTS, Richard M. On satire and parody: The importance of being ironic. *Metaphor and Symbolic Activity*, Routledge, v. 8, n. 2, p. 97–109, 1993.

LAMARRE, Heather; LANDREVILLE, Kristen; BEAM, Michael. The Irony of Satire. *International Journal of Press-politics*, v. 14, p. 212–231, 2009.

LEAL, Sidney Evaldo. Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular. 2021. Tese (Doutorado) — Universidade de São Paulo. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-16072021-115303/. Acesso em: 29 set. 2024.

RUBIN, Victoria et al. Fake news or truth? using satirical cues to detect potentially misleading news. In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, California: Association for Computational Linguistics, 2016. p. 7–17.

RUBIN, Victoria L.; CHEN, Yimin; CONROY, Nadia K. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, v. 52, n. 1, p. 1–4, 2015.

SIMPSON, Paul. On the Discourse of Satire. Amsterdam: John Benjamins Publishing Company, 2003.

SINGH, Raj Kishnor. Humour, irony and satire in literature. v. 3, n. 4, p. 63–72, 2012.

TANDOC, Edson C.; LIM, Zheng Wei; LING, Richard. Defining Fake News: A typology of scholarly definitions. *Digital Journalism*, v. 6, n. 2, p. 137–153, 2018.

TAVARES, Maria Alice. O verbo no texto jornalístico: notícia e reportagem. *Working Papers em Linguística*, n. 11, p. 123-142, 1997.

WARDLE, Claire; DERAKHSHAN, Hossein. Thinking about information diso rder:formats of misinformation, disinformation, and mal-information. p. 12, 2018.

FRASEOLOGIA, LINGUÍSTICA DE CORPUS, TRADUÇÃO DE EXPRESSÕES IDIOMÁTICAS E LEXICOGRAFIA: PARCERIAS DE SUCESSO

Isabela MOREIRA DE OLIVEIRA[101]

ABSTRACT: In this study, we analyzed eighty idioms related to food in Brazilian Portuguese to better understand their meaning in context and, based on that, propose translation strategies and equivalents in English. Our goal was to present a translator-oriented, bilingual (Portuguese - English) glossary of idiomatic expressions (IEs). For each entry, we provide a definition, authentic examples, synonyms, and suggest equivalents, with authentic usage examples. We have also investigated their degree of idiomaticity and fixity and attested their use in general language corpora for both languages. What makes this material unique, though, is the fact that it can be consulted electronically by semantic fields / themes, based on the 600+ distinct words we used to describe and categorize each IE from an onomasiological standpoint. We hope this study can contribute to advancing translation of PortugueseEnglish IEs and, perhaps, inspire the creation of new methodologies and lexicographic products aimed at translators.

KEYWORDS: Portuguese-English contrastive studies; semantic fields; idiomatic expressions; Corpus Linguistics; onomasiology; translator-oriented glossary.

Toda vez que compartilho com alguém os detalhes de como foi feita a nossa pesquisa de mestrado, as pessoas se envolvem; no início achava exagerada essa reação entusiasmada, mas com o tempo percebi que a pesquisa envolvia áreas de conhecimento muito relevantes para o dia-a-dia de todos, que nossa pesquisa tinha muita aplicabilidade; é possível perceber isso a partir dos pilares que compõem nosso estudo: a **Fraseologia**, a **Linguística de Corpus**, a **tradução de expressões idiomáticas** e a **Lexicografia** - está certo que em grande parte das vezes as pessoas não sabiam dar nome àquilo sobre o que estávamos conversando, mas nem por isso se mostravam menos cativadas ao perceber o quão tangível era o assunto. Com os pilares da pesquisa firmados, estabelecemos como objetivo elaborar um glossário bilíngue português brasileiro (PB) - inglês americano (EN) de expressões idiomáticas (EIs) com a temática alimentação[102] – que é, reconhecidamente, uma área que carrega marcas culturais muito fortes (TEIXEIRA, 2003), de consulta onomasiológica e online que tivesse tradutoras/es como principais consulentes, mas que poderia também ser usado por demais aprendizes e estudiosos de inglês como L2, assim como poderia ser usado de maneira bidirecional.

Para que o objetivo pudesse ser alcançado, precisamos investigar cada um dos pilares da nossa pesquisa mais a fundo. Para chegarmos na definição

[101] Mestra em Estudos da Tradução pela Universidade de Brasília (UnB), docente temporária do Departamento de Línguas Estrangeiras e Tradução (LET) da UnB e doutoranda do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS). isabela.oliveira@unb.br

[102] O glossário bilíngue, de consulta onomasiológica e online, apresentado no APÊNDICE 1: *Glossário Português-Inglês de Expressões Idiomáticas com a Temática Alimentação*, pode ser acessado a partir do link: https://tinyurl.com/rcetsv6y

do nosso objeto de estudo (i.e. as EIs) investigamos o que compreende a Lexicologia, Terminologia, **Fraseologia** e Paremiologia. De maneira sucinta, segundo Barros (2004, p. 61), a Lexicologia estuda a palavra no nível do sistema linguístico (língua global) e a Terminologia a estuda em nível da(s) norma(s) de universos de discursos especializados (línguas de especialidade).

Saliba (2000) afirma que a Fraseologia compreende unidades lexicais (UL) formadas por duas ou mais palavras gráficas, podendo chegar à extensão de uma oração e chamadas então de unidades fraseológicas (UF) ou fraseologismos; a Paremiologia, segundo Riva (2009), pode ser vista como uma subdivisão dentro dos estudos fraseológicos e assim como a Fraseologia, tem como objeto de estudo UF. Zavaglia (2006) afirma que as UL maiores que a palavra são UF, a autora estabelece também que as UF compreendem diversos tipos de combinações estáveis e que se caracterizam por sua fixidez e idiomaticidade (ZAVAGLIA, 2017). Depois de estabelecer essas convergências e divergências, definimos nosso objeto de estudo - as expressões idiomáticas (EIs) - pela caracterização proposta por Tagnin (2005), que trabalha com UF e sua interface com a Tradução. Segundo a autora, uma expressão idiomática nada mais é do que uma UF que se caracteriza pela convencionalidade (quando uma UF torna-se consolidada pelo uso) e pela idiomaticidade (o significado da UF não pode ser deduzido através da soma de significados de seus componentes).

A essa altura, já tínhamos coletado mais de uma centena de EIs com a temática alimentação com a ajuda da internet, de amigos e familiares; dessa coletânea, selecionamos 80 EIs em PB para compor nosso banco de dados (Figura 1) e começamos a desenvolver nossa ficha de coleta em formato de planilha eletrônica para propor equivalentes em EN, que seriam então apresentados em forma de glossário bilíngue PB - EN de EIs com a temática alimentação, de consulta onomasiológica e online.

Figura 1. Excerto dos campos de coleta no banco de dados.

A1 fx Cód

	A	B	C	D	E	F	G	H	I
1	**Cód**	**EI PB**	**Signif (fonte)**	**Corpus PB**	**Freq PB**	**Ex uso PB (fonte)**	**Ano pub**	**Variante(s)**	**grau fixidez**
2	1	Algo ser mamão com açúcar	Coisa muito fácil. (https://tinyurl.com/c67a6p9e)	NOW	106	Depois de batida na Áustria, Bottas aplaude brita: "Não é para ser **mamão com açúcar**"(...)"Você freia muito tarde na curva 4, você sabe que vai para a brita. Você freia muito forte e muito rápido na 6, você sabe que está na brita, o mesmo com a 7. Eles também colocaram uma zebras sérias nas curvas 9 e 10. Isso é positivo. Você não deveria escapar e voltar assim, **mamão com açúcar**, sabe?", encerrou.	2019		médio
3	2	Algo ser batata	"É ISSO MESMO!" "É batata" é a expressão perfeita quando se quer definir que algo é certeiro e que não tem chance de errar. Ou seja: se "é batata", não precisa nem ter dúvidas! E no mundo culinário, a gente sabe bem como as batatas são versáteis, práticas e deliciosas, transformando pratos comuns em verdadeiras delícias! (https://tinyurl.com/2z7ajfb3)	NOW	3	Essa **é batata**. James Bond é o personagem mais vezes trocado dentro de uma mesma franquia (sem contar Drácula, que caiu em domínio público e aparece em qualquer filme de diferentes franquias, e Jason de Sexta-Feira 13 – que com o uso de maquiagem e sem falar, pode ser vivido por qualquer um). (https://tinyurl.com/3nheurfp)	2015		médio

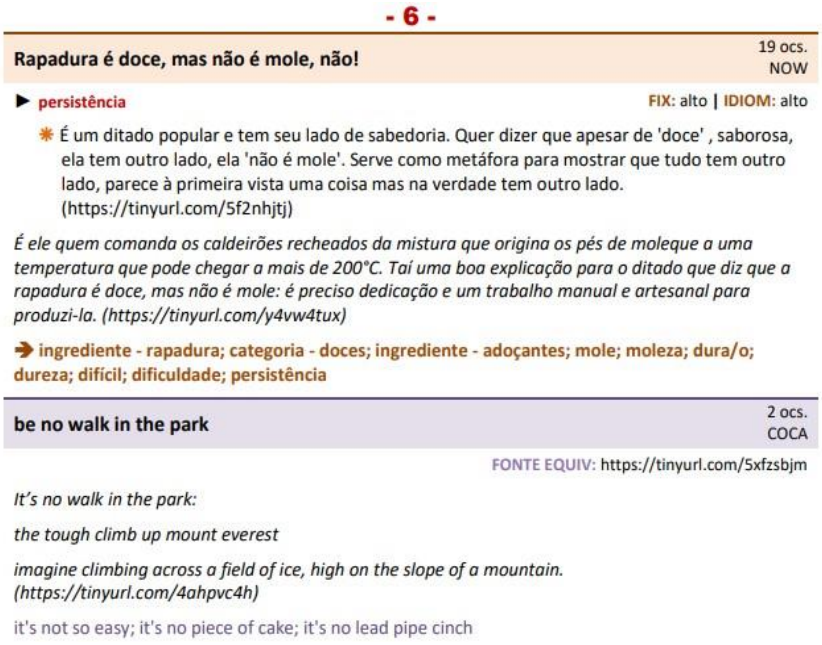
Fonte: Moreira de Oliveira, 2022, p. 58.

Nesse percurso, aplicamos as estratégias de **tradução de EIs** propostas por Baker (1992) e fomos aprimorando nossa ficha de coleta, que por fim era composta por 26 colunas por linha (i.e por EI); em relação aos equivalentes, definimos que apresentaríamos opções de EIs tanto em PB e EN atuais e

convencionais. Acredito que esse objetivo justifica fundamentalmente a nossa abordagem de desenvolver uma ficha de coleta tão extensa, tão detalhada. Para o preenchimento da ficha, tivemos que investigar mais profundamente outro pilar que compõe nossa pesquisa: a **Linguística de Corpus** (LC). Em sua interface com a Tradução, a LC mostrou-se uma grande aliada, já que utilizamos estes três corpora monolíngues como fonte de consulta para atestar o uso das EIs (PB e EN): *Corpus of Contemporary American*

English - COCA e dois subcorpora do *Corpus do Português*, o *Web/ dialects* e o *Now*, todos desenvolvidos pelo mesmo pesquisador, Mark Davies e disponíveis para acesso online e gratuito[103]. Com base na composição da ficha de coleta e no estudo mais aprofundado da **Lexicografia**, definimos a microestrutura do nosso glossário: são 80 verbetes (Figura 2) compostos pelos campos preenchidos nas fichas de coleta, dos quais destacamos grau de fixidez, grau de idiomaticidade e exemplos autênticos de uso nos dois idiomas como campos de extrema relevância para nosso consulente alvo (tradutoras/es); já em relação à sua macroestrutura, os verbetes foram organizados de maneira onomasiológica, i.e. as EIs são consultadas a partir de seus campos semânticos em direção às EIs - esses campos foram definidos no decorrer do preenchimento da ficha de coleta (que resultou em aproximadamente 600 palavras distintas[104]) e acabamos por escolher a consulta ao glossário de forma onomasiológica por priorizar relações de sinonímia por vezes perdidas pela organização semasiológica (que parte das UL em direção ao conceito) e também para que, caso o consultante se esqueça ou desconheça as UL que compõem a EI que está procurando, ainda possa encontrá-la no glossário.

Figura 2: Exemplo de verbete.



- 6 -

Rapadura é doce, mas não é mole, não! 19 ocs. NOW

▶ persistência FIX: alto | IDIOM: alto

✱ É um ditado popular e tem seu lado de sabedoria. Quer dizer que apesar de 'doce' , saborosa, ela tem outro lado, ela 'não é mole'. Serve como metáfora para mostrar que tudo tem outro lado, parece à primeira vista uma coisa mas na verdade tem outro lado. (https://tinyurl.com/5f2nhjtj)

É ele quem comanda os caldeirões recheados da mistura que origina os pés de moleque a uma temperatura que pode chegar a mais de 200°C. Taí uma boa explicação para o ditado que diz que a rapadura é doce, mas não é mole: é preciso dedicação e um trabalho manual e artesanal para produzi-la. (https://tinyurl.com/y4vw4tux)

➔ ingrediente - rapadura; categoria - doces; ingrediente - adoçantes; mole; moleza; dura/o; dureza; difícil; dificuldade; persistência

be no walk in the park 2 ocs. COCA

FONTE EQUIV: https://tinyurl.com/5xfzsbjm

It's no walk in the park:

the tough climb up mount everest

imagine climbing across a field of ice, high on the slope of a mountain. (https://tinyurl.com/4ahpvc4h)

it's not so easy; it's no piece of cake; it's no lead pipe cinch

Fonte: Moreira de Oliveira, 2022, p. 60.

[103] Disponíveis em: https://www.english-corpora.org/coca/ e https://www.corpusdoportugues.org/.
[104] O APÊNDICE 3 *Palavras usadas para caracterizar as temáticas e campos semânticos* pode ser acessado a partir do link: https://tinyurl.com/rcetsv6y

Considerando que o percurso de tradução compreendido pela ficha de coleta que desenvolvemos, aprimoramos e aplicamos na nossa pesquisa foi extenso e meticuloso, conseguimos alcançar bons resultados. O caminho foi exaustivo e demorado, pois foram 80 EIs em PB e seus equivalentes em EN organizados em verbetes e para que chegássemos a cada verbete, 26 colunas foram preenchidas para cada um deles. Outro aspecto que garantiu resultados relevantes e confiáveis foi a consulta aos corpora para atestar o uso das EIs em PB e EN. Nos deparamos com alguns desafios que nos levaram a reflexões que nos guiarão para possíveis melhorias futuramente: usar os corpora, embora extremamente útil para disponibilizar exemplos convencionados de uso, não foi muito eficiente para EIs com grau de idiomaticidade alto, ou baixo grau de fixidez; outro questão foi o limite de consultas estabelecido pelos corpora em suas versões disponíveis online gratuitas, que permitem somente 50 consultas por dia; ao pesquisar colocados que tem ambos significados idiomático e denotativo, os corpora nos apresentavam resultados de ambos e precisamos então filtrá-los; em várias ocasiões, os corpora em PB e em EN nos levaram a páginas inexistentes ou sem relação com o conteúdo alvo; tivemos dificuldade por vezes em atestar o uso de EIs por serem parte da língua falada, e não escrita como os corpora consultados; tivemos que acrescentar à nossa ficha de coleta as colunas “sinônimos” e “variantes” já que nos deparamos diversas vezes com mais de uma opção de equivalente.

A abordagem contrastiva nos permitiu chegar a algumas conclusões que caracterizam as EIs que compõem o glossário: 59% das EIs tem grau de fixidez médio, 37% tem grau alto de fixidez, e 3% grau baixo - EIs com baixo e médio grau de fixidez se misturam mais facilmente no texto, isso pode indicar a dificuldade que tradutores e falantes “ingênuos” (TAGNIN, 2005) tem ao identificá-las em contexto; EIs com grau de fixidez alto apresentam menos variantes quando comparadas àquelas de grau baixo e médio de fixidez, o que também se aplica a flexão de gênero e número; quanto ao grau de idiomaticidade, 49% das EIs tem nível de idiomaticidade alto, 37% médio e somente 13% baixo, esses números podem justificar nossas escolhas por equivalentes ligados ao significado conotativo e não ao denotativo. Ao observarmos as palavras que compõem os campos semânticos, observamos que o campo semântico *ingredientes* aparece em aproximadamente 44% das entradas, o campo semântico dos *pratos,* está presente em cerca de 30% das entradas; os ingredientes que mais apareceram foram *frutas* (10%), *carboidratos* (8%), *tubérculos* (6%) e *carnes* (6%), podemos então perceber que são alimentos que compõem a base da alimentação dos brasileiros; dentre as frutas, são as *frutas tropicais* (10%) que aparecem mais; quanto aos *animais*, que aparecem em 8% das entradas, as *aves* ocorrem em 5% e *porcos* 4%, o que indica que são animais comuns na cultura brasileira, que podem ser criados em nossos quintais; *parte do corpo*, presente em cerca de 19% das entradas nos surpreendeu devida a sua grande ocorrência (não esperávamos uma porcentagem tão alta mesmo que *estômago* e *boca* sejam indicativos da temática alimentação; dentre as palavras que não têm relação com a temática alimentação, *problema* ocorre em 10% das EIs, *superação* em 9% e *dificuldade* ocorre em 6%, o que nos leva ao caráter de ensinamentos ancestrais expressados pelas EIs.

Ao empregar as estratégias propostas por Baker (1992), apresentamos EIs equivalentes em significado, mas não na forma:

54) *ALGUÉM DAR com a língua nos dentes* | *spill the beans*

24) *ALGO ACABAR em pizza* | *come to nought*

52) MANDAR ALGUÉM catar coquinho(s) | *go take a long walk on a short pier* Algumas EIs que tem equivalentes muito similares em relação a sua forma e conteúdo em ambas as línguas:

10)ALGO SER a *cereja* do *bolo* | the *cherry* on the *cake*

11)(não adianta) *chorar* (sobre) o *leite derramado* | *cry* over *spilled milk*

49) Se/Quando a *vida* DAR A ALGUÉM um *limão/limões*, faça *limonada* | when *life* gives you *lemons*, make *lemonade*

Diante disso, concluímos que é possível encontrar EIs que são compartilhadas pelo mundo por conterem ensinamentos ou perspectivas universais sobre determinado assunto e por isso foram adotadas por outras culturas. Por fim, a parceria entre LC e tradução de fraseologismos foi feliz - aliás, construir o glossário nos levou a entender melhor a LC e sua interface com a tradução. Compartilhamos a metodologia aplicada na direção PB > EN[105], que pode ser também usada na direção inversa – esperamos que estudos na área continuem a se expandir.

Agradecimentos: À organização do ELC/ EBRALC 2024, por esse evento tão necessário.

Referências

BAKER, M. *In other words.* Abingdon, Oxon ; New York, NY: Routledge, 1992.

BARROS, Lídia A. *Curso básico de terminologia*. São Paulo: EDUSP, 2004.

RIVA, H. C. Dicionário onomasiológico de expressões idiomáticas usuais na língua portuguesa no Brasil. São José do Rio Preto: 2009, 311 f. *Tese*

(doutorado em Estudos Linguísticos) – Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista.

SALIBA, M. C. Unidades lexicais maiores que a palavra: descrição linguística, considerações psicolinguísticas e implicações pedagógicas. *Dissertação* (mestrado) - Universidade Federal do Paraná. Curitiba: 23/08/2000. Disponível em: https://acervodigital.ufpr.br/handle/1884/24439. Acesso em: 07 ago. 2020.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. São Paulo: Disal, 2005. 223p.

[105] A metodologia completa desta pesquisa está disponível em: https://tinyurl.com/rcetsv6y.

TEIXEIRA, E.D. Em busca de um novo modelo tecno-formal para a construção de dicionários técnicos bilíngues - o exemplo da culinária. *Intercâmbio* (PUCSP), São Paulo, SP, v. XII, p. 243-251, 2003.

ZAVAGLIA, C. Dicionário e Cores. *Alfa*, São Paulo, 50 (2): 25-41, 2006.

ZAVAGLIA, C.; FROMM, G. Fraseologia e Paremiologia: uma entrevista com Claudia Zavaglia. *ReVEL*, v. 15, n. 29, 2017. Disponível em: https:// revel.inf.br. Acesso em: 15 set. 2020.

INPACT - INTERNACIONALIZAÇÃO DA PRODUÇÃO ACADÊMICA COM CORPUS E TECNOLOGIA: A CONSTRUÇÃO DE UMA FERRAMENTA ON-LINE PARA A ESCRITA DE ARTIGOS DE PESQUISA EM INGLÊS NAS HUMANIDADES

Ana Eliza Pereira BOCORNY[106]
Deise Prina DUTRA[107]

RESUMO: O projeto InPACT visa internacionalizar a produção científica brasileira nas Humanidades, desenvolvendo uma ferramenta online de suporte à escrita acadêmica em inglês baseada em linguística de corpus. A ferramenta utiliza elementos fraseológicos extraídos de corpora para auxiliar pesquisadores a atenderem às convenções linguísticas e retóricas das suas disciplinas. Testes preliminares indicam que a ferramenta é eficaz e intuitiva.

Palavras-chave: Linguística de Corpus; Gêneros Acadêmicos; *Lexical Bundles*; *Lexical Frames*.

INTRODUÇÃO

O projeto InPACT tem como objetivo principal contribuir para a internacionalização da produção científica brasileira nas áreas de Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes. Para tanto, o projeto propõe o desenvolvimento de recursos pedagógicos on-line baseados em corpus, que possam auxiliar pesquisadores e estudantes na produção de artigos científicos em inglês. Uma ferramenta on-line de suporte à escrita acadêmica é um dos recursos que está sendo desenvolvido no âmbito do projeto. É importante destacar que diversos trabalhos têm explorado o desenvolvimento de ferramentas para apoiar a produção textual acadêmica, tanto em língua portuguesa quanto em outras línguas.

No Brasil, recursos como o AMADEUS (ALUÍSIO *et al*., 2001), o SciPo (FELTRIN, *et al*., 2003) e o SciPo - Farmácia (SCHUSTER *et al*., 2005) surgiram como esforços pioneiros nessa direção. Internacionalmente, temos exemplos como o AWSUM (MIZUMOTO, et al., 2017) e o Collocaid (FRANKENBERG-GARCIA *et al*., 2019). Neste contexto, a ferramenta que está sendo desenvolvida no âmbito do projeto InPACT emerge como uma resposta às necessidades específicas da produção científica de artigos de pesquisa em inglês por parte da comunidade acadêmica brasileira das áreas-alvo, considerando suas particularidades linguísticas e os padrões discursivos das seções mencionadas.

A presente proposta se alinha a iniciativas passadas e busca avançar na integração da Linguística de Corpus e da tecnologia em prol da escrita acadêmica em inglês, especialmente por buscar elementos fraseológicos recorrentes (*lexical bundles* e *lexical frames*) nos corpora compilados e relacioná-los às funções retóricas que os mesmos realizam nas diferentes

[106] Professora do Magistério Superior, Porto Alegre - RS, Universidade Federal do Rio Grande do Sul (UFRGS). E-mail: ana.bocorny@gmail.com

[107] Professora do Magistério Superior, Belo Horizonte - MG, Universidade Federal de Minas Gerais (UFMG).

seções das 16 disciplinas-alvo[108], no âmbito das Humanidades.

Neste contexto, o presente projeto constituiu-se a partir de pressupostos teóricos de diferentes áreas do conhecimento. Aqui, damos destaque aos (i) estudos sobre gêneros do discurso e (ii) aos princípios da Linguística de Corpus. A concepção de gênero que fundamenta esta pesquisa está alinhada com as perspectivas de Bakhtin (1997), Swales (1990) e Bathia (1997). De Bakhtin (1997), ressalta-se a concepção de gênero do discurso como "tipos relativamente estáveis de enunciados". De Swales (1990, p. 46), destaca-se a ideia de que os gêneros "são veículos de comunicação para atingir um objetivo". Por fim, de Bathia (1997, p.160), sublinha-se o entendimento de análise de gênero como sendo "o estudo do comportamento linguístico situado em contextos acadêmicos ou profissionais".

A Linguística de Corpus parte de uma perspectiva de descrição da língua em uso, seja ela geral ou especializada. A visão da língua como um sistema probabilístico é um dos fundamentos principais da Linguística de Corpus (BERBER SARDINHA, 2004). Assim, os traços linguísticos (lexicais, estruturais, pragmáticos e discursivos) não ocorrem todos com a mesma regularidade (BERBER SARDINHA, 2004). Por esse motivo, a variação dos traços não é aleatória; pelo contrário, existe "um mapeamento regular entre a frequência maior ou menor de um traço e um contexto de ocorrência" (BERBER SARDINHA, 2004, p. 351). Dessa forma, defender que os traços não são aleatórios significa dizer que "a linguagem é padronizada. A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura que se repete significativamente mostra sinais de ser, na verdade, um padrão lexical ou léxico-gramatical." (BERBER SARDINHA, 2004, p. 31).

Por fim, é importante ressaltar que, ao propor a identifição das formas linguísticas que realizam as funções retóricas expressas nas seções de artigos de pesquisa em inglês, este projeto busca preencher um "gap" relatado por Moreno e Swales (2018) e Gray *et al*. (2020). Os referidos autores afirmam que, ainda hoje, poucos estudos são realizados a partir da combinação da perspectiva e princípios dos estudos sobre gêneros do discurso e da Linguística de Corpus para investigar as realizações linguísticas de movimentos retóricos de diferentes gêneros do discurso.

METODOLOGIA

Este estudo conta com uma equipe multidisciplinar que inclui linguistas de corpus, especialistas em *design* e cientistas da informação. A partir desses saberes o desenvolvimento da ferramenta valeu-se de uma combinação de etapas metodológicas que envolveram as três áreas.

A primeira etapa tratou da definição dos objetivos da ferramenta, da identificação dos usuários alvo e de suas necessidades. Como parte dessa etapa

---

[108] **Ciências Humanas (10):** Filosofia (Phil), Sociologia (Soc), Antropologia (Ant), Arqueologia (Arc), Geografia (Geo), Psicologia (Psy), Educação (Edu), Religiões (RelF), Demografia (Dem), **Ciências Sociais Aplicadas (4):** Direito e Ciências Jurídicas (LawLS), Economia (Eco), Comunicação (Com), Ciência Política (PolS), Políticas Públicas (PubP), **Linguística, Letras e Artes (2):** Linguística (Ling), Letras (Lang).

inicial foi criado o *naming* (processo de criação e desenvolvimento de nomes para marcas, produtos, serviços, empresas e outras entidades) e a identidade visual do projeto.

A segunda etapa foi a seleção e organização dos textos utilizados para extração de dados linguísticos usados para informar a construção da ferramenta. Nesta etapa também buscou-se a identificação da estrutura retórica das diferentes seções dos artigos científicos das áreas de Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes.

A terceira etapa foi a definição da arquitetura da ferramenta e a elaboração do design da interface. Nesta etapa, foram realizados testes de usabilidade.

A quarta etapa foi o desenvolvimento de um protótipo da ferramenta e a inserção dos dados linguísticos extraídos na etapa dois. Foram utilizadas tecnologias de programação *web* e bancos de dados linguísticos para criar uma plataforma on-line que fosse intuitiva, que pudesse ser acessada com facilidade e que oferecesse os elementos linguísticos necessários para a produção de artigos de pesquisa em inglês.

RESULTADOS E DISCUSSÃO

Como resultados da primeira etapa da construção da ferramenta tivemos a identificação de três personas que representam os usuários-alvo e suas necessidades. Após o desenvolvimento de estudo, alternativas, seleção e refinamento, o *naming* foi definido como: InPACT - Internacionalização da Produção Acadêmica com Corpus e

Tecnologia. Com o *naming* escolhido, a identidade visual foi elaborada. A Figura 1 mostra o logo definido, a inspiração para o ícone criado a partir do logo, as três aplicações do logo para resultados do projeto e as escolhas de ícones e grafismos para cada disciplina.

Os dados linguísticos extraídos na segunda etapa do estudo foram obtidos a partir do Corpus de Humanidades (CORHUM), um corpus estratificado em 16 disciplinas da área das Humanidades e em quatro seções de artigos de pesquisa (Introdução, Metodologia e Resultados / Discussão e Conclusão - IMR/DC). Com um total de aproximadamente 64 milhões de palavras, o CORHUM conta com 64 subcorpora com aproximadamente 1 milhão de palavras cada. Os subcorpora contêm textos das seções IMR/DC de artigos de pesquisa das disciplinas-alvo Esses artigos foram publicados na plataforma PLOS One, em inglês, de 2013 a 2023. Utilizamos as ferramentas Sketch Engine (KILGARRIFF *et al*., 2004) e AntConc 4.0.10 (ANTHONY, 2022) para extrair conjuntos de *lexical bundles (LBs)* e *lexical frames (LFs)* de cada subcorpus (por exemplo, o conjunto de *LBs* da seção metodologia da disciplina Educação).

Uma vez extraídos, *LBs* e *LFs* foram agrupados por similaridade lexical (unidades com um número de elementos lexicais iguais: *the aim of this paper is to* e *the objective of this paper is to*) e por similaridade retórica (unidades diferentes com funções retóricas iguais: *the aim of this paper is to* e *this paper aims to*). Uma vez agrupados, partiu-se para a análise das funções retóricas de cada conjunto de unidades fraseológicas. A relação entre forma e função se deu a partir de um *framework* representando a estrutura retórica de artigos das

disciplinas-alvo, construído a partir da revisão de 25 estudos prévios (por exemplo, ZHANG; WANNARUK, 2016; YANG; ALLISON, 2003).

Por fim, construiu-se uma base de dados com as informações coletadas que foi incorporada ao protótipo da ferramenta. A metáfora usada para a construção da ferramenta foi a do texto acadêmico como uma parede com tijolos vermelhos e azuis. Os tijolos vermelhos representando os *building blocks of discourse,* formulaicos e convencionais, sob a forma de *LBs* e *LFs* (por exemplo: '*The aim of this paper is to…')* e os tijolos azuis representando o 'conteúdo' relativo ao estudo que o pesquisador está desenvolvendo (por exemplo, '...*extract the most frequent lexical frames from a corpus of abstracts*').

A ferramenta pretende oferecer as opções correspondentes aos tijolos vermelhos relacionando tais opções aos movimentos retóricos das seções onde ocorrem. O texto referente aos tijolos azuis, diz respeito ao conhecimento prévio e aos dados da pesquisa de cada autor de cada artigo de pesquisa. A arquitetura da ferramenta e a elaboração do *design* da interface foram os resultados obtidos na terceira etapa do estudo. A partir do entendimento de que a ferramenta deveria ser capaz auxiliar na produção de um artigo de pesquisa com a estrutura retórica convencionalmente usada nas disciplinas-alvo, utilizando os *LBs* e *LFs* identificados em cada seção, os principais casos de uso da ferramenta foram identificados como: (i) redigir um texto/trecho; (ii) visualizar exemplos; (iii) consultar propósitos da ferramenta; (iv) ver tutoriais.

A partir das constatações descritas foi desenhado o fluxo do usuário simulando o caso de uso 'redigir um texto'. Definido o fluxo do usuário, iniciou-se a geração de alternativas para a ferramenta e o protótipo de uma das opções foi criado. A partir de testes realizados, foram feitas melhorias e aprimoramentos no protótipo inicial da ferramenta.

A ferramenta on-line de suporte à escrita acadêmica em inglês já está em fase de testes. O feedback recebido de pesquisadores das disciplinas-alvo têm sido positivo, indicando que a ferramenta é intuitiva, fácil de usar e eficiente no oferecimento de padrões linguísticos convencionais e específicos das diferentes disciplinas e seções dos artigos de pesquisa da área-alvo.

REFERÊNCIAS

ALUÍSIO, S. M.; BARCELOS, I.; SAMPAIO, J.; OLIVEIRA JR, O. N. How to Learn the Many Unwritten "Rules of the Game" of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases to Support Scientific Writing. **IEEE International Conference on Advanced Learning Technologies**, p. 257-260, 2001.

ANTHONY, L. **AntConc** (Version *4.0.10*) [Computer Software]. Tokyo, Japan: Waseda University, 2021. Disponível em: https://www.laurenceanthony.net/software. Acesso em: 20 jul. 2022.

BAKHTIN, M. **Estética da Criação Verbal**. 2.ed. São Paulo: Martins Fontes, 1997.

BHATIA, V. K. 'Análise de gênero hoje' [Trad. Benedito G. Bezerra]. **Revue Belge de Philologie et d´Historie**, Bruxelles, v. 75, p. 629-652, 1997.

BERBER SARDINHA, T. **Linguística de Corpus**. São Paulo: Manole, 2004.

FELTRIM, V. D. *et al.* A construção de uma ferramenta de auxílio à escrita de resumos acadêmicos em português. In: **Anais do XXIII Congresso da Sociedade Brasileira de Computação**, 2003.

FRANKENBERG-GARCIA, A., REES, G., LEW, R., ROBERTS, J., SHARMA, N. AND BUTCHER, P. ColloCaid: a tool to help academic English writers find the words they need. In: MEUNIER, F.; VAN DE VYVER, J.; BRADLEY, L.; THOUËSNY, S. (Orgs.). **CALL and complexity – short papers from EUROCALL 2019**. Voillans:Research-publishing.net, 2019.

KILGARRIFF, A. *et al*. Itri-04-08 the sketch engine. **Information Technology**, v. 105, n. 116, p. 105-116, 2004.

MIZUMOTO, A; HAMATANI, S; IMAO, Y. Applying the bundle–move connection approach to the development of an online writing support tool for research articles. **Language Learning**, v. 67, n. 4, p. 885-921, 2017.

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.

SCHUSTER, E.; ALUÍSIO, S. M.; FELTRIM, V. D.; PESSOA JR, A.; OLIVEIRA JR, O.N.Enhancing the Writing of Scientific Abstracts: A Two-phased Process Using Software Tools and Human Evaluation. **XXV Congresso da Sociedade Brasileira de Computação**, p. 962-971, 2005.

SWALES, J. **Genre analysis**: English in academic and research settings. Cambridge: Cambridge University Press, 1990.

YANG, R.; ALLISON, D. Research articles in applied linguistics: Moving from results to conclusions. **English for Specific Purposes**, v. 22, p. 365-385, 2003.

ZHANG, B.; WANNARUK, A. Rhetorical Structure of Education Research Article Methods Sections. **PASAA**: Journal of Language Teaching and Learning in Thailand, v.51, p. 155-184, 2016.

ANÁLISE MULTIDIMENSIONAL ADITIVA DOS TEXTOS LEGAIS FEDERAIS BRASILEIROS

Carolina Godoi de Faria MARQUES[109]
Carlos Henrique KAUFFMANN [110]

RESUMO
Neste artigo apresentamos a Análise Multidimensional Aditiva (BIBER, 1988) do LEX-BR-Ius (FERRARI e MARQUES, 2022), um corpus de textos legais brasileiros. Para a sua realização, nosso corpus foi adicionado às dimensões de variação do português brasileiro identificadas por Berber Sardinha, Kauffmann e Acunzo (2014) e comparado com o Corpus Brasileiro de Variação e Registro (CBVR). Os resultados indicam que os textos legais são um registro informacional, letrado e orientado para o futuro.

Palavras chave: Análise multidimensional; LEX-BR-Ius; Legislação federal brasileira; Variação linguística; Linguagem jurídica.

INTRODUÇÃO

As leis, como são popularmente chamados os textos legais, são os pilares da nossa sociedade. Elas estabelecem as normas sob as quais ela se pauta, organizando-a, regulando-a e protegendo-a (SOUZA e SOUZA, 2017). Apesar da sua importância, seu texto é percebido pelo cidadão comum como rebuscado e de difícil compreensão, sendo muitas vezes necessário o auxílio de um profissional do direito para interpretá-lo, o que dificulta o conhecimento e o exercício dos seus direitos e deveres. Essa dificuldade se dá porque a linguagem utilizada nas leis - a linguagem jurídica - é altamente especializada e complexa (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018).

O estudo da linguagem jurídica é um campo de pesquisa em expansão, entretanto são poucos os estudos sobre a linguagem utilizada nas leis e até o momento da realização dessa pesquisa não identificamos corpora compilados para o estudo de textos legais brasileiros nem estudos sobre a sua variação nessa língua (TIERSMA, 1999; PONTRANDOLFO, 2012). Diante dessa lacuna, optamos por investigar a variação linguística na legislação brasileira vigente. Para tanto, partimos da hipótese de que os textos legais são um registro, conforme definição de Biber e Conrad (2009) e realizamos uma Análise Multidimensional (AMD) Aditiva (BIBER, 1988) do LEX-BR-Ius (FERRARI e MARQUES, 2022), um corpus da legislação federal brasileira por nós compilado, adicionando-o às dimensões de variação do português brasileiro (PB) identificadas por Berber Sardinha, Kauffmann e Acunzo (2014).

A ANÁLISE MULTIDIMENSIONAL

[109] Doutoranda, Universidade Federal de Minas Gerais, Belo Horizonte/MG. Bolsista CAPES (n. 939578/2024-00). E-mail: carol.godoi@outlook.com.br

[110] Doutor, pesquisador, Pontifícia Universidade Católica de São Paulo, São Paulo/SP. Bolsista de pós-doutoramento CAPES.

A Análise Multidimensional (BIBER, 1988) é uma abordagem empíricometodológica baseada em corpus para o estudo da variação linguística. Nela adotase a noção de registro, ou seja, uma variedade da língua com traços situacionais, linguísticos e funcionais próprios, utilizada em contextos comunicativos específicos (BIBER e CONRAD, 2009). Biber (1988) propõe que a existência de padrões de coocorrência de traços linguísticos em determinado registro é motivada funcionalmente cuja identificação e a subsequente comparação possibilitaria sua caracterização e descrição (BERBER SARDINHA, 2010). Para tanto, são estabelecidas dimensões de variação a partir da análise estatística de um corpus e da interpretação funcional de seus resultados (Biber 1988).

Segundo Berber Sardinha (2013a) é possível realizar seja a Análise Multidimensional completa, conforme proposta por Biber (1988), quanto a Análise Multidimensional aditiva. As diferenças entre elas se resumem na complexidade dos cálculos estatísticos necessários para a sua realização e no grau de detalhamento da descrição dos registros (BERBER SARDINHA, 2013a; et al, 2019). Ademais, enquanto a AMD completa identifica as dimensões de variação, a aditiva não permite tal identificação, se valendo das dimensões identificadas por uma AMD completa para a sua realização. Considerada por Berber Sardinha et al (2019) como mais simples e flexível, a AMD aditiva fornece um panorama dos registros em análise, obtido a partir da adição do corpus de estudo às dimensões da AMD completa que as identificou e sua comparação com o corpus utilizado para identificá-las.

METODOLOGIA

O LEX-BR-Ius

Para a realização do presente estudo compilamos o LEX-BR-Ius um corpus sincrônico composto por textos legais federais brasileiros em vigência no momento da compilação. Visando garantir uma correta representatividade e não enviesar seu conteúdo e organização interna, optou-se por coletar textos de todos os tipos legais em sua integralidade (SINCLAIR, 2004; BIBER, 1993) no Portal da Legislação, site governamental que disponibiliza as leis atualizadas gratuitamente online, e o balanceamento foi feito segundo a frequência de uso dos textos. Para propiciar buscas e análises linguísticas aprofundadas o corpus foi etiquetado morfossintaticamente com o PALAVRAS (BICK, 2000, 2014) e marcado em Modest XML (HARDIE, 2014) com etiquetas criadas por nós. Quanto à arquitetura do corpus, optamos por manter a divisão do Portal da Legislação separando os textos legais em seis seções: Constituição, Códigos, Estatutos, Emendas à Constituição, Leis complementares e Leis ordinárias., perfazendo um total de 755 normas e 3.300.289 palavras.

A AMD

Para alcançar nossos objetivos, realizamos a AMD Aditiva do nosso corpus. Para tanto, adicionamos o LEX-BR-Ius às dimensões do português brasileiro (PB) identificadas pela AMD do Corpus Brasileiro de Variação e Registro (CBVR) (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014): *(1) Oral versus literate discourse, (2)*

Argumentation, (3) Involved versus informational production, (4) Directive discourse, (5) Future versus past time orientation e *(6) Reported discourse*. Trata-se do estudo mais completo sobre a variação do PB já feito, em que foram analisados 48 registros escritos e orais dessa língua.

O primeiro passo dessa análise foi anotar e contabilizar as ocorrências dos traços linguísticos de cada uma das dimensões analisadas no nosso corpus com o etiquetador PALAVRAS (BICK, 2000 e 2014) e o pós-processador PALAVRAS Tag count (BERBER SARDINHA, 2013b), ambos adotados por Berber Sardinha, Kauffmann e Acunzo (2014) em seu estudo. Em seguida, normalizamos as ocorrências por mil palavras e calculamos seus Z-escores para que as frequências absolutas dos traços linguísticos em analise não enviesassem os dados.

O próximo passo foi calcular a carga fatorial dos textos e a partir dela a carga fatorial do corpus, o que nos permitiu localizá-lo nas dimensões do PB e compará-lo com os registros do CBVR. Para tanto, primeiramente calculamos o escore de dimensão dos textos e, em seguida, a média desses escores. Esta última foi então incorporada à tabela com as médias de dimensão dos registros do CBVR, a nós disponibilizada pelos autores do estudo. Por fim, para constatar a significância estatística dos nossos resultados, realizamos os testes ANOVA e R^2.

RESULTADOS E DISCUSSÃO

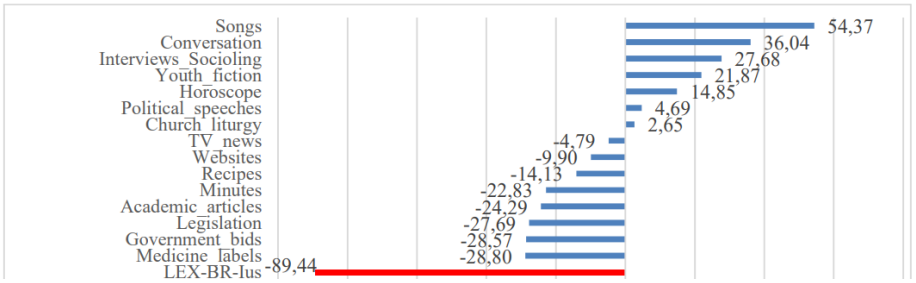
O LEX-BR-Ius pontuou e se distinguiu dos registros do CBVR significativamente em todas as dimensões. Nessa seção, por questões de espaço, apresentaremos apenas as dimensões nas quais nosso corpus se destacou (1, 3 e 5), e reproduziremos nos gráficos apenas parte dos registros abarcados pelo CBVR. Nas demais dimensões (2, 4 e 6) as pontuações obtidas giram em torno de 0. Provavelmente, como os textos legais são impositivos e visam majoritariamente informar e descrever as normas da forma mais clara e detalhada possível, o uso de traços argumentativos, diretivos e do discurso indireto é relegado.

Dimensão 1: Oral vs. literate discourse

Na dimensão 1 os registros são distribuídos segundo o seu grau de oralidade e de letramento conforme o gráfico abaixo. No polo positivo temos os

registros nos quais predominam os traços linguísticos[111] do discurso oral e no negativo aqueles do discurso letrado.

Gráfico 1: Dimensão 1



Fonte: autoras (2024)

O LEX-BR-Ius obteve a maior pontuação negativa dessa dimensão, logo, dentre os registros analisados, é o que carrega mais traços do discurso letrado. Dentre eles destacamos: orações reduzidas de gerúndio (i), passivas sem agente (ii), nominalizações em posição de sujeito (ii), particípios passados e substantivos compostos e abstratos (i, ii), como mostram os exemplos abaixo

(i) I - praticar ato visando fim proibido em lei ou regulamento ou diverso daquele previsto, na regra de competência; (LO8.429_02.06.1992)

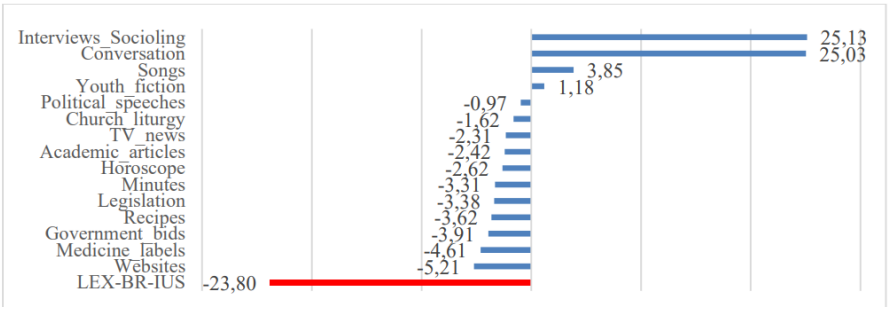
(ii) Art. 15. A participação no CONARE será considerada serviço relevante e não implicará remuneração de qualquer natureza ou espécie. (E9.474_22.12.1997) Os traços linguísticos dessa dimensão contribuem para a complexidade gramatical e densidade informacional dos textos e exercem a função de restringir e detalhar seu conteúdo, compactando um grande volume de informações fornecidas de forma técnica e concisa.

Dimensão 3: Involved vs. informal production

Nessa dimensão temos, no polo positivo, os registros marcados pela interação e, no polo negativo, os registros informacionais, sendo o único traço desse polo o *typetoken ratio*, que mede a densidade lexical dos textos.

[111] Identificados pela AMD realizada por Berber Sardinha, Kauffmann e Acunzo (2014).

Gráfico 2: Dimensão 3



Fonte: autoras (2024)

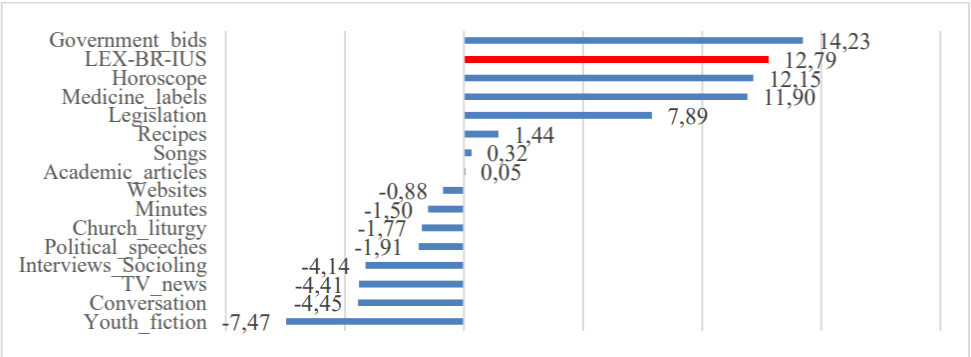
O LEX-BR-Ius obteve a maior pontuação negativa, indicada pela diversidade de vocabulário utilizado nos textos legais, que envolve termos técnicos, termos em latim, e também substantivos em geral que especificam ao máximo o sujeito do texto.

(i) Art. 2º São águas públicas de uso comum: a) os mares territoriais, nos mesmos incluídos os golfos, bahias, enseadas e portos; e) as nascentes quando forem de tal modo consideráveis que, por si só, constituam o "caput fluminis";
(C24.643_10.07.1934)

Dimensão 5: Future vs. past oriented

A dimensão 5, por sua vez, abarca a variação na orientação temporal predominante no discurso que vai do passado (polo negativo) ao futuro (polo positivo).

Gráfico 3: Dimensão 5



Fonte: autoras (2024)

Nessa dimensão nosso corpus apresentou uma das maiores pontuações positivas, sendo caracterizado por um discurso orientado para o futuro. Esse é

marcado por verbos no futuro do subjuntivo e do presente (i), modais haver (ii) e poder (ii), orações subordinadas (i), conjunções coordenadas e advérbios de probabilidade.

(i) Art. 39. O pedido passivo de cooperação jurídica internacional será recusado se configurar manifesta ofensa à ordem pública. (C13.105_16.03.2015)

(ii) Art. 10. O juiz não pode decidir, em grau algum de jurisdição, com base em fundamento a respeito do qual não se tenha dado às partes oportunidade de se manifestar, ainda que se trate de matéria sobre a qual deva decidir de ofício. (C13.105_16.03.2015)

Esses traços exercem a função de descrever e especificar como se dará a aplicação das normas e quais serão suas consequências a partir da sua promulgação.

CONCLUSÃO

A AMD aditiva nos permitiu traçar um perfil linguístico dos textos legais federais brasileiros a partir do seu mapeamento em todas as dimensões de variação do PB identificadas por Berber Sardinha, Kauffmann e Acunzo (2014). Nossos resultados indicam que os textos legais são caracterizados por um discurso letrado, de caráter informacional e orientado para o futuro, sendo que a argumentação e a diretividade exercem um papel secundário nos textos e o discurso direto é favorecido àquele indireto. Ademais, nosso corpus computou cargas fatoriais únicas e variação estatisticamente significativa em todas as dimensões analisadas, confirmando nossa hipótese de que os textos legais seriam um registro.

Agradecimentos

Essa pesquisa foi parcialmente financiada pela CAPES (bolsa nº 88887.626989/202100) e FAPEMIG (bolsas PROBIC e PIC-JR-FAPEMIG) as quais agradecemos.

Referências

BERBER SARDINHA, T. A abordagem metodológica da Análise Multidimensional. *Gragoatá*. Niterói, n. 29, p. 107-125, 2. sem. 2010. Disponível em: https://periodicos.uff.br/gragoata/article/view/33077. Acesso em: 19 jan. 2022.

BERBER SARDINHA, T. Variação entre registros da Internet. In: SHEPHERD, T. G.; SALIÉS, T. G. (Eds.). *Linguística da Internet*. São Paulo: Contexto, 2013a, p. 55–85.

BERBER SARDINHA, T. *Pós-processador PT Tag Count*. 2013b.

BERBER SARDINHA, T.; PINTO, M. V.; MAYER, C.; ZUPPARDI, M. C.;

KAUFFMANN, C. H. Adding Registers to a Previous Multi-Dimensional Analysis. In:

BERBER SARDINHA, T.; VEIRANO PINTO, M. (eds.). *Multi-Dimensional Analysis*: Research Methods and Current Issues. London: Bloomsbury, 2019. p. 165-186.

BERBER SARDINHA, T.; KAUFFMANN, C.; ACUNZO, C. M. Dimensions of register variation in Brazilian Portuguese. In: VEIRANO PINTO, M. (Eds.). *Multi-dimensional analysis: 25 years on a tribute to Douglas Biber*. John Benjamins Publishing Company, 2014.

BIBER, D. Representativeness in Corpus Design. *Literary and Linguistic Computing,* v. 8, n. 4, Oxford: Oxford University Press, p. 243-257, 1993.

BIBER, D. *Variations across speech and writing***.** Cambridge: CUP, 1988.

BIBER, D; CONRAD, S**.** *Register, genre, and style*. Cambridge: CUP, 2009.

BICK, E. PALAVRAS, a constraint grammar-based parsing system for Portuguese. In: T. SARDINHA, Berber, e FERREIRA, T. São Bento (Eds.), *Working with Portuguese corpora*. London: Bloomsbury, p 279–302, 2014.

BICK, E. *The Parsing System "Palavras":* Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dr. Phil. thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press, 2000.

CARAPINHA, C. A linguagem jurídica. Contributos para uma caracterização dos Códigos Legais***.*** *REDIS: Revista de Estudos do Discurso*, n. 7, 2018. Disponível em:

https://ojs.letras.up.pt/index.php/re/article/view/6200. Acesso em: 10 set. 2021.

GOŹDŹ-ROSZKOWSKI, Stanisław. Legal Language. In: CHAPELLE, Carol A. (Org.*). The Encyclopedia of Applied Linguistics.* John Wiley e Sons, 2012, p. 3281-3287.

HARDIE, A. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal*, v. 38, n. 1, 73-103, 2014. Disponível em: https://doi.org/10.2478/icame2014-0004. Acesso em: 20 jul. 2021.

FERRARI, L. A.; MARQUES, C. G. F. O LEX-BR-Ius: arquitetura e decisões na compilação de um corpus representativo das leis federais brasileiras. *ANTARES*, v.14, n.34, 2022. Available at: http://www.ucs.br/etc/revistas/index.php/antares/article/view/11150/5328. Access:19 dez. 2022.

PONTRANDOLFO, Gianluca. Legal Corpora: an overview. *Rivista Internazionale di Tecnica della Traduzione*, Trieste, v. 14, p. 121-136, 2012. Disponível em: https://www.openstarts.units.it/bitstream/10077/9783/1/12Pontrandolfo.pdf. Acesso em: 6 set. 2020.

SINCLAIR, John. Corpus and Text. In: WYNNE, M (eds.). *Developing Linguistic Corpora:* a Guide to Good Practice. Oxford: Oxbow Books, 2005, p. 1-16. Disponível em: 6 set. 2020.

SOUZA, C. F. de; SOUZA, P. de T. F. de. Direito e democracia - o significado das leis e do legislativo na teoria da democracia. *Revista Do Direito*, (51), 145-156, 2017. Disponível em: https://doi.org/10.17058/rdunisc.v1i51.7784. Acesso em: 6 set. 2020.

TIERSMA, P. *Legal Language*. Chicago: The University of Chicago Press, 1999.

A CRIAÇÃO DO MACHADO DE ASSIS CATÁLOGO & CORPUS (MACC)

Ursula Puello SYDIO[112]

RESUMO O MACC, ou Machado de Assis Catálogo & Corpus, é um recurso digital que permite que pesquisadores acessem um abrangente catálogo e um corpus bilíngue da obra machadiana traduzida para o inglês. Enquanto o seu Catálogo reúne 39 títulos, incluindo romances, contos e antologias do autor, o seu Corpus conta com 6 romances (11 traduções) e 90 contos (373 traduções).

Palavras-chave Linguística de corpus; Machado de Assis; Estudos da tradução; catálogo; Literatura brasileira traduzida.

INTRODUÇÃO

Machado de Assis é o autor brasileiro mais pesquisado no exterior de acordo com Costa (2016). No entanto, ainda não existia de um catálogo completo e atualizado das traduções para o inglês, tampouco um corpus eletrônico que facilitasse a consulta a essas traduções. Com essa lacuna em mente, foi desenvolvido o website MACC (SYDIO, 2023a), o Machado de Assis Catálogo & Corpus, que está disponível no endereço https://macc.fflch.usp.br/pt-br/.

O Catálogo do MACC lista 39 publicações traduzidas no período de 1921 a 2023, incluindo romances, contos e antologias da obra machadiana. Já o Corpus traz 6 romances (11 traduções) e 90 contos (373 traduções).

Este trabalho tem como objetivo apresentar os métodos utilizados para a criação do catálogo e do corpus, além de explicar como os pesquisadores podem usufruir desse recurso digital.

FUNDAMENTAÇÃO TEÓRICA

O Catálogo surgiu antes do Corpus, afinal, antes de iniciar a compilação do corpus, era necessário estabelecer quantas e quais obras do Machado de Assis foram traduzidas. Embora a obra do autor em português esteja amplamente catalogada, a mesma situação não se aplica às traduções, que são mais dinâmicas e sujeitas a frequentes retraduções.

A obra do Machado de Assis chegou em terras anglófonas em 1921, com a tradução de três de seus contos na coletânea *Brazilian Tales* (ASSIS, 1921). No entanto, os seus romances só chegaram às prateleiras das livrarias dos Estados Unidos e Reino Unido na década de 1950, desde então, os seus títulos continuaram a ser traduzidos e retraduzidos. Por exemplo, o autor recebeu duas

[112] Doutoranda em Letras Estrangeiras e Tradução (LETRA), Universidade de São Paulo, São Paulo - SP, bolsista CAPES, contato através do e-mail ursulapuellosydio@gmail.com.

retraduções de *Memórias Póstumas de Brás Cubas* (1881) em 2020 e outra retradução de *Dom Casmurro* (1889) em 2023. As recentes retraduções demonstram a importância de um catálogo atualizado, ainda que tenhamos a consciência da impossibilidade de atingir à completude (Pym, 2014). Apesar de sua incompletude, Pym (2014) também enfatiza que os catálogos são fundamentais para a criação de corpora.

O próximo passo foi a construção do Corpus. Como define Tagnin (2015, p.1), os "corpora são bancos de textos de linguagem autêntica, criteriosamente construídos, destinados à pesquisa e legíveis por computador". Além disso, os corpora eletrônicos são o objeto de estudo da Linguística de corpus, que é "uma abordagem empírica para o estudo da língua [...] especialmente útil no estudo da Tradução" (Tagnin, 2015, p.1), pois as ferramentas computacionais permitem a análise de corpora mais volumosos.

A decisão de disponibilizar o Catálogo e o Corpus em um recurso digital voltado para pesquisadores foi tomada com base em projetos similares que precederam o MACC. Entre os websites com catálogos de tradução ou corpora eletrônicos que serviram de inspiração para o MACC, estão o CorTrad do projeto CoMET (TAGNIN; TEIXEIRA; SANTOS, 2009), o COMPARA, da Linguateca (FRANKENBERG-GARCIA; SANTOS, 2002) e o website Poesia Traduzida no Brasil (ASEFF, 2018).

METODOLOGIA

A pesquisa iniciou pela criação do catálogo, ou seja, o primeiro passo foi o levantamento das traduções de obras machadianas para língua inglesa, uma vez que os títulos em português já estavam bem catalogados. Os títulos foram levantados a partir de artigos, teses, catálogos de tradução online e, principalmente, de busca por obras do autor em grandes sites de comércio de livros. Cada título identificado foi registrado em um banco de dados que reunia informações como título em inglês, título em português, gênero literário, tradutor(a), ano de publicação, editora e país. Essa base de dados permitiu traçar um panorama histórico das traduções de Machado de Assis para o inglês, desde 1921 até os dias atuais. Em seguida, o banco de dados foi dividido por gênero literário. Uma tabela era dedicada aos romances, como estes geralmente são publicados fora de antologias, foi fácil estabelecer a obra correspondente em português. Por outro lado, na tabela dedicada aos contos, o processo foi um pouco mais longo, uma vez que a maioria dos contos está reunida em coletâneas e foi necessário investigar conto a conto para concluir a fase de catalogação.

Com as obras catalogadas, a próxima etapa foi a compilação do corpus e ela foi dividida nas seguintes fases:

a) Conversão dos textos dos livros eletrônicos em arquivos .txt;
b) Pré-processamento, limpeza e organização dos textos;
c) 1ª parte do alinhamento: foi criado um programa computacional em Python para extrair cada .txt e converter um único arquivo .csv por obra, com o texto em português na primeira coluna e as traduções nas colunas seguintes;

d) 2ª parte do alinhamento: verificação e correção manual (obra por obra) do alinhamento por parágrafos dos arquivos .csv;
e) Criação de um banco de dados em PostgreSQL, que seria usado para buscas no website do MACC.

Por fim, com o corpus eletrônico, literário (contos e romances machadianos), bilíngue (português e inglês) e paralelo (alinhado por parágrafos) pronto, foi possível dar início a última etapa da pesquisa: a construção de website gratuito e fácil de navegar que contaria com as ferramentas para buscas tanto no Corpus quanto no Catálogo, em palavras, a criação do MACC.

DISCUSSÃO DOS DADOS

O MACC traz os dados mais relevantes levantados durante a pesquisa de forma estruturada, pesquisável e acessível para a comunidade acadêmica em um website gratuito, basta realizar um cadastro prévio.

A seção do Catálogo pode ser consultada de três formas. Na primeira, temos “Linha do tempo” que lista os 39 títulos catalogados em inglês em ordem cronológica. Na segunda, temos a “Busca por obras em língua portuguesa”, que permite o visitante pesquisar a partir do título em português ou do gênero literário. Na terceira, temos a “Busca por obras em língua inglesa”, que permite o visitante filtrar a sua busca a partir de informações como ano de publicação, título em inglês, título em português, gênero literário e país.

A seção de “Busca no Corpus” traz um corpus machadiano que totaliza 2.105.695 palavras, divididas em um subcorpus em português, composto por 6 romances e 90 contos, e outro em inglês, composto por 11 traduções dos romances e 373 traduções dos contos. Através da ferramenta de busca desenvolvida especialmente para o MACC, o visitante pode pesquisar a partir do subcorpus em inglês ou em português.

Em suma, o MACC é fruto de uma pesquisa de mestrado (SYDIO, 2023b), reunindo um catálogo atualizado de traduções da obra machadiana para inglês e o maior corpus paralelo bilíngue da obra machadiana disponível para consulta. A sua principal contribuição para os Estudos da Tradução, os Estudos Machadianos e para a Linguística de Corpus é a disponibilização desses para a comunidade científica.

Agradecimentos

Agradeço à Profa. Dra. Luciana Carvalho Fonseca, por sua orientação durante o mestrado, à Profa. Dra. Stella Esther Ortweiler Tagnin, por sua contribuição durante a qualificação e por me orientar agora no doutorado, e à Profa. Dra. Marlova Gonsales Aseff e à Profa. Dra. Elisa Duarte Teixeira pelas contribuições enquanto a banca examinadora do mestrado.

REFERÊNCIAS

ASEFF, Marlova. Catálogo da poesia traduzida no Brasil (1960-2009). 1. Ed. Brasília, 2018. ISBN: 978-85-540456-0-9. Disponível em: http://poesiatraduzida.com.br/ Acessado em: 10 jul. 2024.

ASSIS, Machado de. **Brazilian Tales.** Trad. Isaac Goldberg. Boston: The Four Seas Company, 1921.

ASSIS, Machado de. **Obra Completa de Machado de Assis**. Rio de Janeiro: Nova Aguilar, 1994. Disponível em: http://machado.mec.gov.br/ Acessado em: 10 jul. 2024.

COSTA, C. B. **DOM CASMURRO EM INGLÊS: TRADUÇÃO E RECEPÇÃO DE UM CLÁSSICO BRASILEIRO**. [s.l.] Universidade Federal de Santa Catarina, 2016.

FRANKENBERG-GARCIA, A.; SANTOS, D. COMPARA, um corpus paralelo de português e de inglês na Web. **Cadernos de Tradução IX**.1 Florianópolis, 2002, pp. 61-79.

PYM, A. **Method in Translation History.** Routledge, 2014.

SYDIO, Ursula Puello. MACC: Machado de Assis Catálogo & Corpus. [S. l.], 1 jan. 2023a.

SYDIO, Ursula Puello. **Machado de Assis Catálogo & Corpus (MACC):** A construção de um catálogo e um corpus paralelo das traduções da obra machadiana para língua inglesa. 126 f. Dissertação (Mestrado) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2023b.

TAGNIN, S. E. O. A Linguística de Corpus na e para a Tradução. Em: **Corpora na Tradução**. São Paulo: HUB, 2015.

TAGNIN, S. E. O.; TEIXEIRA, E. D.; SANTOS, D. CorTrad: a multiversion translation corpus for the Portuguese-English pair. **Arena Romanistica**, v. 4, p. 314-323, 2009.

ANOTAÇÃO SEMÂNTICA MULTIMODAL A PARTIR DO CORPUS AUDITION: UMA CONTRIBUIÇÃO DA SEMÂNTICA DE FRAMES PARA A PESQUISA EM TRADUÇÃO AUDIOVISUAL ACESSÍVEL

Maucha Andrade GAMONAL[113]
Adriana Silvina PAGANO[2]
Tiago Timponi TORRENT[114]

ABSTRACT: This work presents the multimodal semantic annotation conducted through the Audition corpus. The corpus consists of a series of short films from different genres and includes semantic annotation of the original audio transcription, subtitles and closed captions, overlay text, and audio description. The annotation methodology follows the approach of Belcavello et al. (2024), and the theoretical framework supporting the research is Frame Semantics (Fillmore, 1982).

Palavras-chave: Semântica de Frames; Tradução Audiovisual Acessível; audiodescrição; anotação semântica; corpus multimodal.

Introdução

O estudo e o desenvolvimento de metodologias e tecnologias que relacionem conteúdo audiovisual a práticas de acessibilidade é um dever da comunidade científica. Em conformidade com o paradigma dos direitos humanos, a inclusão é um direito social, como ratificado em 2008 pela Convenção sobre os Direitos das Pessoas com Deficiência (Brasil, 2008). Nesse sentido, aliar a Semântica de *Frames* pode ser útil para a Tradução Audiovisual Acessível, especialmente para a prática da audiodescrição, uma vez que possibilita avaliar os efeitos das escolhas lexicais em uma audiodescrição a partir de uma abordagem teórico-metodológica de *frames* semânticos.

Neste trabalho, apresentamos a prática de anotação semântica multimodal no corpus Audition, um corpus compilado com diversos objetos modais, incluindo material audiovisual em formato curta-metragem, transcrição das audiodescrições de cada curta, além de legendas e outras informações visuais. Selecionamos anotações de diversos curtas, dentre eles aquelas produzidas para o curta-metragem Eu não Quero Voltar Sozinho tanto da transcrição da audiodescrição quanto das imagens do curta em que há tais inserções.

---

[113] Residente de pós-doutoramento no Programa de Pós-Graduação em Linguística na Universidade Federal de Minas Gerais, Minas Gerais, atualmente é bolsista Capes. mgamonal@ufmg.br. [2] Professora Titular de Linguística Aplicada da Universidade Federal de Minas Gerais, Minas Gerais, bolsista de produtividade em Pesquisa IC do CNPq.

[114] Professor Associado do Departamento de Letras e do Programa de Pós-Graduação em Linguística da Universidade Federal de Juiz de Fora, Minas Gerais, bolsista de Produtividade em Pesquisa 2, CNPq.

Além da audiodescrição oficial produzida pelo grupo Tramad, cuja anotação foi realizada por Dornelas (2023), nós incluímos a anotação de outra versão de audiodescrição produzida para o mesmo material audiovisual (Vieira, 2015). A metodologia do trabalho segue os procedimentos de anotação de *frames* semânticos da FrameNet Brasil (Torrent et al., 2022) voltadas à prática multimodal (Belcavello et al., 2024).

Os dados de anotação deste trabalho compõem o dataset semântico de objetos multimodais da FrameNet Brasil. Tais dados serão úteis para a aplicação tecnologia linguística em Processamento Automático de Linguagem Natural, como algoritmos de inteligência artificial para rotulação semântica automática.

Fundamentação teórica *Frames* semânticos

O conceito de *frame* na Semântica advém de Fillmore (1985) por meio da longa trajetória de estudos que se consolidou com a proposição da teoria Semântica de *Frames*. Sua proposta se distancia dos estudos formais e se aproxima dos estudos empíricos, uma vez que, como ele mesmo destaca, o interesse está em investigar as continuidades entre a linguagem e a experiência (Fillmore, 1982, p.112). Para ele, as escolhas linguísticas de uma comunidade fala revelam categorias de experiência codificadas por seus membros.

O autor utiliza a analogia da gramática e um conjunto de ferramentas. Assim como as ferramentas são identificadas por suas especificidades de forma e composição, assim é a fonologia e a morfologia de uma língua. E, semelhantemente à gramática, as ferramentas servem a diversos propósitos tendo em vista a grande quantidade de situações para as quais são úteis. Desse modo, Fillmore convida a pensar o texto não como um registro de “pequenos significados” em busca de um “significado maior”, mas como um registro de ferramentas utilizadas para uma determinada atividade (Fillmore, 1982, p.113). Ao *frame* cabe a função de representar tal sistema de conceitos.

> frame é um sistema de conceitos relacionados de modo que, para entender qualquer um deles, é necessário entender toda a estrutura de conceitos na qual se enquadram, quando um dos elementos é introduzido em um textoou em uma conversa, todos os outros serão disponibilizados automaticamente[115]. (Fillmore, 1982, p.111)

As unidades lexicais *construção*.n, *montar*.v, *reformar*.v, *erguer*.v, *reforma*.n se reúnem no *frame* `Construir`[116], de acordo com o repositório de *frames* da FrameNet Brasil. Em sua definição, há ações de montagem ou de construção, em que o AGENTE [117] une um COMPONENTE para formar a ENTIDADE_CRIADA (FrameNet Brasil, 2024). Sabe-se que o *frame* é definido a partir de seus elementos, os chamados Elementos de *Frame* (FE), e esses

[115] A frame is any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available.
Tradução dos autores.

[116] Seguindo convenções estabelecidas, *frames* são registrados em fonte Courier.

[117] Seguindo convenções estabelecidas, elementos de *frames* são registrados em letras maiúsculas.

elementos, por suas vezes, são agrupados por meio da afinidade semântica atribuída ao conjunto de unidades lexicais (LU) que instanciam o *frame*.

1) Até que [alguém AGENTE] construiu CONSTRUIR [um barzinho ENTIDADE_CRIADA], e, pouco a pouco, se tornou um vilarejo completo. [#181606]
2) Em outras partes da cidade, quando queriam reformar [um bairro ENTIDADE_CRIADA], eles demoliam tudo e construíam prédios bem feios no lugar. [#181367]

As sentenças (1) e (2) integram o conjunto de anotações de corpus da FrameNet Brasil. Diz-se que tanto *construir*.v como *reformar*.v são LUs que evocam o *frame* `Construir` nos exemplos ilustrados. Por mais que cada uma tenha suas especificidades semânticas, elas dispõem de afinidades que as conectam a uma categoria da experiência. E isso é visto por intermédio de seus FEs.

Audiodescrição

A AD é uma forma de tradução audiovisual acessível da informação visual em linguagem auditiva verbal (Pagano et al., 2016). Conforme Fryer (2016) destaca, consiste em um comentário verbal que fornece informações visuais para aqueles que não as percebem por conta própria.

A Tradução Audiovisual Acessível (TAV), segundo Fryer (2016, p.2), refere-se à tradução de todos os produtos audiovisuais, incluindo filmes, documentários, programas de televisão e conteúdo online. Como a autora menciona, ao contrário das legendas, dublagens ou *voice-over*, a audiodescrição não se desenvolve a partir de um texto verbal preexistente. Tal realidade, para alguns autores, caracteriza a AD como uma mediação intersemiótica, intermodal ou cross-modal (Jiménez Hurtado, 2007, apud Fryer, 2016).

No Brasil, conforme Franco e Araújo (2022), a efetividade do acesso à audiodescrição e a outras práticas de acessibilidade estão diretamente ligadas ao cenário político do país. Elas destacaram a agenda política do ano vigente da publicação da obra, em que investimentos em projetos culturais foram extintos, além do próprio Ministério da Cultura. Mesmo assim, as autoras enfatizam a resiliência de todos os agentes de cultura e acessibilidade em garantir inclusão por meio da TAV.

Metodologia

Corpus Audition

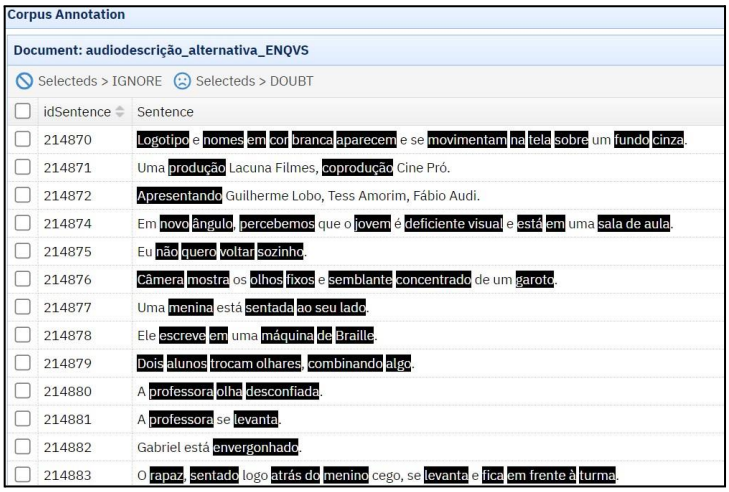
O curta-metragem Eu não quero voltar sozinho é uma obra audiovisual produzida em 2010 pela Lacuna Filmes, cuja audiodescrição foi elaborada pelo grupo Tradução, Mídia e Audiodescrição (TRAMAD). O curta é uma das obras audiovisuais que compõem o Audition, corpus compilado para as tarefas de anotação multimodal da FrameNet Brasil (Silva et al., 2023).

No Audition, os curtas-metragens são de diferentes gêneros com documentários, animações e *live actions*. Em comum, todos possuem a opção de audiodescrição. Para este trabalho, nosso material de análise inclui a audiodescrição oficial do curta e a anotação multimodal de outra versão produzida por Vieira (2015).

Frames semânticos em anotação multimodal

Os procedimentos de anotação seguem conforme a metodologia da FrameNet Brasil (Torrent el al., 2022; Belcavelo et al., 2024). Na Figura 1, as sentenças transcritas da audiodescrição são mostradas no software de anotação webtool. As marcações destacam as Unidades Lexicais que evocam algum *frame* no banco de dados disponível.

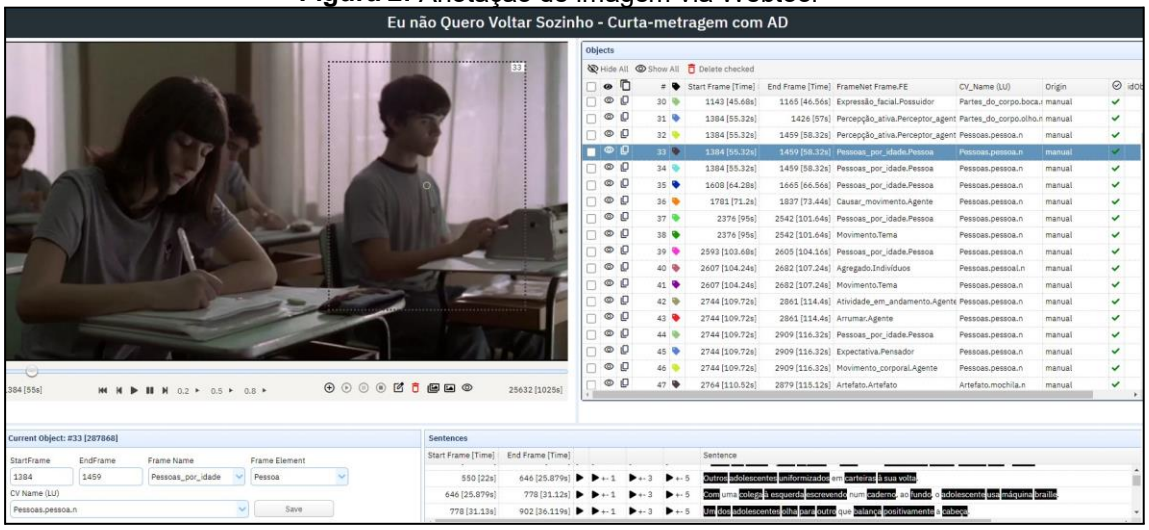
Figura 1: Sentenças transcritas para anotação linguística via Webtool



Fonte: Captura de tela da Webtool (FrameNet Brasil, 2024)

A Figura 2 exibe a tela de anotação de imagem também via Webtool. Por meio dela, o anotador identifica os objetos visuais e os relaciona aos *frames* do banco por meio dos elementos que atuam no *frame*, os chamados Elementos de *Frame*.

Figura 2: Anotação de imagem via Webtool



Fonte: Captura de tela da Webtool (FrameNet Brasil, 2024)

Discussão dos dados

Os resultados da anotação multimodal das duas transcrições das audiodescrições geraram um agrupamento de *frames* semânticos, Elementos de *Frame* e Unidades Lexicais. A partir do produto das anotações, verifica-se a construção de sentido nesses dois materiais audiodescritos, identificando similaridades e particularidades acerca de cada opção tradutória.

A percepção do audiodescritor acerca da obra audiovisual mostra a centralidade da perspectiva adotada na construção de uma AD. A Semântica de *Frames* e a FrameNet Brasil são apresentadas neste trabalho como aparatos teórico-metodológicos tanto para análise quanto para criação de conteúdo audiovisual acessível com o interesse de ser útil para a criação de recursos que possibilitem inclusão e acessibilidade no âmbito audiovisual.

Agradecimentos

Este trabalho recebeu suporte do CNPq por meio dos processos 151361/2023-1 and 313103/2021-6, e da CAPES por meios dos processos 88887.936139/2024-00 and 8887.683333/2022-00. Os autores expressam agradecimento a todos os estudantes e demais pesquisadores que atuaram na criação do corpus Audition.

Referências bibliográficas

BRASIL. Decreto Legislativo nº 186, de 9 de julho de 2008. Aprova o texto da Convenção sobre os Direitos das Pessoas com Deficiência e seu Protocolo Facultativo, assinados em Nova York, em 30 de março de 2007. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/decreto/d186.htm. Acesso em: 15 jul. 2024.

BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. T. Charon: A FrameNet Annotation Tool for Multimodal Corpora *In*: Proceedings of The 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022. Marseille, France: ELRA, 2022, p. 91-96.

DORNELAS, L. A audiodescrição sob a perspectiva da semântica de frames: análise dos frames evocados pelo texto da audiodescrição e pelas imagens dinâmicas num curta-metragem. 2023. Dissertação de mestrado - Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

FILLMORE, C. J. Frame Semantics. *In* The Linguistic Society of Korea (org.). Linguistics in the Morning Calm. Seoul: Hanshin, 1982, p. 111-137.

_____. Frames and the semantics of understanding. *In*.: Quaderni di Semantica. Vol. VI, nº 2, Dezembro de 1985.

FRANCO, E. P. C. ; ARAÚJO, V. L. S. . Audio Description in Brazil. In: Taylor, Christopher; Perego, Elisa. (Org.). The Routledge Handbook of Audio Description. 1ed.Londres: Routledge, 2022, v. 1, p. 596-612.

FRYER, L. (ed.). (2016). An Introduction to Audio Description a Practical Guide. London: Routledge

FRAMENET BRASIL. Software de anotação Webtool. Disponível em: https://webtool.framenetbr.ufjf.br/. Acesso em:15 de jul. 2024.

PAGANO, A. S.; TEIXEIRA, A. L. R.; MAYER, F. A. Accessible Audiovisual

Translation. In: JI, Meng; LAVIOSA, Sara (ed.). *The Oxford Handbook of Translation and Social Practices*. Oxford: Oxford University Press, 2020. cap. 4, p. 66-82.

TORRENT, T. T.; MATOS, E. E.; BELCAVELLO, F.; VIRIDIANO, M.; COSTA, A. D.; GAMONAL, M. A.; MARIM, M. C. Representing context in FrameNet: *a multidimensional, multimodal approach*. Frontiers in Psychology - Language Sciences, 13, article 838441.

SILVA, A. C.; RABELO, I.; OLIVEIRA, I. M.; SOUZA, M.; GAMONAL, M.; ROZA, R. Coleta, composição e etapas de pré-processamento de corpus: procedimentos para a anotação multimodal da FrameNet Brasil. *In*: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 14. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 362-366. VIEIRA, G. M. Audiodescrição do curta Eu não Quero Voltar Sozinho. *In*: Trabalho acadêmico para Núcleo Temático sobre Cinema e Representações Sociais.Universidade Federal do Vale do São Francisco (UNIVASP), 2015.

O PROCESSAMENTO DA LINGUAGEM NATURAL NO ÂMBITO DA PROMOÇÃO DA ACESSIBILIDADE TEXTUAL E TERMINOLÓGICA

Heloísa Orsi Koch DELGADO[118]
Bruna Rodrigues da SILVA[119]

RESUMO: Este recorte de pesquisa apresenta análise de texto sobre Transtorno de Humor Bipolar, sob viés do Processamento da Linguagem Natural. O objetivo foi verificar se o trecho seria compreendido pelo leitor médio brasileiro. Como o texto se mostrou complexo, foram adotados preceitos de tradução intralinguística, com vistas a promover sua acessibilidade textual e terminológica.

Palavras-chave: Acessibilidade Textual e Terminológica; linguagem simples; Transtorno do Humor Bipolar; NILC-Metrix; escolaridade limitada.

INTRODUÇÃO

Este trabalho apresenta recorte de projeto de pesquisa realizado no estágio pós-doutoral da primeira autora e inserido no âmbito do Grupo de Pesquisa Acessibilidade Textual e Terminológica (GEATT) da Universidade Federal do Rio

Grande do Sul (UFRGS). A pesquisa se insere nas áreas de Terminologia, Tradução Intralinguística e Acessibilidade Textual e Terminológica (ATT), tendo como enfoque metodológico o Processamento da Linguagem Natural (PLN).

A partir desse direcionamento teórico-prático, analisamos os elementos textuais e terminológicos — de possível complexidade — de textos sobre o Transtorno do Humor Bipolar (THB)[120], escritos em português do Brasil. Esses trechos fazem parte do dicionário on-line sobre esse transtorno (DicTrans), voltado para estudantes de Medicina e profissionais da saúde. A análise de dificuldade textual teve como referência o perfil de leitor adulto com baixo nível de instrução formal (Ensino Fundamental Completo) e pouca experiência de leitura.

Tal investigação, de natureza qualiquantitativa, baseou-se nos índices de complexidade da linguagem, obtidos por meio de ferramentas linguístico-computacionais. Dessa forma, verificamos se o *corpus* utilizado está adequado ao perfil de leitor desejado ou se mudanças textuais precisariam ser realizadas para elaborar um texto mais acessível. Avaliamos quais trechos deveriam ser reformulados e quais índices seriam apontados pelas ferramentas. Salientamos que, tanto os traços globais, quanto os particulares dos textos foram analisados em um todo de significação e de comunicação, possibilitando que nosso *corpus*

[118] Linguista, tradutora, professora universitária. Universidade La Salle, Canoas, Rio Grande do Sul. heloisa.orsi.koch.delgado@gmail.com.br.

[119] Doutoranda pelo PPG-Letras/UFRGS, professora da rede pública de ensino, Porto Alegre, Rio Grande do Sul.

[120] Pode ser também chamado de Transtorno Afetivo Bipolar (TAB).

servisse para ilustrar dados comparativos e ilustrativos sobre os níveis de complexidade textual nele encontrados.

A escolha pelo THB justifica-se pela condição grave que costuma ser acometida por episódios maníacos e depressivos, afetando 140 milhões de pessoas no mundo, conforme resultados de 2019 da OMS. No Brasil, a Associação Brasileira de Transtorno Bipolar (ABTB) estima que cerca de 8% da população adulta brasileira sofre do quadro psicopatológico (AMARO, 2020). Dados do Ministério da Saúde mostram que foram registrados 4.839.833 procedimentos no Sistema de Informação Ambulatorial (SIA/SUS) entre os meses de março e maio de 2019 (FLORES, s.d.).

FUNDAMENTAÇÃO TEÓRICA

Os textos escritos são os campos naturais de termos de áreas específicas do saber. Krieger & Finatto (2004, p.106) já afirmavam que

> a relevância do texto está diretamente vinculada ao princípio comunicacional que postulam. Isso corresponde a considerar o texto como *habitat* natural das terminologias dotados de elementos linguísticos, pragmáticos, comunicativos e discursivos, e como objeto de comunicação entre destinador e destinatário.

Nesse sentido, Ciapuscio (2003) observa que o uso de determinadas terminologias varia de acordo com o nível para o qual os textos serão destinados: no caso de textos escritos por ou para especialistas, o conceito de um termo é pleno, enquanto para o público geral e leigo, apenas os traços que são relevantes para a caracterização dos termos permanecem. As considerações dessas autoras guiaram a reformulação dos trechos aqui apresentados, buscando a acessibilidade textual e terminológica.

Quanto à reescrita dos trechos originais, fizemos uso das estratégias da Tradução Intralinguística - cuja função é, primordialmente, explicar os signos verbais de uma língua utilizando outros signos da mesma língua - visto que almeja subsidiar uma crescente demanda pela compreensão de discursos técnicos e científicos (ZETHSEN, 2009; JAKOBSON, 1959). Ao descrever a tradução intralinguística e apontar micro estratégias para serem utilizadas, Zethsen (2009, p.16) destaca que, “entre os achados mais importantes, há uma forte tendência à simplificação”.

Com relação ao enfoque aplicado, utilizamos os princípios do PLN, marcado pela observação de índices em textos individuais, de forma que programas computacionais, tais como o NILC-Metrix[121], analisem um texto e disponham de uma quantidade determinada de informações. Portanto, a possibilidade de um tratamento individualizado de cada texto é válida e necessária para pesquisas que tenham como objetivo examinar textos, um a um, para compará-los, de maneira multifatorial, com outros (FINATTO, 2018). Pode-se definir o PLN, então, como uma vertente da inteligência artificial, que incorpora inúmeras técnicas para interpretação da linguagem com base em métodos estatísticos e de aprendizado de um determinado número de regras de

[121] Disponível em http://fw.nilc.icmc.usp.br:23380/nilcmetrix. Acesso em set. 2024.

funcionamento de uma língua por meio de análises de *corpora* de exemplos típicos do mundo real (MANNING & SCHÜTZE, 1999).

METODOLOGIA

Em linhas gerais, os passos metodológicos compreenderam a observação de índices numéricos do *corpus* de estudo, submetido ao NILC-Metrix, de acordo com os preceitos do PLN.

As etapas realizadas foram: *i)* escolha dos excertos textuais disponíveis no DicTrans e envio à ferramenta; *ii)* análise dos índices obtidos sob uma perspectiva quantitativa; *iii)* averiguação dos trechos de potencial complexidade, considerando o perfil do leitor (olhar qualitativo); *iv)* aplicação da tradução intralinguística; *v)* envio da nova versão à ferramenta; e *vi)* análise dos novos resultados. Também foi realizada a repetição dos passos anteriores caso o texto ainda não se mostrasse adequado ao propósito de pesquisa. Ao final, foi realizada a comparação dos índices obtidos entre os textos originais e simplificados.

As métricas de interesse, além da frequência de palavras, são: **simplicidade textual** (reflete a quantidade de palavras em cada sentença do texto, bem como usos mais simples e estruturas sintáticas mais familiares, que apresentem menor desafio para a compreensão) e **simplicidade lexical** (verifica se o texto contém palavras que possuam significado complexo ou evoquem imagens mentais fáceis de processar e de entender). A ferramenta utilizada agrupa métricas desenvolvidas em mais de uma década no NILC, iniciadas com o Coh-Metrix-Port. Trata-se de um sistema computacional que contém por volta de 200 métricas propostas em estudos de discurso, psicolinguística, linguística cognitiva e computacional, que tem o objetivo de analisar a complexidade textual para o português (WICK-PEDRO E SANTOS, 2021). As métricas analisadas aqui são:

Tabela 1 – Métricas do NILC-Metrix

Nome da métrica	Interpretação	Descrição
Índice de Leiturabilidade Flesch **(1)**	Quanto maior o resultado da métrica, menor a complexidade textual.	Busca uma correlação entre tamanhos médios de palavras e sentenças.
Média dos valores das frequências das palavras de conteúdo do texto via Corpus Brasileiro **(2)**	Quanto maior a frequência das palavras, menor a complexidade do texto.	Média dos valores das frequências das palavras de conteúdo do texto, variando entre 1 e 7.
Proporção de sentenças longas em relação a todas as sentenças do texto **(3)**	As longas são mais complexas do que as sentenças curtas e médias; as muito longas, mais complexas do que as longas.	Proporção de sentenças longas em relação a todas as sentenças do texto.
Proporção de sentenças curtas em relação a todas as sentenças	Quanto maior a proporção de sentenças curtas, menos	Proporção de sentenças curtas em relação a todas as sentenças do

do texto **(4)**	complexo é o texto.	texto.

Fonte: NILC-Metrix (2024)

RESULTADOS E CONSIDERAÇÕES

Os resultados apontaram para a complexidade dos textos originais. Sendo assim, de fato, foi preciso fazer uso da tradução intralinguística para que os textos se tornassem potencialmente compreensíveis ao público.

Quadro 1 – Exemplo de trecho sobre episódio maníaco

Episódio maníaco
Reconhece-se um episódio maníaco quando as seguintes evidências comportamentais se manifestam. Período distinto de humor anormal e persistentemente elevado, expansivo ou irritável e aumento anormal e contínuo da energia ou da atividade dirigida a objetivos ou com duração mínima de uma semana e presente na maior parte do dia, quase todos os dias (ou qualquer duração se a hospitalização se fizer necessária). Durante o período de perturbação do humor e aumento da energia ou atividade, três (ou mais) dos seguintes sintomas (quatro dias se o humor é apenas irritável) estiverem presentes em grau significativo e representem uma mudança notável do comportamento habitual: autoestima inflada ou grandiosidade; redução da necessidade de sono; mais loquaz que o habitual; fuga de ideias ou experiência subjetiva de que os pensamentos estão acelerados; distratibilidade, ou seja, a atenção é desviada muito facilmente por estímulos externos insignificantes ou irrelevantes; aumento da atividade social, profissional, escolar ou sexual e agitação psicomotora; envolvimento excessivo em atividades com elevado potencial para consequências dolorosas (surtos desenfreados de compras, indiscrições sexuais ou investimentos financeiros insensatos).

Fonte: Dictrans (2012)

O índice Flesch (1) corresponde ao índice -23.1875, que extrapola o índice mais alto de complexidade, indicando um texto consideravelmente difícil para pessoas com grau de alfabetização limitado e condizente com a de um especialista da área. A métrica (2) indica um valor de 4.48262, apontando para uma maior frequência de palavras de conteúdo (substantivos, verbos, adjetivos e advérbios), ou seja, teoricamente, menor a complexidade do texto. Para a métrica (3), o seguinte cálculo foi feito: o texto registra 3 sentenças de 11, 51 e 112 palavras (1 frase curta e 2 consideravelmente longas). A proporção é 2/3, resultando no índice de 0,66 e indicando que o texto é complexo (as diretrizes da acessibilidade textual postulam que as frases devem conter, no máximo, 25 palavras). A última métrica (4) aponta para o valor de 0,5 feita com base no cálculo da métrica (3), isto é, 1/3. Aqui, evidenciamos um índice um pouco maior do que 0,33 (resultado esperado), o qual pode significar que a diferença (0,17) esteja associada ao excedente de palavras nas frases muito longas (principalmente a de 112 palavras) em comparação à curta. Vejamos, abaixo, a tradução intralinguística com base no texto original.

Quadro 2 – Tradução intralinguística sobre o episódio maníaco

Episódio maníaco
O episódio maníaco acontece quando a pessoa tem o humor diferente do normal. Fica muito alegre e feliz ou irritada e com muita energia por muitas horas durante o dia e por, pelo menos, uma semana. Quando a pessoa fica assim, ela pode se achar melhor do que os outros, dormir menos, falar muito e mudar de um assunto para o outro rapidamente e estar desatenta. Pode gastar demais e conversar sobre sexo sem sentir-se envergonhada. A pessoa pode ficar fora de si e ter que parar de trabalhar, sair com amigos e namorar. Pode ter que ir para o hospital para ficar mais segura e não fazer mal a si e nem aos outros.

Fonte: As autoras.

O conteúdo do texto original consiste em levar o conhecimento sobre o episódio maníaco a pessoas inseridas no mundo médico ou da saúde em geral. Assim, é aceitável que termos da área estejam presentes sem que seja necessário aplicar subsídios da tradução intralinguística para deixar o texto mais fácil de entender.

Porém, quando se trata de textos para uma fatia significativa da população – que possui um baixo nível de instrução e letramento -, estratégias linguísticas que favoreçam à compreensão precisam ser levadas em consideração. Assim, a primeira ação tomada foi a de enxugar o texto. Em seguida, optamos pela paráfrase, ou seja, renúncia a componentes formais ou funcionais do texto que possam levar à falta de acesso ao conhecimento. Essa estratégia resultou em um número significativamente menor de frases e palavras, em que se evidenciam ponderações sobre o todo do texto. Nota-se, portanto, que facilitar um texto extrapola os limites da simples troca de palavras, sugere-se "uma dada escrita por uma alternativa entre várias possíveis" (FINATTO; TCACENCO, 2021).

O índice Flesch (1) foi alterado para 68.45462 (o 100 indica que o texto é muito fácil), revelando um grau de leiturabilidade fácil. A métrica (2) variou para 5.005 (maior do que no texto original), apontando para uma maior frequência de palavras de conteúdo, indicando um texto menos complexo. Para o cálculo da métrica (3), observamos que o texto registra 6 sentenças de 13, 23, 30, 10, 18 e 21 palavras, então, a proporção é 4/6 (quatro frases muito longas[122] para 6 frases no total), resultando no índice de 0,66 e indicando que o texto é complexo. No entanto, dois pontos merecem consideração: i) a tolerância de duas palavras explicadas na nota 9 e ii) as diretrizes da acessibilidade que sugerem o limite de 25 palavras aceitável e não de apenas 15 para frases longas e mais de 15 para muito longas. A última métrica (4) mudou para 0,28571, isto é, 1/6. Aqui, evidenciamos um índice menor do que 0,5, indicando a existência de mais frases longas do que curtas. Cabe, da mesma forma, considerar o exposto na métrica (3) e salientar que o uso de outras ferramentas poderia evidenciar resultados diferenciados (por conta de diferentes descrições e interpretações das métricas), oferecendo contrapontos relevantes. Dessa forma, seria possível afirmar que, para fins deste recorte de estudo, a reformulação simplificada apontou para pontos de reflexão dos resultados, mostrando que deve haver uma ponderação

[122] Como o texto não apresenta frases longas e a classificação entre as sentenças curtas, médias, longas e muito longas possui uma tolerância de apenas duas palavras, pode haver um desvio de classificação na contagem das frases em função dessa tolerância.

sobre os dados quantitativos. Estes devem contar com a reflexão qualitativa baseada no olhar e na experiência de tradutores intralinguísticos, direcionados para a simplificação textual, como forma de encontrar pontos de equilíbrio essas duas abordagens. Por fim, reforçamos a relevância da testagem entre leitores reais, para validar estratégias e identificar problemas ainda impensados.

REFERÊNCIAS

AMARO, D. **Entre a euforia e a depressão: 8% da população é bipolar**. Edição do Brasil. Disponível em http://edicaodobrasil.com.br/2020/03/20/entre-euforia-ehttp://edicaodobrasil.com.br/2020/03/20/entre-euforia-e-depressao-8-da-populacao-brasileira-e-bipolar/depressao-8-da-populacao-brasileira-e-bipolar/, 2020.

CIAPUSCIO, G. **Textos especializados y terminología**. Barcelona: IULA, 2003.

DICTRANS, Dicionário sobre o Transtorno do Humor Bipolar (DELGADO, H.O.K.; VERNETTI, C. L. ; SANTOS, C. A. dos). Porto Alegre, 2019. Disponível em:

https://www.dictrans.org/conheca.php. Acesso em: junho de 2023.

FINATTO, M. J. B. Corpus-amostra português do século XVIII: textos antigos de Medicina em atividades de ensino e pesquisa. DOMÍNIOS DE LINGU@GEM , v. 12, p. 435, 2018.

FINATTO, M. J. B.; TCACENCO, L. M. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. **Tradterm,** 37 (1), p. 30-63, 2021. Disponível em:https://www.revistas.usp.br/tradterm/article/view/168327. Acessado em: 02 jun. 2021. DOI https://doi.org/10.11606/issn.2317-9511.v37p30-63

FLORES, J. **Entre crises e likes**. Blog da UOL. Disponível em:

https://www.uol.com.br/vivabem/reportagens-especiais/transtorno-afetivo-bipolar-oque-e-por-que-ele-pode-piorar-com-a-pandemia. Acesso em: jan. de 2023.

JAKOBSON, R. **On linguistic aspects of Translation**. Massachusetts: Harvard University Press, 1959.

KRIEGER, M. G., FINATTO, M. J. B. **Introdução à Terminologia**: teoria & prática. 1a ed. São Paulo: Contexto, 2004.

WICK-PEDRO, G.; SANTOS, R. L. S. Complexidade textual em notícias satíricas: uma análise para o português do Brasil. *In*: **SIMPÓSIO BRASILEIRO**

DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL). Porto
Alegre: Sociedade Brasileira de Computação, 2021. p. 409-415.

MANNING, C.D., SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Massachussets: MIT Press, 1999.

ZETHSEN, K. K. (2009). **Intralingual Translation**: An Attempt at Description. Meta, 54 (4), 795–812. https://doi.org/10.7202/038904ar

CORPUSCRIPT: AN AUTOMATED TEXT-CLEANING TOOL FOR CORPUS LINGUISTICS

Jhonatan Henrique LOPES Alves[123]
Ana Eliza Pereira BOCORNY[124]
Deise Prina DUTRA[125]
Carolina Godoi de Faria MARQUES [126]
Gustavo Leal TEIXEIRA[127]
Danilo Duarte COSTA[128]

Introduction

The process of corpus compilation remains a significant challenge in the field of corpus linguistics. This paper introduces CorpuScript, an innovative text-cleaning software aimed at aiding researchers in the process of corpus preparation. By combining software engineering with corpus linguistics methods, this tool can significantly improve the workflow for corpora compilation, specifically in the task of corpus cleaning.

The necessity for CorpuScript emerged from recurring challenges experienced by our research team, particularly during our current corpus research project, in which a considerable large number of texts needed to be cleaned before being used for data analysis.

Considering the pressing need for an automated solution that could improve the text-cleaning process in our research project, CorpuScript was carefully developed to help us accelerate the corpus compilation, while meeting the requirements outlined in our corpus design.

THEORETICAL FRAMEWORK

The importance of clean, well-prepared corpora in linguistic research is wellestablished in the literature. Biber, Conrad, and Reppen (1998) emphasize that corpusbased investigations rely on empirical analysis of large, principled collections of natural texts.

Notably, a standard procedure in corpus building is the conversion of the selected texts into plain text (ASCII or UTF-8), since this type of file format can be run in most corpus analysis tools, as mentioned in Ädel (2020) and Reppen (2022).

---

[123] Undergraduate Student, Universidade Federal de Minas Gerais, Belo Horizonte, MG. Scholarship: FAPEMIG (APQ-01173-22)

[124] Professor, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[125] Professor, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

[126] Doctorate Student, Universidade Federal de Minas Gerais, Belo Horizonte, MG. Scholarhip: CAPES

(n. 88887.939578/2024-00)

[127] Professor, Universidade Federal de Minas Gerais, Montes Claros, MG

[128] Professor, Universidade Federal do Vale do Jequitinhonha e Mucuri, Diamantina, MG

However, Gries and Newman (2013, p. 263), point out that files will still “almost invariably require some editing for them to be used most effectively”. To ensure they can be read by computer software that are used for corpus analysis, Coxhead (2020, p. 470) stresses that text files “should be as clean as possible”.

In this context, corpus cleaning refers to the task of removing extraneous material from text data, such as headers, footers, special characters, line breaks, and other non-linguistic elements that do not contribute to the actual linguistic content (Weisser, 2016). Cleaning a corpus is a fundamental step in the corpus compilation process, since those unwanted items “may adversely affect the accuracy of the analytical procedures we intend to carry out, as well as impinging on the corpus’s representativeness” (McEnery and Brooks, 2022, p. 43).

Although cleaning many texts manually is rather time-consuming, it remains a frequently used method. For instance, when instructing English language students to use corpora for improving their learning, Poole (2018) suggests that they clean their texts by using the ‘find and replace tool’ in a text processor. Similarly, a non-automatic text cleaning approach was adopted in a study by Charles (2015) with students of English for Academic Purposes (EAP).

Automating the text cleaning process is certainly a highly welcomed advancement. To this end, according to Anthony (2020), high-level, functional programming languages are well-suited for the creation of brief, straightforward programs aimed at expediting the cleaning and processing of corpora. Languages such as Pearl, Python, and R have been extensively employed in corpus linguistics applications, encompassing tasks involved in corpus cleaning.

METHODS

CorpuScript was developed using Python, incorporating a suite of robust libraries to manage various aspects of text processing. The primary libraries utilized include:

1) Regular Expressions (“re” module): Fundamental for executing complex pattern matching and substitution operations essential for text cleaning.

2) SpaCy: A comprehensive natural language processing library employed for tasks such as tokenization, lemmatization, part-of-speech tagging, and stop word removal, thereby enhancing the linguistic accuracy of the text processing pipeline.

3) BeautifulSoup (bs4): Utilized for parsing and stripping HTML content from textual data, ensuring that only plain text is processed.

4) PySide6: Leveraged to develop the graphical user interface (GUI), facilitating user interaction and accessibility for researchers without programming expertise.

Additional Python Standard Libraries: Modules such as os, sys, unicodedata, logging, json, urllib.request, time, random, multiprocessing, and threading were integrated to handle file operations, system interactions, logging mechanisms, JSON data processing, network requests, time management, concurrency, and synchronization.

The core text cleaning functionality of CorpuScript is structured through a modular preprocessing pipeline, comprising the following key steps to standardize and prepare textual data for analysis:

HTML Stripping: Implemented via BeautifulSoup, this step removes any embedded HTML tags within the text, ensuring that only unformatted text is retained for subsequent processing.

Character Filtering: This module removes specified characters or sequences from the text based on user-defined parameters, allowing for the exclusion of unwanted symbols or tokens that may interfere with text analysis.

Diacritic Removal: Utilizing the unicodedata module, this process eliminates diacritical marks from characters, normalizing the text to its basic alphabetic form and enhancing consistency across different text inputs.

Script Filtering: Specific modules are employed to remove characters from nonLatin scripts, such as Greek and Cyrillic, maintaining a uniform character set within the corpus and eliminating potential noise from multilingual data.

Unicode Normalization: Applied using unicodedata.normalize with the NFKC (Normalization Form KC) standard, this step ensures that characters are represented in a consistent and compatible form, reducing discrepancies caused by varied Unicode encodings.

Whitespace Normalization: This process involves adjusting whitespace by removing unnecessary spaces preceding punctuation marks, standardizing spacing around punctuation, brackets, and braces, and collapsing multiple consecutive whitespace characters into a single space, thereby enhancing the readability and uniformity of the text.

Line Break Removal: By replacing newline characters with spaces, this module transforms multiline text into a continuous flow, which is beneficial for certain types of text analysis.

Bibliographical Reference Removal: Through the use of regular expressions, this step detects and removes bibliographical references embedded within the text, such as in-text citations, to focus the analysis on the main content.

Lowercasing: Converting all text to lowercase ensures uniformity, facilitating case-insensitive processing and comparison in subsequent analysis stages.

Lemmatization: Utilizing SpaCy’s lemmatization capabilities, this module reduces words to their base or dictionary forms, which aids in consolidating different morphological variants of a word, thus improving the semantic consistency of the corpus.

Tokenization: This process involves splitting the text into sentences or words using SpaCy's tokenization tools, enabling more granular analysis and manipulation of the textual data.

Stop Word Removal: SpaCy's predefined stop word list is employed to filter out common, non-informative words, thereby focusing the analysis on more meaningful and content-rich terms.

Unicode Category Filtering: This module removes characters belonging to specific Unicode categories, such as superscript and subscript characters, further refining the text and eliminating potential formatting artifacts.

Regular Expression Substitutions: Advanced pattern matching and replacements are conducted using user-defined regular expressions, allowing for customizable and flexible text cleaning operations tailored to specific dataset requirements.

The preprocessing pipeline is designed to be highly modular and configurable, enabling users to selectively apply cleaning steps based on their specific research needs. Each preprocessing module is implemented as a distinct component, facilitating ease of maintenance, scalability, and the ability to extend or modify the pipeline as needed. This modular architecture ensures that CorpuScript can accommodate a wide range of text processing tasks, from simple cleaning operations to more complex linguistic transformations, thereby supporting comprehensive corpus preparation for subsequent linguistic analysis.

Furthermore, CorpuScript's GUI, developed with PySide6, provides an intuitive interface for configuring processing parameters, selecting files or directories for processing, and monitoring progress through real-time feedback mechanisms. Concurrent processing capabilities, managed via Python's multiprocessing and threading modules, enable efficient handling of large datasets by leveraging multiple CPU cores. Logging functionalities ensure that all processing activities are meticulously recorded, facilitating debugging and audit trails.

RESULTS AND DISCUSSION

The implementation of CorpuScript has the potential to profoundly impact corpus compilation and research. The most striking advantage is the considerable reduction in time spent on pre-processing time.

A prime example of this efficiency gain was observed in our large-scale corpus research project. Initially, we estimated a six-month period for corpus preparation alone. However, with the introduction of CorpuScript midway through the project, we were able to complete the preparation phase in a matter of days, demonstrating the software's significant impact on research productivity.

The software's ability to maintain consistency across large volumes of text has also improved the quality of prepared corpora. By minimizing human error and ensuring uniform application of cleaning rules, CorpuScript can contribute to the reliability and validity of corpus-based studies.

CONCLUSION

CorpuScript represents a significant advancement in the field of corpus linguistics. By automating and streamlining the text cleaning process, it addresses long-standing challenges in corpus preparation, considerably reducing processing time, while minimizing human error and ensuring consistency across large corpora.

The software's impact extends beyond time-saving, enabling researchers to work with larger corpora and conduct more comprehensive analyses, thereby contributing to the advancement of corpus linguistics research.

As we continue to refine and expand the capabilities of CorpuScript, we invite collaboration and feedback from both the linguistic and software engineering communities. The goal is to further enhance its functionality and broaden its applicability to other scientific domains, ultimately contributing to more efficient, accurate, and comprehensive linguistic research. While the current version of CorpuScript has already demonstrated significant value, several points for future enhancement have been identified.

These future developments aim to further enhance the software's functionality, adaptability, and integration with existing research workflows, solidifying its role as an essential tool in corpus linguistics research.

AKNOWLEDGEMENTS

The authors wish to thank the following organizations for supporting and funding the research reported here: Federal University of Minas Gerais and Grant #APQ01173-22, Minas Gerais, Research Foundation (FAPEMIG).

REFERENCES

ÄDEL, Annelie. Corpus compilation. In: PAQUOT, Magali; GRIES, Stefan Th (Eds.). **A practical handbook of corpus linguistics**. Springer Nature, 2020.

CHARLES, Maggie. Same task, different corpus. In: BOULTON, Alex; LEŃKOSZYMAŃSKA, Agnieszka. **Multiple affordances of language corpora for datadriven learning**. John Benjamins 2015, p. 131-154, 2015.

GRIES, Stefan; NEWMAN, John. Creating and using corpora. In: PODESVA, Robert J.; SHARMA, Devyani (Ed.). **Research methods in linguistics**. Cambridge University Press, 2014.

MCENERY, Tony; BROOKES, Gavin. Building a written corpus: what are the basics?. In: O'KEEFFE, Anne; MCCARTHY, Michael (Ed.).**The Routledge handbook of corpus linguistics**. 2nd Edition. Routledge, 2022. p. 35-47.

POOLE, Robert. **A guide to using corpora for English language learners**. Edinburgh University Press, 2018.

REPPEN, Randi. Building a corpus: what are key considerations?. In: O'KEEFFE, Anne; MCCARTHY, Michael (Ed.).**The Routledge handbook of corpus linguistics**. 2nd Edition. Routledge, 2022. p. 13-20.

HOW TO USE SHAPE AND STEM CORPORA TO HELP RESEARCH-PAPER WRITING IN ENGLISH FOR ACADEMIC PURPOSES CLASSES[129]

Paula Tavares PINTO[130]
Luciano Franco da SILVA[131]
Talita SERPA[132]
Diva Cardoso de CAMARGO[133]

RESUMO: Corpora do tipo "faça-você-mesmo" são bancos de dados linguísticos poderosos que podem ser usados para apoiar a redação acadêmica e a tradução nas áreas de Humanidades, Ciências e Matemática (VANTAROLA, 2002; MAIA, 2002; FRANKENBERG-GARCIA, 2019; CARVALHO et al., 2021). Este trabalho discutirá as possibilidades de compilar rapidamente dois corpora especializados nas áreas de SHAPE e STEM com a ferramenta AntCorGen (ANHONY, 2019). Ambos os corpora serão explorados com o Sketch Engine (KILGARIFF, 2004) para mostrar como os pesquisadores podem usá-los para redigir seus próprios artigos acadêmicos. Os leitores aprenderão maneiras de encontrar adjetivos e verbos frequentes e relevantes, bem como blocos lexicais usados para dar ênfase à escrita acadêmica. Além disso, eles encontrarão formas específicas de explorar os corpora de SHAPE e STEM para identificar estruturas acadêmicas recorrentes para cada seção de um artigo acadêmico, ou seja, introdução, metodologia, discussão e conclusões.

Palavras-chave: linguística de corpus; redação de artigos científicos; corpora DIY; disciplinas SHAPE e STEM.

ABSTRACT: Do-it-yourself corpora are powerful language databases that can be used to support academic writing and translation in the areas of Humanities, Science and Math (VANTAROLA, 2002; MAIA, 2002; FRANKENBERG-GARCIA, 2019; CARVALHO et al., 2021). This paper will discuss the possibilities of quickly compiling two specialized corpus in the areas of SHAPE and STEM with the tool AntCorGen (ANHONY, 2019). Both corpora will be explored with Sketch Engine (KILGARIFF, 2004) to show how researchers can use them to write their own research papers. Readers will learn about ways to find frequent and relevant adjectives and verbs, as well as lexical bundles that are used to bring emphasis to their academic writing. Also, they will find specific ways to explore SHAPE and STEM corpora to find recurrent academic structure for each research paper section, that is to say, the introduction, methodology, discussion and conclusions.

Keywords: corpus linguistics; research paper writing; DIY corpora; SHAPE and STEM disciplines.

---

[129] Based on Pinto et al. (2024), available at <https://lume.ufrgs.br/bitstream/handle/10183/272634/001197497.pdf?sequence=1> Access on Oct. 12th, 2024.

[130] 2 Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo(CNPq).paula.pinto@unesp.br.

[131] Instituto Federal do Paraná (IFPR), Goioerê, Paraná.

[132] Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo(CAPES).

[133] Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo(CNPq).

Introduction

Writing research papers in English may be a challenge for newcomer authors at the beginning of their academic careers. For those who are non-native speakers of English and did not have the chance to use academic English with frequency it may be even harder. Most of the time these researchers are used to reading scientific papers, but do not have much experience in writing them, which may take years of experience and hard work.

Some of the scholars who have studied academic writing in depth are Swales and Feak (2004, 2009), Hyland (2004, 2014), Lee and Swales (2006), Cortes (2007), Flowerdew (2010). Even though these authors have widely described the features of academic writing, there are some characteristics that may still not be as salient for novice researchers such as the use of academic collocations and lexical bundles. Some authors use word combinations that do not sound natural to their scientific community and this may impair their article acceptance. Some of the scholars who have pointed out the academic issues found in research papers of non native speakers of English are Charles (2012), Howarth (2013), Chang and Swales (2014), Karpenko-Seccombe, (2020), Pinto et al. (2021) and Pinto et al. (2024).

In this context, Corpus Linguistics has played an important role in providing a range of writing tools to help researchers from different fields to find language patterns in academic discourse that are recognized by their peers. This happens because authors will rely on large collections of academic texts, hereafter, corpora, which can show them how their research community generally writes and the specific terminology and frequent patterns that can be rapidly identified and retrieved for writing purposes. This methodological approach can be used in different areas such as Math, Humanities and Biological areas. In order to do that, authors can use pre-compiled specialised corpora or compile their own collection of research papers published in high impact journals and use them as a Do-it-Yourself corpus (VANTAROLA, 2002; MAIA, 2002; FRANKENBERG-GARCIA et al., 2019; PINTO et al., 2024). By using corpora, the writer will be able to observe the useful information according to his specific needs and will develop an autonomous process of learning that will lead him to mastering the academic English based on his interpretation of his peers' writing.

This paper will discuss how specialised corpora can be explored by researchers who want to compile their own language database to help them write different sections of their own research papers. We will illustrate our proposal by taking examples from SHAPE disciplines, which involve Social Sciences Humanities, Arts for People and Economy, as well as STEM disciplines, which involve Science, Technology, Engineering, and Mathematics.

AntCorGen for the compilation of SHAPE and STEM areas

AntCorGen (ANTHONY, 2019) is a tool used to quickly compile specialised corpora with research papers from the PLOS one platform. A tutorial video of this tool was recorded by its creator in a short video[134]. Below we will talk about the

[134] AntCorGen tutorial <<https://www.youtube.com/watch?v=WrsIzE9to4o> access on June 30th, 2023. [7] PLOS available at <https://plos.org/about/> access on June 27th, 2023.

compilation of SHAPE Plos and STEM Plos and their exploration for academic writing.

SHAPE disciplines stand for Social Sciences Humanities, Arts for People and Economy. All these disciplines and subareas can be found at PLOS, which is a nonprofit, open access multi-disciplinary publisher[7]. All areas of SHAPE can be easily accessed in AntCorGen and the researcher can choose the parts of research papers he wants to analyse. Since we wanted to have mostly written material we selected the articles' abstracts, introduction, materials & methods, results & discussion and conclusions.

We called this corpus SHAPE Plos and, since it was compiled for describing the process in this chapter, we set the maximum of 100 articles , but it is possible to have a much larger study corpus if we wanted to. After this compilation we had a study corpus of 445,291 words to be explored.

STEM disciplines are related to both Biology and Hard Sciences. Although the figure below seems to have only Biology and Life Sciences, the actual list of disciplines selected was longer and we could include areas such as Math and Computer Sciences as well. In the same way, we selected 100 articles for STEM Plos corpus.

After this compilation, we had a specialised corpus of STEM disciplines with a total number of 297,255 words to be observed and compared to the results from SHAPE Plos.

Analyses with Sketch Engine

We uploaded both corpora, SHAPE and STEM to Sketch Engine (KILGARIFF, 2014) so we would be able to observe the frequent adjectives and verbs in each broad area and see the similarities and differences between them. We could also generate concordance lines with search words, terms and phrases that can be used by researchers to explore and observe how international researchers in their area have been writing different sections of their research papers.

Building your Research paper with SHAPE Plos and STEM Plos corpus

If a researcher wants to have examples of research papers in SHAPE and STEM disciplines, he can search for common expressions in the corpus. In our case, we have divided both subcorpora into research sections that are usually found in research articles. Based on Karpenko-Seccombe (2020), we are going to discuss how researchers can use their own specialised corpora for writing their research papers.

Writing the Introduction Section

According to Swales and Feak (2009, 2011), a research paper introduction typically contains three main steps or *moves*: a) establishing the area of research, where the author will show the importance of a field and introduce previous research in his area; b) establishing a gap in the knowledge or problem to be solved and c) presenting his paper, where he will identify his objectives, introduce expected outcomes and describe the structure of his work.

In order to explore introductions in SHAPE Plos and STEM Plos corpora, we searched for concordance lines with the query phrase " this paper" and we selected some of the lines to be used as examples here:

1. **This paper** attempts to fill the gap of existing research concerning the link between public pension and fertility. [SHAPE Plos]

2. In **this paper** ,we perform a comprehensive survey of the worldwide linguistic landscape as emerging from mining the Twitter microblogging platform. [SHAPE Plos]

3. In **this paper** , we are interested in measuring linguistic regularities both at the level of word structure and at the level of word order. [SHAPE Plos]

4. **This paper** explores the ways abortion attitudes intersect with causal beliefs about gender categories, within the unique social context of a national referendum held to legalise abortion in the Republic of Ireland. [SHAPE Plos]

5. In **this paper**, we introduce a novel mobile application called "Medikamentenplan" ("Medication Plan"), which was developed to support medication compliance and vital sign documentation. [STEM Plos]

6. In **this paper**, we propose a concise, improved and effective privacy framework for wearable device manufacturers, as well as application developers, capable of providing greater privacy and security to the wearable device owners. [STEM Plos]

7. **This paper** innovatively proposes countermeasures to improve the innovation of e-commerce practitioners in rural areas. [STEM Plos]

8. The objective of **this paper** is to outline our approach of establishing and implementing this IT infrastructure. [STEM Plos]

We can see that authors from SHAPE and STEM use similar strategies to introduce their research papers. In 1, 4 and 7, authors used the structure *This paper + [adverb] + verb (infinitive)*. In examples 2, 3, 5 and 6, authors opted to use *In this paper + we + verb (infinitive)*. Finally, in example 8, the author preferred to introduce his paper by using the structure *The objective of this paper is + to + verb (infinitive)*.

We can see a pattern in the previous examples that can be used in a more confident way by researchers of SHAPE and STEM.

Final Considerations

In this paper we presented an overview on how to compile specialised corpora in SHAPE and STEM with the AntCorGen tool and how researchers can use those corpora to access the academic language used by their peers. By doing so, researchers will confirm or refute ways of presenting their studies according to each research paper section, as well as the best way of describing their methodological approach, and call attention to their studies contribution. We hope this chapter may inspire research teams to start building their own language database that can be used by future members and can be constantly updated.

Acknowledgments
The authors would like to thank the support by CNPq (Process Number #307287/2021-1); FAPESP (Process Number # 2022/05908-0); CAPES.

References

ANTHONY, L.. AntCorGen (Version 1.1.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software, 2010.

FRANKENBERG-GARCIA, A.; BOCORNY, A. E. P.; TAVARES-PINTO, P.; SARMENTO,
S. Supporting the internationalization of Brazilian research. *Workshops delivered at the Federal University of Rio Grande do Sul and at São Paulo State University*, Porto Alegre and São José do Rio Preto, April-June 2019. 2019.

CARVALHO, C. T. de; LARANJA, L. A. N.; PINTO, P. T. DIY Corpora: o que são e para quem são? *Tradterm*, v. 37, n. 1, p. 64-87, 2021. Available at: https://doi.org/10.11606/issn.2317-9511.v37p64-87. Access on Oct. 12th,. 2024.

CHANG,, Y. Y., & SWALES, J. M. Informal elements in English academic writing: threats or opportunities for advanced non-native speakers?. In *Writing: Texts, processes and practices* (pp. 145-167). Routledge, 2014

CHARLES,, M. 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, *31*(2), 93-102, 2012.

CORTES, V. Exploring genre and corpora in the English for academic writing class. *The ORTESOL Journal*, *25*, 8-14, 2007.

FLOWERDEW, L. Using corpora for writing instruction. *The Routledge handbook of corpus linguistics*, 444-457, 2020

HOWARTH, P. A. *Phraseology in English academic writing*. Max Niemeyer Verlag, 2013

HYLAND, K. *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press 2004.

HYLAND, K. Disciplinary discourses: Writer stance in research articles. In: _____. *Disciplinary discourses*: Social interactions in academic writing. 2. ed. Londres: Routledge, 2014. p. 99-121.

KARPENCO-SECCOMBE, T. *Academic writing with corpora: A resource book for data-driven learning*. Routledge, 2020.

KILGARIFF, A., BAISA, V., BUSTA, J., JAKUBÍČEK, M., KOVÁR, V., MICHELFEIT, J., ... & SUCHOMEL, V. The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36, 2014.

LEE, D., & SWALES, J. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for specific purposes*, *25*(1), 56-75, 2006.

MAIA, B. Do-it-yourself, disposable, specialised mini corpora–where next? Reflections on teaching translation and terminology through corpora. Cadernos de tradução, 1(9), 221-235, 2002.

PINTO, P.T.; SILVA, Luciano Franco da ; SERPA, TALITA ; CAMARGO, DIVA CARDOSO DE . Do-It-Yourself Corpora to Support SHAPE and STEM Research Paper Writing. In: SARMENTO, S., REBECHI, R., MATTE M. L.. (Org.). *English for Academic purposes: reflections, description e pedagogy*. 01ed.Porto Alegre: Zouk, 2024, v. 01, p. 97-126.

PINTO, P. T.; CAMARGO, D. C. de; SERPA, T.; SILVA, L. F. da. Analysing the behaviour of academic collocations in a corpus of research-papers: a data-driven study/Analisando o comportamento de colocações acadêmicas em um corpus de artigos científicos: um estudo dirigido por dados. *Revista de Estudos da Linguagem*, v. 29, n. 2, p. 1229-1252, 2021.

SKETCH ENGINE <https://auth.sketchengine.eu/#login> Access on March 13th, 2024.

SWALES & FEAK, C. B. *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). Ann Arbor, MI: University of Michigan Press, 2004.

SWALES & FEAK, C. B. *Abstracts and the writing of abstracts* (Vol. 2). University of Michigan Press ELT, 2009.

TAVARES-PINTO, P.; REES, G.; FRANKENBERG-GARCIA, A. Identifying collocation issues in English L2 research article writing. In: CHARLES, Maggie; FRANKENBERG-GARCIA, Ana (org.). *Corpora in ESP/EAP Writing Instruction*: Preparation, Exploitation, Analysis. 1. ed. Londres: Routledge, 2021. p. 01-20.

"VOCÊ ESTÁ TENDO PRAZER COM SEU TORTURADOR?" A CONDIÇÃO FEMININA NOS RELATOS DE TORTURA À COMISSÃO NACIONAL DA VERDADE

Eu sou leve, sabe, eu tô viva, estamos vivos, vamos ficar vivos. Por que olhar pra trás? Não vive quem fica arrastando cordéis de caixões

Regina Duarte

Giovana de Castro MARCHESE [135]
Luciana Carvalho FONSECA [136]

RESUMO: Esta pesquisa faz uma análise testemunhos de mulheres que foram presas políticas durante o regime ditadura empresarial-militar no Brasil à Comissão Nacional da Verdade, com o fim de infestigar como o gênero é performado em seus discursos ao narrarem a violência sexual sofrida nas sessões de tortura. Este estudo lança mão de Foucault (1970, 1975), Butler (2018, 2004) e Segato (2022) para as análises, e da ferramenta Sketch Engine para investigação do corpus.

Palavras-chave: estupro; ditadura militar; patriarcado; memória; Sketch Engine.

Introdução

Em tempo de revisionismo histórico, trabalhos em prol da memória pública são um imperativo ético no Brasil para que possamos recuperar a força das lutas sociais que trazem à luz a violência do regime ditatorial de 1964 a 1985. A epígrafe deste resumo é um trecho retirado de uma entrevista concedida por Regina Duarte, secretária de Cultura do governo Bolsonaro em 2020, à CNN

Brasil no dia 7 de maio de 2020[3]. Quando questionada pelo jornalista Daniel Adjuto sobre tortura durante a ditadura militar no Brasil, a secretária ri e afirma que "na humanidade, não para de morrer" e que "sempre houve tortura[137]". Ao minimizar as mortes e sevícias causadas pelo Estado brasileiro, Regina Duarte contribui para a naturalização da violência no país, fazendo eco com os discursos do então presidente Jair Bolsonaro.

---

[135] Doutoranda no Programa de Estudos Linguísticos e Literários em Inglês do Departamento de Letras Modernas – DLM, Faculdade de Filosofia, Letras e Ciências Humanas – FFLCH, Universidade de São Paulo – USP, São Paulo, SP. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Email: giovana.cmr@usp.br

[136] Professora Doutora nos Programas de Pós-Graduação Estudos Linguísticos e Literários em Inglês (ELLI) e Letras Estrangeiras e Tradução (LETRA), Departamento de Letras Modernas – DLM, Faculdade de Filosofia, Letras e Ciências Humanas – FFLCH, Universidade de São Paulo – USP, São Paulo - SP

[137] A entrevista pode ser assistida na íntegra no canal da CNN Brasil no YouTube: https://www.youtube.com/watch?v=v9gLHrP7RNw Acesso em: 07 de outubro de 2024

A normalização do regime ditatorial e, consequentemente, da tortura e a desqualificação das memórias críticas à ditadura e dos atores a ela relacionados nos impossibilitam de lidar com nossos erros históricos, o que contribui para a perpetuação das estruturas de violência e de opressão. Pauta (não tão) velada do governo Bolsonaro (2019-2022), que tinha no torturador Carlos Brilhante Ustra um herói. Um exemplo dessa perpetuação da violência é o aumento de 10,8% dos casos de feminicídio no primeiro semestre de 2021 em relação a 2019, primeiro ano do governo bolsonarista.

Dessa forma, buscando desempenhar o meu papel de cidadã brasileira e pesquisadora em consonância com a perspectiva delineada por Pedretti (2021, p. 54), que postula a responsabilidade das instituições públicas e da sociedade civil, incluindo acadêmicos e movimentos sociais, no esforço de reorganizar as demandas por *memória, verdade e justiça* após quatro anos de desarticulação dessas iniciativas, apresento esta pesquisa que examina os relatos de tortura durante a ditadura militar submetidos à Comissão Nacional da Verdade (doravante CNV, 2012 – 2014). O enfoque deste estudo está nas narrativas femininas que delineiam a violência de gênero perpetrada contra mulheres, a fim de investigar, com base em uma análise crítica do discurso, como gênero é performado nesses discursos.

Fundamentação teórico-metodológica

O objeto de estudo desta pesquisa é um corpus monolíngue em português, *offline*[138], intitulado CNV_Mulheres, composto pela transcrição dos depoimentos de presas políticas à CNV entre 2012 e 2014[139]. O corpus é composto por 103 depoimentos.

A ferramenta *Sketch Engine*, software de análise linguística desenvolvido pela Lexical Computing CZ, foi usada para, através das palavras-chave, levantar o tema central do corpus. Durante a investigação, me chamou atenção o grande número de palavras relacionadas à violência sexual, como *estuprar, violentar, abuso, (chamar de) puta*. Foi feito um levantamento das palavras-chave dentro desse campo lexical e, acessando as linhas de concordância correspondentes a elas, foi dado início à análise dos relatos.

Assim, com base em Foucault (1970, 1975), Butler (2018, 2004) e Segato (2022) esta pesquisa investiga como gênero é performado nos discursos sobre tortura sexual de ex-presas políticas em seus testemunhos à CNV. Discussões preliminares lançam luz sobre discursos de incerteza do estupro, negação do estupro, dessubjetivação política da mulher militante e menstruação como instrumento de tortura.

---

[138] Para os vários tipos de corpora, ver TAGNIN, Stella E. O**.** A Linguística de Corpus na e para a Tradução. In: TAGNIN, Stella E.O.; VIANA, Vander (Org.). **Corpora na Tradução**. São Paulo: Hub Editorial, 2015. p. 19-56.

[139] As transcrições dos depoimentos podem ser encontradas no site da própria CNV: http://cnv.memoriasreveladas.gov.br/todos-volume-1.html Acesso em: 08 de setembro de 2023.

Discussão de dados: negação do estupro

No corpus, muitas das depoentes que negam estupro nas sessões de tortura fazem uso da dupla, ou até mesmo tripla, negação. Por exemplo, Karen Leslie Raborg Sage Keilt ao ser questionada por Mezarobba, da CNV, se ela havia sido estuprada na cela onde aconteciam as torturas, responde: "*Não, não*". Já, Dagmar Pereira da Silva faz uso da tripla negação ao ser questionada por um interlocutor não identificado da CNV se houve abuso sexual durante as torturas, respondendo: "*Não, não, não*". Ana Maria Ramos Estêvão, por sua vez, explica que foi ameaçada de estupro, mas reitera que "*não* chegaram a cumprir, *não*." Ainda, Leslie Denise Beloque, menciona que havia diferenças na intensidade das torturas a depender da equipe responsável, "mas *nunc*a, por exemplo, *nenhuma* tentativa de assédio sexual, insinuações ou ameaças de estupro" (grifos da autoria).

A dupla negação poderia estar relacionada tanto à recusa de ingresso no tema abuso sexual quanto à ênfase da negação do abuso sexual em si, com o propósito de evidenciar que a violência sexual de fato não ocorreu. Nos dois casos, a escolha das depoentes pode estar relacionada à culpabilização das vítimas de estupro na sociedade, que as estigmatiza e as coloca como merecedoras dessa violência devido à sua recusa à performance de gênero esperada pelo patriarcado (BUTLER, 2018). Essa formação discursiva da mulher violentada como merecedora da violência aparece muito claramente na fala de Lúcia Maria Sálvia Coelho quando explica por que acredita não ter sido violentada em suas sessões de tortura:

Essa parte sexual não me fizeram, porque eu estava em tamanho pânico. Mas eu acho que eu devia, no começo, estar com uma cara muito realmente do tipo que me criaram, de professora séria.

Se, de acordo com Orlandi (1999) "há sempre no dizer um não-dizer necessário", não é difícil descortinar o não dito no discurso da depoente. Aqui, Lúcia Maria performa gênero em seu discurso de acordo com o esperado pela ordem patriarcal: a mulher que tem cara de "professora séria" não é passível de violência sexual. Mas, se a mulher com cara de "professora séria" não sofre violência sexual, quem são as outras que sofrem? Segundo Maria Dalva Leite de Castro de Bonet, as putas. A militante explica que os militares buscavam convencê-la de que algumas mulheres, por serem putas, não ligavam para a violência sexual que sofriam. Sua fala começa sendo interrompida pela voz do próprio torturador, na forma de discurso direto, enfatizando que a opinião é de uma terceira pessoa, não a dela:

Tem mulher que chega aqui nem se liga pra isso" pra tentar formular na tua cabeça que você é especial. E elas são as putas e você...na dor acredita em qualquer coisa.

Nessa formação discursiva ecoa a crença de que a mulher que não se desvia das regras de comportamento social impostas a ela pelo patriarcado tem menos chance de ser vítima de violência sexual. Ou seja, aquelas mulheres que

> foram violentadas ou estupradas teriam dado alguma razão para que o crime ocorresse. Dessa forma, associada à negação da violência sexual, temos a justificativa da violência sexual. Em muitos casos, a mulher está consciente do crime sexual que sofreu durante o encarceramento, mas, após a liberdade, se silencia, pois acredita tê-lo merecido. Esse ato de silenciar-se não é em si uma escolha, mas uma resposta aos mecanismos de culpa e vergonha que foram inculcados nessas mulheres por meio de discursos patriarcais que estabelecem quem sofreu o estupro e por que o sofreu. A justificativa da violência sexual, nesse sentido, serve como ferramenta de manutenção da ordem social e da dominação (FOUCAULT, 1975), consolidando ainda mais o poder patriarcal sobre corpos e subjetividades.

Ainda nessa direção, era comum que, durante a tortura sexual, os militares buscassem induzir a vítima a acreditar que estava tendo prazer com eles, e que, se não colaborasse, seus companheiros ficariam sabendo do ocorrido. É o que relata Ana de Miranda Batista à CNV:

[...] a tortura era a seguinte, também, além de todas as outras: "Você sabe onde você está?" Voz bem cava, "Você sabe onde você está?", "Você está tendo prazer com o seu torturador?" E começava a bolinar o teu corpo todo. "E você sabe que o que seus companheiros vão dizer, que você gozou com um torturador?", "Você não vai poder sair da prisão, você vai ter que ficar do nosso lado porque se não nós vamos contar para os seus companheiros o que você fez aqui".

De acordo com TEGA (2019), essa estratégia contribui para a desorganização da mulher torturada e prejudica o trabalho de resolução do trauma. A pergunta “Você está tendo prazer com o seu torturador?” pode levar à mulher torturada a questionar se houve ou não consentimento na violência sofrida. Isso ocorre de tal maneira que muitas mulheres ainda acreditam que houve consentimento na violência sexual a qual foram submetidas, como explica Miriam Lewin (2013) a respeito de suas companheiras presas na Escola Superior de Mecânica da Armada (ESMA), em Buenos Aires. Ainda, ao ameaçar contar aos companheiros de Ana que ela teria “gozado” com seu torturador, o torturador faz de seu corpo um espaço não apenas de submissão, mas também de silenciamento, uma vez que o medo da estigmatização por parte da vítima é condição recorrente em crimes de estupro, como discutido acima. É importante comentar que a estratégia discursiva de Ana de trazer para seu relato a voz do torturador por meio da citação direta descortina, como afirma Seligmann-Silva (2008), a memória como um misto de verbalidade e imagens. A imagem do torturador do passado interrompendo o discurso de Ana no presente em primeira pessoa indica a atemporalidade da situação traumática. Isso se dá porque, segundo Levi (1990), o trauma é a memória de um passado que não passa.

Considerações finais

Voltemos ao convite de Regina Duarte em nossa epígrafe: “vamos ficar vivos”. Sem memória, verdade ou justiça, vivos sim, mas em quais condições? Quando nos é negado o direito de olhar para os erros do passado, ou quando esses erros são minimizados, a possibilidade de transformação é anulada.

Os relatos que estão sendo analisados mostram que as sessões de tortura são um microcosmo do patriarcado. Os donos da vida (SEGATO, 2022) irão disciplinar os corpos femininos, buscando torná-los submissos e “dóceis” (FOUCALT, 1975) para que ocupem o lugar social relegado às mulheres dentro do patriarcado: o lugar do espaço privado, procriando, voltadas para o cuidado familiar e subalternizadas. Para Kehl (2010), tornar públicas as lutas que foram esquecidas é primordial “na elaboração de traumas sociais” (p. 128), afinal, aquilo que não somos capazes de elaborar, tendemos a repetir. A violência institucionalizada no Brasil contra as mulheres é um exemplo disso.

Referências bibliográficas

BUTLER, Judith. **Problemas de gênero**: feminismo e subversão da identidade. Rio de Janeiro: Civilização Brasileira, 2018. 303 p.

ESTEVEZ, Alejandra (org.). **Lembrar é agir**: memória, verdade e direitos humanos. São Paulo: Letra e Voz, 2021. p. 53-68.

FOCAULT, Michel. **A ordem do discurso**: aula inaugural no college de france, pronunciada em 2 de dezembro de 1970. 24. ed. São Paulo: Edições Loyola, 1970. 74 p. Edição 2014, tradução de Laura Fraga de Almeida Sampaio.

FOCAULT, Michel. **Vigiar e punir**: nascimento da prisão. 16. ed. Petrópolis: Vozes, 1975. Edição de 2014. Tradução de Raquel Ramalhete.

KEHL, Maria Rita. Tortura e sintoma social. In: TELES, Edson; SAFATLE, Vladimir (org.). **O que resta da ditadura**. São Paulo: Boitempo, 2010. p. 123 - 132.

LEVI, Primo. **Os afogados e os sobreviventes**: os delitos, os castigos, as penas. São Paulo: Paz & Terra, 1990. 168 p.

TEGA, Danielle. **Tempos de dizer, tempos de escutar**: testemunhos de mulheres no Brasil e na Argentina. São Paulo: Intermeios, 2019. 271 p.

PEDRETTI, Lucas. Entre políticas de memória e camadas de esquecimento. In: VIÑAR, Maren e Marcello. **Exílio e Tortura**. São Paulo: Escuta, 1992. 154 p.
SEGATO, Rita. **Cenas de um pensamento incômodo**: gênero, cárcere e cultura em uma visada decolonial. Rio de Janeiro: Bazar do Tempo, 2022. 256 p. Tradução de Ayelén Medail.

SELIGMANN-SILVA, Márcio (org.). **O Espaço Biográfico**: catástrofe e representação. [S.l]: Editora Escuta, 2008. 264 p.

ANÁLISE MULTIDIMENSIONAL LEXICAL EM CORPORA DE RESENHAS E VIDEORRESENHAS *ONLINE*: UMA ABORDAGEM DA LINGUÍSTICA DE CORPUS COMO ÁREA AUTÔNOMA DE PESQUISA CIENTÍFICA

Mauricio José Ferreira LOPES[140]

Resumo: Este estudo analisa variações léxico-discursivas em resenhas escritas e videorresenhas de influenciadores literários no Instagram e YouTube. A Linguística de Corpus (LC) e a Análise Multidimensional (AMD) foram usadas para identificar padrões linguísticos (Biber, 1988; Berber Sardinha, 2000). A Análise do Discurso (AD) de Pêcheux auxiliou na interpretação das práticas discursivas ideológicas e sociais (Pêcheux, 2010). A combinação de LC com Inteligência Artificial (IA) ampliou as possibilidades de análise, especialmente em relação à formação de comunidades digitais (Silva, 2019).

Palavras-chave: Linguística de corpus; análise multidimensional; práticas discursivas; redes sociais; resenhas literárias.

Introdução

O estudo investiga as variações léxico-discursivas em corpora de resenhas escritas e videorresenhas literárias, produzidas por cinco influenciadores digitais literários (IDLs) cujos perfis e canais são hospedados nas redes sociais Instagram e YouTube. A pesquisa posiciona a Linguística de Corpus (LC) como uma área de investigação científica autônoma, utilizando a Análise Multidimensional (AMD) para identificar padrões linguísticos em registros distintos, com base nos métodos de Biber (1988) e Berber Sardinha (2000). Para além da análise quantitativa oferecida pela AMD, a Análise do Discurso (AD) de Pêcheux é incorporada ao estudo, a fim de interpretar as práticas discursivas, observando-se suas formações ideológicas e sociais, conforme abordado por Pêcheux (2010). A intersecção entre LC e AD permite uma análise integrada dos registros, revelando como as práticas discursivas refletem contextos sociais e como as dimensões discursivas emergentes mostram formações ideológicas subjacentes aos discursos dos influenciadores. Além disso, o uso de técnicas de Inteligência Artificial (IA) possibilita uma análise mais sofisticada de grandes volumes de dados linguísticos, oferecendo novas oportunidades para investigar os discursos produzidos em plataformas digitais, como observado por Silva (2019).

O estudo examina como influenciadores literários configuram suas práticas discursivas de acordo com o público-alvo e as características das diferentes plataformas. Resenhas publicadas no Instagram tendem a apresentar uma abordagem mais introspectiva e analítica, enquanto as videorresenhas no YouTube enfatizam a comunicação direta e a interação com o público. A combinação entre LC e IA permite uma análise mais precisa de gêneros e subgêneros literários, oferecendo insights valiosos sobre as dinâmicas discursivas em plataformas digitais. A pesquisa destaca o papel fundamental da LC como ciência autônoma e interdisciplinar, que fornece uma compreensão

[140] Professor de Língua Estrangeira na rede pública municipal de São Paulo. Mestre e doutorando em Linguística Aplicada e Estudos da Linguagem pela PUC-SP, bolsista CAPES.

crítica das práticas discursivas contemporâneas e suas implicações sociais e ideológicas. O estudo, assim, contribui para a consolidação da LC como uma área científica que dialoga com outras disciplinas, em função de sua natureza transdisciplinar, ampliando o escopo da análise linguística em contextos digitais e colaborando para a formação de comunidades discursivas online e a disseminação do conhecimento literário.

Questões de Pesquisa

Como se caracterizam os discursos subjacentes às dimensões discursivas que emergem da análise fatorial dos corpora de resenhas e videorresenhas literárias?

De que forma a Linguística de Corpus, integrada com a Inteligência Artificial, pode ser reconhecida como uma área autônoma de pesquisa científica, além de uma mera abordagem metodológica?

Quais são as tendências e preferências literárias dos IDLs no Instagram e no YouTube, e como essas escolhas refletem as interações culturais e sociais das comunidades de leitores?

Objetivos

Analisar e interpretar as dimensões discursivas emergentes das variáveis fatoriais nos corpora CLIRI e CLIVY.

Argumentar a favor da LC como área autônoma de pesquisa científica, explorando sua integração com técnicas de IA para análise de grandes volumes de dados linguísticos.

Fundamentação Teórica

A LC, desde o seu início com o desenvolvimento do Brown Corpus nos anos 1960, evoluiu de uma metodologia de análise de linguagem baseada em corpora para uma área de pesquisa com princípios teórico-epistemológicos robustos. Biber (1988) introduziu a Análise Multidimensional (AMD), que permite identificar dimensões de variação linguística em corpora, utilizando análise fatorial para mapear padrões coocorrentes de características léxico-discursivas. Esta abordagem é fundamental para compreender como registros distintos, como resenhas e videorresenhas, se diferenciam em função de suas características discursivas.

A LC fornece uma base empírica para identificar padrões de uso linguístico, enquanto a AD corrobora as dimensões discursivas subjacentes a esses padrões à luz das práticas sociais e formações ideológicas. Ambas reconhecem a dimensão social e ideológica da linguagem, destacando como as práticas discursivas refletem e moldam as estruturas sociais. A integração dessas duas abordagens permite uma análise mais rica e multifacetada dos textos e discursos, proporcionando uma compreensão mais profunda e abrangente de como a linguagem funciona como um fenômeno social e ideológico nas redes sociais (BIBER, 1995; PÊCHEUX, 2010).

Ademais, a integração com a AD, particularmente a de linha francesa, enriquece as interpretações ao fornecer um quadro teórico para entender os significados subjacentes aos padrões identificados. Essa combinação permite

uma análise mais holística, profunda e abrangente dos dados, considerando tanto a quantificação linguística quanto a interpretação dos contextos sociais e culturais em que esses discursos estão inseridos.

Metodologia

Design dos Corpora

Foram criados dois corpora:

- CLIRI (Corpus de Resenhas do Instagram): Composto por 517 resenhas de IDLs, abrangendo um total de 144.358 palavras.
- CLIVY (Corpus de Videorresenhas do YouTube): Composto por 504 videorresenhas, transcritas automaticamente, totalizando 1.252.978 palavras.

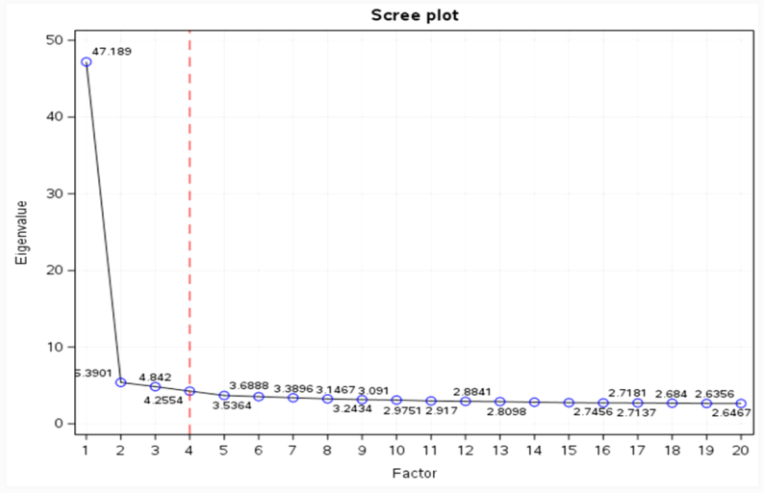
Procedimentos Metodológicos

Coleta de Dados: As resenhas foram coletadas de perfis e canais de IDLs no Instagram e YouTube, focando em influenciadores com alto engajamento e produção crítica.

Processamento dos Dados: As videorresenhas foram transcritas usando software de reconhecimento de fala, seguido de uma normalização linguística para remover ruídos textuais.

A análise fatorial identificou quatro fatores principais com base no critério de índice de variância (*eigenvalues*) acima de 1. A linha vermelha pontilhada no gráfico a seguir marca a posição do ponto de corte, indicando que os fatores além do quarto têm pouca relevância para a explicação da variabilidade total dos dados.

Gráfico 1: *Eigenvalues* ou índice de variância associado ao respectivo fator extraído da análise processada pelo Fonte SAS *University Edition*.



Os fatores com *eigenvalues* mais altos (à esquerda do ponto de inflexão) são os mais significativos, explicando a maior parte da variância

nos dados. Neste gráfico, os primeiros quatro fatores apresentam *eigenvalues* substancialmente maiores, indicando que são os mais relevantes para descrever as variações léxico-discursivas nos corpora analisados. A linha vermelha pontilhada sugere que os primeiros quatro fatores devem ser retidos para uma interpretação mais eficaz, pois representam a maior quantidade de variação explicada nos dados.

Dimensões Discursivas: Aplicação da AMD lexical para identificar padrões lexicais e discursivos. Quatro dimensões principais foram identificadas:

Dimensão 1: Comunicação e Interatividade vs. Introspecção e Análise. Dimensão 2: Motivação e Engajamento Cultural vs. Reflexão e Realismo Existencial. Dimensão 3: Análise Crítica e Psicológica vs. Contextualização Histórica e Descritiva.

Dimensão 4: Análise Pontual da Narrativa vs. Foco na Comercialização do Livro.

Interpretação dos Fatores: As dimensões foram interpretadas discursivamente, capturando as propriedades comunicativas primordiais detectadas nos registros pela AMD lexical.

Discussão dos Dados

Os resultados indicam que há variações significativas nas abordagens de resenhas e videorresenhas literárias:

Dimensão 1: Videorresenhas focam na comunicação direta e criação de uma comunidade, enquanto as resenhas do Instagram são mais introspectivas e analíticas.

Dimensão 2: Videorresenhas abordam aspectos culturais e de engajamento, enquanto as resenhas escritas exploram reflexões existenciais e sociais.

Dimensão 3: Resenhas no Instagram enfatizam a análise crítica e psicológica, enquanto as videorresenhas priorizam a contextualização histórica e descritiva das obras.

Dimensão 4: Resenhas escritas tendem a ser mais detalhadas na análise narrativa, enquanto videorresenhas focam em aspectos comerciais e logísticos da leitura.

Considerações Finais

O estudo revelou que as resenhas no Instagram tendem a adotar uma abordagem analítica e reflexiva, explorando aspectos individuais e críticos das obras literárias. Em contrapartida, as videorresenhas no YouTube priorizam a interação e o engajamento, com foco na narrativa e na comunicação com a audiência. Essa diferenciação reforça a ideia de que a profundidade intelectual e a acessibilidade não são mutuamente excludentes no domínio discursivo das redes sociais, mas podem coexistir de maneira complementar. A análise demonstrou que as redes sociais, apesar de sua natureza fragmentada e dinâmica, têm o potencial de disseminar conhecimento literário e promover debates críticos.

O estudo também sustenta a importância de reconhecer a LC como uma área científica autônoma. A abordagem da LC permite explorar criticamente a linguagem em seu contexto de uso, superando a visão limitada de que a pesquisa linguística se restringe a aspectos técnicos ou metodológicos. Assim, ao posicionar a LC como campo independente, os pesquisadores têm a oportunidade de desenvolver estudos transdisciplinares, contribuindo para uma compreensão mais profunda da linguagem e de suas manifestações na sociedade contemporânea.

Ao reafirmar a LC como área autônoma, este estudo científico destaca a importância de uma abordagem crítica e reflexiva na condução dos estudos linguísticos, evidenciando a capacidade da LC de dialogar com diversas disciplinas e enriquecer o entendimento sobre as dinâmicas discursivas nas redes sociais. Em suma, o estudo demonstrou que a LC, além de fornecer insights valiosos sobre práticas discursivas específicas, possui um papel fundamental na construção de uma teoria crítica da linguagem, consolidando-se como uma área científica capaz de produzir conhecimento inovador e relevante para o estudo da linguagem em diferentes contextos.

Essa abordagem multidisciplinar contribui para uma compreensão mais aprofundada dos discursos literários em redes sociais, mostrando como influenciadores digitais configuram suas práticas discursivas para diferentes audiências e situações comunicacionais e interativas no domínio discursivo digital.

Referências

ARAÚJO, J.; SOUSA, M. M. N.; CAVALCANTI, J. M. Comunidade discursiva e redes sociais: os resenhadores do Skoob. *Revista Intercâmbio*, v. 45, p. 28-51, 2020. Disponível em: https://revistas.pucsp.br/index.php/intercambio/article/view/50439. Acesso em: 12 jul. 2024.

BAKER, P.; McENERY, T. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Palgrave Macmillan, 2015.

BERBER SARDINHA, T. Linguística de Corpus: Histórico e Problemática. *Revista Delta*, v. 16, n. 2, p. 45-67, 2000.

BERBER SARDINHA, T.; VEIRANO PINTO, M. *Multi-Dimensional Analysis: Research Method and Current Issues*. Bloomsbury Academic, 2019.

BIBER, D. *Variation Across Speech and Writing*. Cambridge University Press, 1988.

BIBER, D.; CONRAD, S. *Register Genre and Style*. Cambridge University Press, 2009. PÊCHEUX, M. *Semântica e Discurso*. Editora Unicamp, São Paulo, 2010.

SILVA, E. Análise do Discurso e Linguística de Corpus. *Revista de Estudos Linguísticos*, v. 37, n. 2, p. 112-134, 2019.

PEDAGOGIA DA TRADUÇÃO E OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL (ODS)

Emiliana FERNANDES BONALUMI [141]
Diva CARDOSO DE CAMARGO[142]

RESUMO: Esta comunicação a respeito da pedagogia da tradução (DÍAZ FOUCES, 1999, 2001; LAVIOSA, 2008, 2010, 2020; YYY, 2016a, 2016b; SERPA et al, 2021; THOW, 2022) teve por intuito analisar termos multi-palavras em um corpus compostos de notícias originais em línguas inglesa e portuguesa sobre os ODS, em especial, item (11) Cidades e Comunidades Sustentáveis. Também fizemos uso da aprendizagem movida por dados (JOHNS, 1991, 2002; GRANGER, 1998, 2002; BERBER SARDINHA, 2004, 2006, 2010; BOULTON, 2010; LAVIOSA, 2008, 2010, PINTO, 2021; PINTO & YYY, 2021; PINTO et al. 2022). A fim de gerar os termos analisados, utilizamos a ferramenta on-line Sketch Engine. Após examinarmos as listas dos primeiros doze termos em línguas inglesa e portuguesa, foi compilado um glossário composto das similaridades e diferenças encontradas respectivamente em quatro termos de línguas inglesa e portuguesa. Outrossim, compilamos um glossário menor constituído de cinco termos, por meio de um texto previamente selecionado em língua portuguesa a respeito do item (11) dos ODS, bem como discutimos alguns dos traços de normalização de Scott (1998), fazendo uso da versão em língua inglesa elaborada pelos discentes.

Palavras-chave: Pedagogia da Tradução, Objetivos de Desenvolvimento Sustentável (ODS), Termos Multi-palavras, Glossário, Traços de Normalização.

INTRODUÇÃO

A pedagogia da tradução aborda o emprego de teorias e práticas de tradução em seu ensino (LEONARDI, 2010; YYY, 2016a, 2016b; SERPA et al, 2021). Sendo assim, julgamos importante utilizá-la na disciplina de Prática de Tradução em Língua Inglesa III em uma das unidades da Universidade Estadual do Estado de São Paulo, uma vez que acreditamos que para efetuar uma tradução, é necessário o conhecimento de sua teoria. Já, a opção de utilizar os Objetivos de Desenvolvimento Sustentável - ODS das Nações Unidas neste trabalho se deve ao fato de que se trata de um tema de saliência desde sua adoção em setembro de 2015, e vem sendo discutido por pesquisadores em âmbito nacional e internacional.

Por seu turno, a aprendizagem movida por dados tem sido desenvolvida em sala de aula desde sua criação por Johns em 1986. É uma abordagem que

[141] Docente do Curso Letras-Inglês da UFR (Universidade Federal de Rondonópolis, MT).

[142] Docente do Departamento de Letras Modernas da UNESP, São José do Rio Preto.

utiliza textos autênticos extraídos dos corpora para diversas finalidades. Recorremo-nos à aprendizagem movida por dados a fim de elaborar as listas de frequência dos termos multi-palavras, fazendo uso da ferramenta on-line Sketch Engine (https://www.sketchengine.eu/).

Com esse propósito, compilamos um corpus jornalístico composto respectivamente de oito textos originais em língua inglesa e oito textos originais em língua portuguesa, a respeito do item (11) Cidades e Comunidades Sustentáveis dos ODS, para a investigação em corpora, cujo objetivo deste trabalho é observar a versão de cinco termos multi-palavras, bem como discutir alguns dos traços de normalização de Scott (1998), por meio da versão de um texto previamente selecionado.

FUNDAMENTAÇÃO TEÓRICA

No que tange à pedagogia da tradução, podemos mencionar que as primeiras investigações neste campo foram a de Díaz Fouces (1999, 2001), nas quais "buscam criar metodologias de ensino que observem as competências da tradução". Por meio desses métodos, "discentes devem estar aptos a codificar e sistematizar informações presentes nos textos" (DÍAZ FOUCES, 1999, 2001 apud SERPA et al (2021). Por seu turno, Laviosa (2008) comenta que "os corpora pequenos e especializados são feitos e usados não apenas como recursos de busca de equivalentes na tradução, mas também como repositórios de dados a fim de aperfeiçoar a compreensão dos discentes a respeito das regularidades da tradução" (LAVIOSA, 2008 apud YYY 2016b, p. 159). No trabalho intitulado "A Corpus-based Proposal for Teaching a Translational Habitus: Initial dialogues with Bourdieu's sociological approaches", Serpa et al (2021) apresentam atividades didáticas utilizando a aprendizagem movida por dados e a pedagogia da tradução. Em 2022, Thow publica o estudo "Translation Pedagogy in the Comparative Literature Classroom: Close Reading and the Hermeneutic Model of Translation" nos mostrando como podemos utilizar a pedagogia da tradução em uma sala de aula de literatura comparada, nos indicando meios para realizá-la.

Acerca dos ODS das Nações Unidas, sabemos da importância desse tema atualmente e por este motivo, além das investigações de Pinto (2021) e Pinto et al (2023) e do Seminário de Estudos Linguísticos da UNESP, realizado em setembro de 2023, com o tema "A Linguística face aos Objetivos de Desenvolvimento Sustentável da Agenda 2030", decidimos por também trabalhá-lo em sala de aula, acreditando que seja de extrema relevância apresentar ao graduando uma variedade de temas que possam vir a enriquecer sua formação. Já, a aprendizagem movida por dados foi criada por Tim Johns e é, de acordo com Berber Sardinha,

> uma das propostas mais sólidas para a utilização de material de corpus na sala de aula. [...] A ênfase é desenvolver no aluno a habilidade de descoberta (discovery learning), e o papel do professor é propiciar meios para que os alunos adquiram estratégias de descoberta. O computador entra como elemento central da aprendizagem, no papel de informante, e não de substituto do professor (BERBER SARDINHA, 2004, p. 290- 291 – grifo nosso).

A fim de elaborar as listas de frequência dos termos multi-palavras, utilizamos a ferramenta on-line Sketch Engine, na qual se fizermos o upload de textos em línguas inglesa e portuguesa, ela nos fornecerá os termos multi-palavras. além de nos trazer linhas de concordância em seu contexto.

METODOLOGIA DE INVESTIGAÇÃO

Apresentamos, a seguir, a composição dos corpora, bem como os procedimentos e as formas de análise adotadas para o nosso estudo.

MATERIAL EMPREGADO NA COMPILAÇÃO DOS CORPORA

Em razão à limitação de espaço, o material empregado na compilação dos corpora será apresentado durante o evento.

PROCEDIMENTOS DE ANÁLISE

PASSOS PARA A ANÁLISE COM BASE NAS CONCORDÂNCIAS

Empregamos a ferramenta on-line Sketch Engine para os TOs em línguas inglesa e portuguesa. Após termos examinado as listas dos primeiros doze termos em línguas inglesa e portuguesa, foi compilado um glossário de línguas inglesa e portuguesa composto das similaridades e diferenças encontradas em quatro termos. Também, elaboramos um glossário menor constituído de um texto em língua portuguesa acerca do item (11) dos ODS e, por meio da versão em língua inglesa do referido texto realizada pelos discentes da disciplina, discutimos alguns dos traços de normalização de Scott (1998).

DISCUSSÃO E ANÁLISE DOS RESULTADOS

A seguir, apresentamos os dados extraídos do texto “Cidades portuguesas a caminho da sustentabilidade”, bem como de suas versões para a língua inglesa efetuadas por sete discentes da disciplina, além da discussão a respeito de alguns traços de normalização de Scott (1998), a saber: tamanho da sentença, reordenação de elementos, pontuação, padrão de repetição simples, mudança em palavras menos comuns, e adição no texto traduzido, .

Como podemos notar por meio do glossário menor constituído de cinco termos extraídos do referido texto, os termos multi-palavras smart city, capital verde e cidades inteligentes apresentaram apenas uma versão para a língua inglesa, respectivamente o próprio termo smart city, green capital e smart cities.

Referindo-se ao termo multi-palavra ecossistema da sociedade, verificamos que houve variações em suas versões para a língua inglesa realizada pelos alunos: Society’s ecosystem (1); ecosystem of Society (4); ecosystems of the Society (1); e societal ecosystem (1).

No tocante ao termo linha orientadora no país, percebemos que as versões para a língua inglesa também foram variadas pelos discentes: guideline in the country (5); guiding line in the country (1); e guiding direction in the country (1).

Devido à limitação de espaço, não será possível apresentar os dados com detalhes nesta seção, porém, durante a apresentação, estes serão efetivamente discutidos e analisados.

CONSIDERAÇÕES INICIAIS

Esperamos que por meio deste trabalho seja possível observar as semelhanças e diferenças nos TOs em línguas portuguesa e inglesa a respeito do léxico contido no item (11) Cidades e Comunidades Sustentáveis dos ODS, por meio da compilação de dois glossários de línguas inglesa e portuguesa.

Acreditamos que a pedagogia da tradução é uma abordagem que merece destaque, uma vez que trata da teoria e da prática da tradução, sendo muito relevante para os discentes, uma vez que por meio da teoria, podemos ampliar nosso conhecimento e transmiti-lo no exercício da versão para a língua inglesa.

REFERÊNCIAS

BAKER, M. Corpus-based translation studies: the challenges that lie ahead. In: SOMERS, H. (ed.). Terminology, LSP and translation studies in language engineering, in honour of Juan C. Sager. Amsterdam/Philadelphia: John Benjamins, 1996, p.175-186.

BERBER SARDINHA, T. Lingüística de Corpus. São Paulo: Manole, 2004.

BURNETT, S. A corpus-based study of translational English. Manchester: UMIST, 1999. Dissertação de mestrado.

YYY. Uso de Corpora para uma Pedagogia da Tradução. Revista Língua & Letras. V. 17:36, p. 188-205, 2016a.

YYY. Language of Translation and Interculturality for a Corpus-based Translation Pedagogy. In: FONTANILLE, J. (Org.). Traduire: signes, textes, pratiques. Liège: Presses Universitaires de Liège, p. 155-173, 2016b.

DÍAZ FOUCES, Ó. Didáctica de la traducción (português – español), Vigo: Servicio de Publicacións da Universidade de Vigo, 1999.

DÍAZ FOUCES, Ó. Sociologia de la traducción, Quaderns: Revista de traducció v. 6, p. 63-77, 2001.

JOHNS, T. MicroConcord: a language learner's research too I. System, Oxford, Pergamon, v.14, n.2, jun/ 1986, p.l5l-62.

LAVIOSA, S. Discovery and Justification Procedures in the Corpus-Based Translation Classroom. Translation Challenges: From Training to Profession, Hammamet, Tunisia, 28-29 November 2008.

LAVIOSA, S. A transcultural conceptual framework for corpus-based translation pedagogy, 2010 In: Proceedings of Using corpora in contrastive and translation studies - UCCTS. Ormskirk, 2010. v. 01.

LAVIOSA, S. Translation and Language Education: Pedagogic Approaches Explored. New York: Routledge/Taylor & Francis, 2014.

LAVIOSA, S. The Instrumental and Hermeneutic Models of Translation in Higher Education. In: Engel,
N., Köngeter, S. (eds) Übersetzung. Springer VS, Wiesbaden, 2020. https://doi.org/10.1007/978-3-658-
20321-
4_3

SERPA, T.; PINTO, P.T.; YYY. A Corpus-based Proposal for Teaching a Translational Habitus: Initial dialogues with Bourdieu's sociological approaches. Trans. Revista de Traductología. V. 25, p. 507-525, 2021.

SCOTT, M. N. Normalisation and Reader's Expectation: A Study of Literary Translation with Reference to Lispector's A Hora da Estrela. Tese (Doutorado em Filosofia). Liverpool: Universidade de Liverpool, 1998.

THOW, E. Translation Pedagogy in the Comparative Literature Classroom: Close Reading and the
Hermeneutic Model of Translation. L2 Journal, V. 14 (2), p. 91-106, 2022. DOI: 10.5070/L214252048

O C-ORAL-ESQ, corpus brasileiro de fala espontânea de pessoas com esquizofrenia

Bruno Nevis Rati de Melo ROCHA
Tommaso RASO

Introdução

O trabalho visa apresentar (a) o C-ORAL-ESQ (RASO et al., 2023), Corpus Oral de Esquizofrênicos, (b) algumas medidas descritivas da atual fase de compilação do corpus e (c) perspectivas futuras do corpus. O C-ORAL-ESQ, em estágio avançado de compilação, contará com 43 registros de interações entre psiquiatras e pacientes com esquizofrenia durante consultas médicas. O corpus será transcrito e segmentado segundo os critérios estabelecidos para o corpus C-ORAL-BRASIL (RASO; MELLO, 2012) e alinhado no Elan (WITTENBURG et al., 2006). Grande parte das gravações será etiquetada informacionalmente segundo os preceitos da Language into Act Theory (CRESTI, 2000). Por fim, o C-ORAL-ESQ contará com uma seção multimodal composta de 18 gravações em áudio e vídeo que seguirá os parâmetros estabelecidos para o C-ORAL-BGEST, apresentados por Barros (2021), permitindo estudos de expressões faciais e gestos.

A Language into Act Theory

A L-AcT é uma teoria que tem como objetivo principal descrever a estrutura informacional da fala espontânea. Em particular, a teoria se interessa por explicar a maneira pela qual o falante codifica prosodicamente as diferentes funções que as unidades tonais do enunciado podem assumir, partindo do entendimento de que todo enunciado é um ato de fala (AUSTIN, 1962). Os enunciados são entendidos como a menor unidade superior ao nível da palavra que possui autonomia prosódica e interpretabilidade pragmática (ou seja, é percebido como prosodicamente concluído e veicula uma ilocução). Um enunciado pode ser formado por uma ou mais unidades tonais, as quais desempenham funções diversas. A função básica, expressa por ao menos uma unidade em todo e qualquer enunciado, é justamente a de realizar uma ilocução. Assim, em enunciados compostos por uma única unidade tonal, essa é necessariamente sua unidade ilocucionária. Além da função ilocucionária, existem outros tipos de funções que uma unidade pode desempenhar, como aquela de estabelecer um domínio cognitivo para a ilocução (cf. CRESTI, 2000 e RASO; MONEGLIA, 2014 para uma descrição detalhada de todas as funções).

Metodologia

O setting das gravações

As gravações do C-ORAL-ESQ são realizadas em instituições hospitalares públicas de Belo Horizonte. De 2019 a 2023, foram feitas gravações em áudio de consultas ocorridas no ambulatório do Instituto Raul Soares (IRS/FHEMIG). A partir de 2023, começaram a ser feitas gravações em áudio e vídeo seja no IRS, seja no Hospital das Clínicas da UFMG (HC/EBSERH).

As gravações são conduzidas por uma equipe de dois pesquisadores que se dirigem ao hospital com os equipamentos (dois conjuntos de microfones sem fio omnidirecionais Sennheiser EK100/SK100, um gravador digital de alta resolução Tascam DR-100 e duas câmeras GoPro Hero7). A equipe se encontra com o residente responsável pela consulta a ser gravada, o qual conversou antecipadamente com o paciente que atenderá naquele dia sobre a possibilidade de ser gravado. Em seguida, a equipe coloca os microfones no residente e no paciente e posiciona as câmeras de vídeo. As duas câmeras são usadas para gravar imagens do paciente, sendo uma posicionada frontalmente a ele e outra posicionada de forma levemente lateral, para permitir observar os gestos com maior profundidade. Antes de sair do consultório, a equipe lê os TCLEs do médico, do paciente e de eventuais acompanhantes e recolhe as assinaturas necessárias. Fora do consultório, a equipe conecta os microfones sem fio ao gravador e inicia a gravação. Durante toda a consulta, a equipe permanece fora do consultório e não interage direta ou indiretamente com os participantes.

Esse desenho metodológico foi feito para potencializar a naturalidade da situação, com o objetivo de garantir a validade ecológica dos dados.

As etapas de tratamento dos dados

Todas as gravações d o C-ORAL-ESQ passam por uma série de etapas obrigatórias de tratamento dos dados: (a) transcrição e segmentação prosódica segundo os critérios adotados pelo C-ORAL-BRASIL (RASO; MELLO, 2012), (b) primeira e segunda revisões da transcrição e da segmentação, (c) alinhamento texto-som-espectrograma no *software* Elan e (d) revisão final da transcrição. As consultas que possuem gravações em vídeo possuem procedimentos adicionais, sendo eles (a) a anonimização da face do paciente e (b) a inserção dos vídeos no alinhamento do Elan. A anonimização é feita por meio de um procedimento computadorizado que captura pontos do rosto importantes para a veiculação das expressões faciais e os reproduz em um rosto criado por inteligência artificial. Essa versão anonimizada será disponibilizada com o corpus e poderá ser usada em apresentações e publicações científicas sem permitir que se descubra a identidade do participante

Por fim, um conjunto de gravações do C-ORAL-ESQ será selecionado para passar pelo procedimento de etiquetagem informacional segundo a L-AcT.

O corpus

Estado atual

O C-ORAL-ESQ irá registrar 43 consultas entre pacientes com esquizofrenia e seus psiquiatras, com uma média de 1.180 palavras produzidas pelo paciente em cada gravação (além das palavras dos médicos e de eventuais acompanhantes dos pacientes).

A Tabela 1 mostra o número de palavras produzidas por cada grupo de participantes (pacientes, profissionais da saúde, acompanhantes e outros). Nela, lê-se que as gravações têm, em média, 2.360 palavras, sendo que a menor possui 749 palavras e a mais extensa, 4.868, totalizando 101.606 palavras. Os

pacientes produzem em média 1.180 palavras por gravação, mas com uma variação muito grande (DP 810), indo de 236 palavras a 2.717.

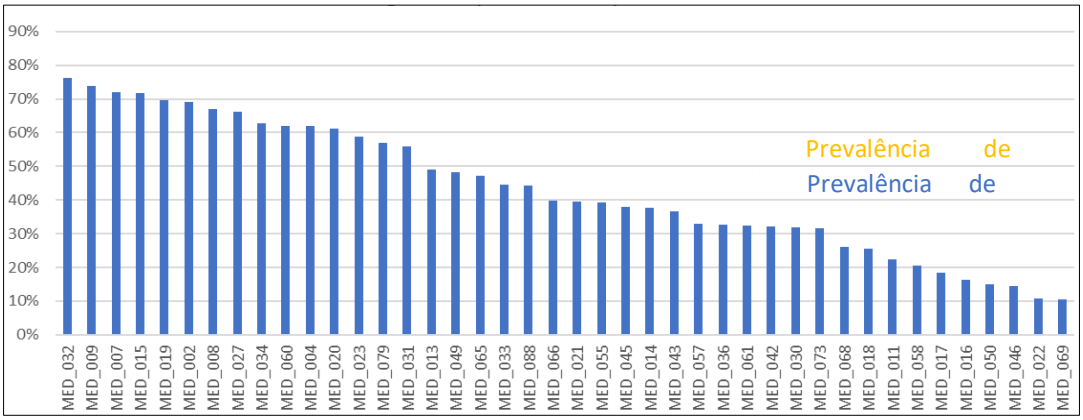
Tabela 1 – Medidas descritivas do corpus

		Palavras	Média	DP	Mediana	Mínimo	Máximo
Pacientes		48.086	1.118	810	836	236	2717
	Médicos	46.522	1.082	475	974	260	2157
Acompanhantes		6.642	332	300	265	10	801
	Outros	3561	57	67	17	2	172
	Total	101.606	2.360	1.031	2.237	749	4.868

Fonte: os autores.

O alto desvio padrão no número de palavras produzidas por todos os grupos de participantes, em especial o de pacientes, é o reflexo de características primordiais das consultas: a variação de tipos e de número de assuntos a serem tratados no dia, a disponibilidade do paciente para interagir com o psiquiatra, a estratégia adotada pelo psiquiatra para lidar com o paciente no dia, a presença ou ausência de acompanhantes e o grau de participação deles nas consultas etc. Com efeito, o alto desvio padrão no número de palavras dos participantes é tanto uma característica esperada quanto desejada em um corpus que busca documentar a fala espontânea produzida em interações médico-paciente não roteirizadas.

Figura 1 – Percentual de palavras de pacientes com relação às palavras de outros participantes

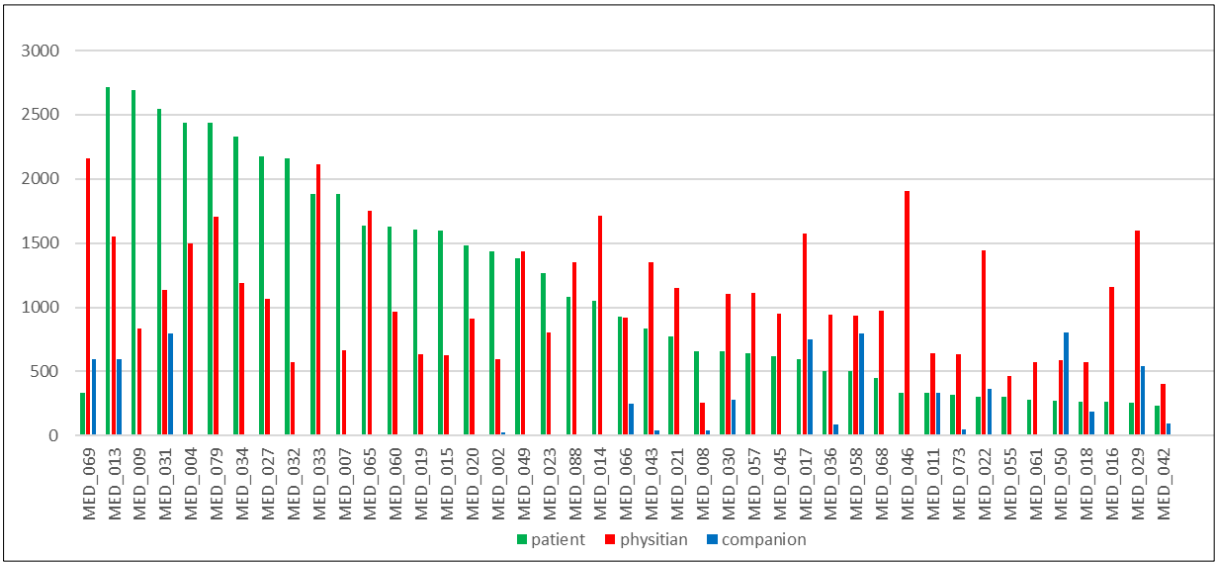


Fonte: Elaboração própria.

A figura 1 mostra o percentual de palavras de pacientes com relação às palavras de outros participantes (psiquiatras, psicólogos, acompanhantes e demais pessoas) em cada gravação. As 15 primeiras colunas, cujo limite superior vai além da linha horizontal tracejada, identificam gravações em que a fala do paciente é predominante na consulta, chegando a ocupar 79% de uma gravação no caso mais extremo. As demais são casos em que o paciente produz um número de palavras inferior ao dos demais participantes, chegando ao caso de uma consulta em que o paciente produz aproximadamente 11% do total de palavras. É interessante notar que existe uma grande variação com relação a esse parâmetro, podendo ser observada uma tendência aproximadamente linear na distribuição dos valores percentuais de palavras dos pacientes com relação às palavras dos outros participantes.

A figura 2, por sua vez, indica o número de palavras por participante em cada gravação (destacando palavras de pacientes em verde, de médicos em vermelho e de acompanhantes em azul). Vale notar que, nas gravações com o menor número de palavras de pacientes (últimas colunas à direita), é maior a quantidade de consultas em que os acompanhantes produzem uma quantidade expressiva de palavras.

Figura 1 – Número de palavras de pacientes (verde), médicos (vermelho) e acompanhantes (azul)



Fonte: Elaboração própria.

Perspectivas futuras

Paralelamente à compilação do C-ORAL-ESQ, está sendo elaborado um corpus de controle específico para esse corpus. O corpus de controle possuirá a mesma quantidade de gravações do C-ORAL-ESQ e será balanceado em gênero, idade e escolaridade. Suas gravações serão multimodais (áudio e vídeo) e registrarão consultas médicas de pacientes com doenças crônicas e sem histórico pessoal e familiar de transtornos mentais. As gravações do corpus de controle passarão pelos mesmos procedimentos metodológicos aplicados ao C-ORAL-ESQ.

Conclusões

O corpus C-ORAL-ESQ fornecerá um material de grande ineditismo não somente no panorama nacional, mas também internacional. Grande parte das pesquisas sobre a fala de pessoas com esquizofrenia é feita a partir do exame de fala eliciada. O C-ORAL-ESQ, por outro lado, apresenta um grande conjunto de dados de fala espontânea obtidos em interações reais, não roteirizadas. O corpus já tem sido usado em uma série de pesquisas preliminares que tem permitido observar com mais atenção a estrutura informacional da fala dos pacientes (COSTA JR., 2022) e pode ser usado tanto para pesquisas de cunho linguístico como médico.

Referências bibliográficas

AUSTIN, J. L. *How to do things with words*. Oxford University Press, Oxford 1962.

BARROS, C. A. *A relação entre unidades gestuais e quebras prosódicas: o caso da unidade informacional Parentético*. Dissertação (Mestrado em Estudos Linguísticos), Universidade Federal de Minas Gerais, 2021.

CAVALCANTE, F. A., *The topic unit in spontaneous american English: a corpus-based study*. Dissertação (Mestrado em Estudos Linguísticos), Universidade Federal de Minas Gerais, 2016.

COSTA JR. J. C., *Padrão informacional de stanzas de pacientes com esquizofrenia*. Tese (Doutorado em Linguística), Universidade Federal de Minas Gerais, 2022.

CRESTI, E., *Corpus di Italiano parlato*, Accademia della Crusca, Firenze 2000.

MONEGLIA, M.; RASO, T. Notes on language into act theory. In: RASO, T.; MELLO, H. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins, 2014, 468-495.

RASO, T.; MELLO, H. *C-ORAL-BRASIL I*. Corpus de referência do português falado informal. Belo Horizonte: UFMG, 2012.

WITTENBURG, P.; BRUGMAN, H.; RUSSEL, A.; KLASSMANN, A.; SLOETJES, H. ELAN: a Professional Framework for Multimodality Research. Em: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation, 2006.

Agradecimentos

Nossa sincera gratidão à presença, contribuição e empenho de todes.

Foi o esforço colaborativo que permitiu a realização deste evento.

Comissão Organizadora

UNIVERSIDADE DE BRASÍLIA

ANAIS ELETRÔNICOS DO XVI ENCONTRO DE LINGUÍSTICA DE CORPUS E DA XII ESCOLA BRASILEIRA DE LINGUÍSTICA COMPUTACIONAL

BRASÍLIA-DF

2024

EXPEDIENTE

Comissão organizadora

Elisa Duarte Teixeira (Presidenta - UnB)

Andréa Geroldo dos Santos (IBFE-SP)

Célia Maria Magahães (UFMG / UnB)

Cláudio Corrêa e Castro Gonçalves (UnB)

Elaine Alves Trindade (PUC-SP)

Flávia Cristina Cruz Lamberti Arraes (UnB)

Joacyr Tupinambás (Unicamp)

Nilson Roberto Barros da Silva (UERN)

Patrícia Tuxi dos Santos (UnB)

Rafaela Araújo Jordão Rigaud Peixoto (DECEA-FAB)

Rodrigo Garcia Rosa (USP)

Rozane Rodrigues Rebechi (UFRGS)

Stella Esther Ortweiler Tagnin (USP)

Thiago Blanch Pires (UnB)

Vander Paula Viana (U. Edinburgh)

Comissão científica

Adriana Zavaglia (USP)

Adriane Orenha Ottaiano (UNESP)

Alessandra Matias Querido (UnB)

Ana Eliza Pereira Bocorny (UFRGS)

Andréa Geroldo dos Santos (IBFE-SP)

Angela Maria Tenório Zucchi (USP)

Ariel Novodvorski (UFU)

Camila Höfling (UFSCar)

Celia Maria Magalhães (UFMG)

Claudia Zavaglia (UNESP-IBILCE)

Cleci Regina Bevilacqua (UFRGS)

CAPES

POSTRAD

COMPLETT UnB
Corpus Multilíngue para Pesquisas em Línguas Estrangeiras, Tradução e Terminologia da Universidade de Brasília

TermiTraDiCo
Terminologia e Tradução Direcionadas por Corpus

ABraLC
Associação Brasileira de Linguística de Corpus

Cristiane Krause Kilian (Inst. Sup. de Ed. Ivoti - ISEI)

Deise Prina Dutra (UFMG)

Elaine Alves Trindade (PUC-SP)

Gleiton Malta (UFBA)

Guilherme Fromm (UFU)

Heliana Ribeiro de Mello (UFMG)

Heloísa Orsi Koch Delgado (La Salle / UFRGS)

Igor Antônio Lourenço da Silva (UFU / UFMG)

Joacyr Tupinambás de Oliveira (UNICAMP)

Luciana Carvalho Fonseca (USP)

Luciana Latarini Ginezi (Tradutora / Intérprete)

Luciane Leipnitz (UFPel)

Malila Carvalho de Almeida Prado (BNU-HKBU)

Marcos de Campos Carneiro (UnB)

Maria José Bocorny Finatto (UFRGS)

Nilson Roberto Barros da Silva (UERN)

Patrícia Tosqui Lucks (ICEA)

Paula Tavares Pinto (UNESP-IBILCE)

Rafaela Araújo Jordão Rigaud Peixoto (DECEA / USP)

Renato Rodrigues Pereira (UFMS)

Rodrigo Garcia Rosa (USP)

Rozane Rodrigues Rebechi (UFRGS)

Sandra María Pérez López (UnB)

Shirlene Bemfica de Oliveira (IFMG - Ouro Preto)

Simone Sarmento (UFRGS)

Stella Esther Ortweiler Tagnin (USP)

Thiago Alexandre Salgueiro Pardo (USP)

Vander Viana (University of Edinburgh)

APRESENTAÇÃO

A Linguística de Corpus (LC), ciência que estuda a linguagem por meio da análise de grandes quantidades de textos em formato eletrônico organizados na forma de corpora, avançou muito no mundo e também no Brasil desde os primeiros trabalhos publicados, que foram surgindo com a introdução do computador no ambiente acadêmico, na década de 1960. De lá para cá, é cada vez maior o número de campos do conhecimento que se valem da LC em suas pesquisas, seja como abordagem, seja como metodologia – inclusive para além das áreas de Letras e Linguística.

A Linguística Computacional, ou Processamento de Linguagem Natural (PLN), é um campo de estudos multidisciplinar que aplica a ciência da computação à análise e compreensão da linguagem humana. Os primeiros trabalhos publicados datam da década de 1950. Mas, assim como a LC, o PLN teve um crescimento exponencial nas últimas décadas, à medida que o uso da tecnologia foi evoluindo e se fazendo presente em praticamente todas as esferas da experiência humana.

O Encontro de Linguística de Corpus (ELC) teve sua primeira edição em 1999, na Universidade de São Paulo - em 2024, comemoramos 25 anos de sua criação! A Escola Brasileira de Linguística Computacional (EBRALC) surgiu alguns anos depois, em 2007, à medida que participantes do ELC constataram uma necessidade premente de ampliar os conhecimentos computacionais e tecnológicos de pesquisadora(e)s brasileira(o)s trabalhando com Linguística de Corpus nas áreas de Humanas. Assim, o objetivo da EBRALC é sempre oferecer diversas oficinas práticas, para as quais a(o)s interessada(o)s deverão se inscrever oportunamente. O ELC, por sua vez, é um evento científico que tem por tradição não possuir sessões paralelas, para que toda(o)s possam assistir a todas as comunicações orais, proporcionando um convívio mais profícuo entre a(o)s participantes das várias áreas envolvidas – um convite ao diálogo e à colaboração.

O ELC/EBRALC 2024, cujo tema será “Linguística de Corpus e Inteligência Artificial: interfaces com Letras e Linguística”, acontecerá de 21 a 24 de outubro, na Universidade de Brasília. Organizado por docentes do Instituto de Letras membros dos grupos de pesquisa TermiTraDiCo (Terminologia e Tradução Direcionadas por Corpus) e COMPLETT (Corpus Multilíngue para Pesquisas em Línguas Estrangeiras, Tradução e Terminologia), e colegas de várias outras universidades, o evento tem o apoio do Programa de Pós-Graduação em Tradução da UnB (POSTRAD). Visa fomentar discussões sobre os impactos da Inteligência Artificial na formação acadêmica e na pesquisa nas áreas de Letras e Linguística, com foco no papel que a Linguística de Corpus tem ocupado e pode ocupar, futuramente, no estabelecimento desse diálogo transdisciplinar.

Esperamos, assim, estimular parcerias e promover uma necessária valorização dos estudos da linguagem com o auxílio do computador, no ambiente acadêmico-científico e na sociedade brasileira.

LISTA COM NOME DE AUTORES
Ordem por sobrenome

ALENCAR, Leonel Figueiredo de

BARROS, Cláudia Dias de

BATISTA, Julia de Souza

BOCORNY, Ana Eliza Pereira

BOHORQUEZ, Carolina

BRAGA, Jasper Vilan

CARDOSO DE CAMARGO, Diva

COSTA, Danilo Duarte

DELGADO, Heloísa Orsi Koch

DURAN, Magali Sanches

FEKETE, João

FERNANDES BONALUMI, Emiliana

FONSECA, Luciana Carvalho

FRANGIOTTI, Graziele Altino

FREITAG, Patrícia Helena

FURTADO, Anna Beatriz Dimas

GAMONAL, Maucha Andrade

GIL, Cristina Borges

GUEDES DE SOUZA, Bianca Mara

KAUFFMANN, Carlos Henrique

KILIAN, Cristiane Krause

KUHN, Tanara Zingano

LIU, Ziyang

LOPES, Jhonatan Henrique

LOPES, Mauricio Jose Ferreira

MAIA-PIRES, Flávia de Oliveira

MARCHESE, Giovana de Castro

MARQUES, Carolina Godoi de Faria

MOREIRA DE OLIVEIRA, Isabela

MURTA, Lucas Renato dos Santos

NUNES, Wagner da Cunha

O'CONNOR, Anne

OLIVEIRA, Shirlene Bemfica de

OLIVEIRA, Simone

ORENHA-OTTAIANO, Adriane

PAGANO, Adriana Silvina

PARDO, Thiago Alexandre Salgueiro

PINTO, Paula Tavares

PIRES, Thiago Blanche

RABELO, Iasmin Valéria Miranda

RAMOS, Camila Alves

RASO, Tommaso

RASO, Tommaso

REBECHI, Rozane R.

REBECHI, Rozane Rodrigues

ROCHA, Bruno Neves Rati de Melo

ROCHA, Bruno Nevis Rati de Melo

ROCHA, Ísis Beber de Souza Fiorilo

SANTOS, Amanda Letícia Valadares dos

SARDINHA, Tony Berber

SILVA, Bruna Rodrigues da

SILVA, Luciano Franco da

SOUSA, Jackson Wilke da Cruz

SOUSA, Luan Daniel dos Santos

SYDIO, Ursula Puello

TAGNIN, Stella Esther Ortweiller

TAMAGNO, Júlia

TAVARES DA SILVA, Priscila

TEIXEIRA, Elisa Duarte

TEIXEIRA, Gustavo Leal

TELES OLIVEIRA, Helena Cid

TOLEDO, Gabriela Dias

TORRENT, Tiago Timponi

VALE, Oto Araújo

VALVERDE DA SILVA, Júlia Cristina

VICTOR, Ana Clara Taborda de Paula

VIEIRA, Marcelo Augusto

VITAL, Átila Augusto Soares

WICK-PEDRO, Gabriela

ZUCCHI, Angela Maria Tenório

Ordem por prenome

Adriana Silvina PAGANO

Amanda Letícia Valadares dos SANTOS

Ana Clara Taborda de Paula VICTOR

Ana Eliza Pereira BOCORNY

Angela Maria Tenório ZUCCHI

Anne O'CONNOR

Átila Augusto Soares VITAL

Bianca Mara GUEDES DE SOUZA

Bruna Rodrigues da SILVA

Bruno Neves Rati de Melo ROCHA

Carlos Henrique KAUFFMANN

Carolina BOHORQUEZ

Carolina Godoi de Faria MARQUES

Cláudia Dias de BARROS

Cristiane Krause KILIAN

Cristina Borges GIL

Danilo Duarte COSTA

Diva CARDOSO DE CAMARGO

Elisa Duarte TEIXEIRA

Emiliana FERNANDES BONALUMI

Flávia de Oliveira MAIA-PIRES

Gabriela Dias TOLEDO

Gabriela WICK-PEDRO

Giovana de Castro MARCHESE

Graziele Altino FRANGIOTTI

Helena Cid TELES OLIVEIRA

Heloísa Orsi Koch DELGADO

Iasmin Valéria Miranda RABELO

Isabela MOREIRA DE OLIVEIRA

Ísis Beber de Souza Fiorilo ROCHA

Jackson Wilke da Cruz SOUSA

Jasper Vilan BRAGA

Jhonatan Henrique LOPES

João FEKETE

Júlia Cristina VALVERDE DA SILVA

Julia de Souza BATISTA

Júlia TAMAGNO

Luan Daniel dos Santos SOUSA

Lucas Renato dos Santos MURTA

Luciana Carvalho FONSECA

Luciano Franco da SILVA

Magali Sanches DURAN

Marcelo Augusto VIEIRA

Maucha Andrade GAMONAL

Mauricio José Ferreira LOPES

Oto Araújo VALE

Patrícia Helena FREITAG

Paula Tavares PINTO

Rozane R. REBECHI

Rozane Rodrigues REBECHI

Shirlene Bemfica de OLIVEIRA

Simone OLIVEIRA

Stella Esther Ortweiller TAGNIN

Tanara Zingano KUHN

Thiago Alexandre Salgueiro PARDO

Thiago Blanch PIRES

Tiago Timponi TORRENT

Tommaso RASO

Tony Berber SARDINHA

Ursula Puello SYDIO

Wagner da Cunha NUNES

Ziyang LIU

LISTA COM PALAVRAS-CHAVE

Academic writing

Acessibilidade

Acessibilidade comunicacional

Agrarian science corpus

Agrarian sciences

Agrupamentos

Análise multidimensional

Anotação

Anotação de córpus

Antconc

Aposições predicativas

Argumentação

Artificial intelligence

Árvore de domínio

Audiodescrição

Automação com inteligência artificial

Blog de coworking

Brazilian television

Catálogo

Chavicidade

Ciências agrárias

Classificadores semânticos

Colocações

Colocações acadêmicas

Compilação

Compilação de corpus

Computational linguistics

Conceitos

Conceptual domain identification

Contraste

Convencionalidade

Corpora

Corpus

Corpus de aprendizes

Corpus linguistics

Corpus multimodal

Corpus oral

Corpus paralelo

Corpus

Corpus-based metaphor studies

Deficiência

Dicionário de colocações

Dilma Vana Rousseff

Discurso de ódio

Discursos presidenciais

Disfluências

Ditadura militar

Divulgação científica

DIY corpora

English language

Ensino de línguas estrangeiras

Ensino de línguas

Ensino médio técnico

Escolaridade limitada

Escrita acadêmica

Esquizofrenia

Estudos da Tradução

Estudos Descritivos da Tradução

Estudos Feministas da Tradução

Estupro

Expressões idiomáticas

Extração

Extração automática e semiautomática

Extração terminológica

Extratores automáticos e semiautomáticos de terminologia

Fala espontânea

Feminismo

Ferramentas de auxílio à tradução

Formas verbais

Fraseologia

Fraseologia português-inglês

Fraseologismos

Função adversativa

Gênero

Gêneros acadêmicos

Glossário

Glossário bilíngue

Idiomatic expressions

Ilocuções

Inclusão surda

Inglês

Inglês acadêmico

Inteligência artificial

Interview analysis

Introdução

Introdutor locutivo

Jair Messias Bolsonaro.

L-act

Legislação federal brasileira

Lei geral de proteção de dados pessoais

Lex-br-ius

Lexical bundles

Lexical frames

Léxico

Léxico tabu

Limpeza de textos.

Língua inglesa

Língua italiana

Linguagem jurídica

Linguagem simples

Linguistic features

Linguística

Linguística computacional

Linguística de corpus

Linguística de corpus de aprendiz

Literatura brasileira traduzida

Literatura estrangeira

Luiz Inácio Lula da Silva

Machado de Assis

Memória

Metáfora

Metaphor identification

Modal verbs

Moderação de plataformas

Mulheres

Multi-dimensional analysis

Multimodal corpus

Multi-palavras

Natural language processing

N-grama de classe semântica.

Nheengatu

Nilc-metrix

Nominalization

Normas de tradução

O joio e o trigo

Objetivos de desenvolvimento sustentável (ODS)

Onomasiology

Padronização textual

Parsing sintático

Patriarcado

Pedagogia da tradução

Personal pronouns

Plataforma de dicionários

PLN

Português

Portuguese-english contrastive studies

Práticas discursivas

Pré-processamento de corpus

Processamento de linguagem natural

Prosódia

Python

Quadros e pacotes lexicais

Receitas

Recorrentes e preferenciais

Recursos lexicográficos

Redes sociais

Research articles

Research paper writing

Resenhas literárias

Resumo científico

Roda viva corpus

RST

Satire

Satirical news

Semantic fields

Semântica

Semântica de frames

Shape and stem disciplines

Simplificação textual e terminológica

Sintaxe

Sketch engine

Social media

Terminologia

Terminologia para tradução

Termos

Traços de normalização

Tradução

Tradução audiovisual acessível

Tradução automática

Tradução cultural

Tradução especializada

Tradução

Translator-oriented glossary

Transtorno do humor bipolar

Treebank

Tupinologia

UD

Universal dependencies

User- generated content

Variação

Variação diacrônica

Vocábulos

SUMÁRIO

RESUMOS
EBRALC-2024

RODA VIVA: UM CORPUS ORAL E A UNIVERSAL DEPENDENCIES

Cláudia Dias de BARROS[1]
Oto Araújo VALE[2]

Neste resumo é apresentado o trabalho sobre a construção de um subcorpus do Corpus Roda Viva (MIRANDA JR. et al., 2024), que é formado por 713 entrevistas de vários anos do programa Roda Viva da TV Cultura, transcritas por jornalistas de forma textualizada, nas quais há complementações das falas, por meio de inserções textuais, informações enciclopédicas, entre outros, o que faz com que percam o status de língua oral, passando a língua escrita. Dessa forma, nesta pesquisa tomou-se a decisão de construir o subcorpus com quatro entrevistas e, a fim de manter o status de língua oral, decidiu-se realizar a transcrição automática das entrevistas por meio de um ASR (Sistema de Reconhecimento Automático de Fala) chamado Whisper (RADFORD et al., 2023). Os textos transcritos apresentaram alguns problemas como transcrição equivocada de algumas palavras e erro de segmentação das sentenças, que precisaram ser corrigidos manualmente posteriormente. A escolha das quatro entrevistas se deu baseada na possível diversidade sintática apresentada pelos quatro entrevistados, sendo eles: uma governadora, um desenhista de história em quadrinhos, um jogador de futebol, e um rapper. A partir dos textos transcritos revisados foi realizada a anotação automática com o formalismo da Universal Dependencies (UD) (DE MARNEFFE et al., 2021), um projeto que tem como objetivo uma anotação gramatical consistente (etiquetas morfossintáticas, características morfológicas e dependência sintática), entre línguas humanas diferentes. Atualmente, a UD possui dezessete etiquetas morfossintáticas ou Part-of-Speech (PoS) tags e 37 etiquetas de relações de dependência – deprel (de dependency relation), que ligam dois a dois os elementos (tokens) de uma sentença. Um deles é chamado de head (núcleo), que é sempre uma palavra de conteúdo e o outro é chamado de dependente. A anotação UD foi realizada pelo parser PortParser (LOPES et al., 2024) e após isso, foi feita uma revisão manual por meio da ferramenta Arborator-Grew ElizIA (GUIBON et al., 2020) e foram identificados alguns fenômenos característicos da língua falada, como a presença de vocativos e marcas discursivas, como né, hein, entendeu, entre outros. A entrevista com o rapper foi a que apresentou menor formalidade e se mostrou desafiadora para o parser anotar corretamente as relações sintáticas. Na entrevista com a governadora do estado, observou-se uma grande quantidade de orações subordinadas e coordenadas, fruto de um discurso mais prolixo, característico de um político. A entrevista do jogador de futebol apresentou a ocorrência de muitas etiquetas dislocated, as quais marcam a presença de um item que poderia ser classificado como o sujeito da oração, mas, por estar longe do verbo, é substituído por um outro sujeito mais próximo, como ‘João’ em: “O João, ele sempre foi uma pessoa desconfiada”. O objetivo dessa

[1] Docente do Curso de Licenciatura em Letras, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Câmpus Sertãozinho

[2] Docente do Curso de Licenciatura em Letras e Bacharelado em Linguística, Universidade Federal de São Carlos – UFSCar.

anotação é fornecer um corpus de língua oral (a princípio as 5 entrevistas e, posteriormente, as outras 708 do Projeto Roda Viva) ao projeto Porttinari (PARDO et al., 2021), um grande corpus multigênero do Português do Brasil, composto por textos escritos, como artigos de jornal, tweets do mercado financeiro brasileiro, revisões de consumidores de e-commerce e revisões de livros.

Palavras-chave: Universal Dependencies; sintaxe; Linguística de Corpus; PLN.

ANÁLISE LINGUÍSTICA DE DISCURSOS PRESIDENCIAIS: UM ESTUDO BASEADO EM CORPUS

Ísis Beber de Souza Fiorilo ROCHA[3]
Bruno Neves Rati de Melo ROCHA[4]

Este trabalho mapeia temas relacionados ao uso das palavras “Brasil”, “Deus” e "governo" em discursos presidenciais do primeiro ano do primeiro mandato de Luiz Inácio Lula da Silva (2003), Dilma Vana Rousseff (2011) e Jair Messias Bolsonaro (2019). A análise visa entender a maneira pela qual cada presidente lida com questões evocadas por esses termos, usando como arcabouço teórico-metodológico a Linguística de Corpus (McEnery; Wilson, 1996; Sardinha, 2004). Para tanto, compilou-se um corpus de 1001 textos de discursos presidenciais, retirados do site da Biblioteca da Presidência, totalizando aproximadamente 1.600.000 tokens. O corpus é formado pelas transcrições de todos os discursos proferidos pelos presidentes no período estudado e exclui os discursos de vice-presidentes e outros representantes governamentais. Usando o Corpus Brasileiro (Sardinha, 2010) como corpus de referência, gerou-se listas de frequência (com/sem stopwords), listas de palavras-chave (com/sem stopwords), colocados das palavras alvo, linhas de concordância das palavras alvo e de colocados das palavras alvo. A escolha das palavras alvo se baseou em suas posições nas listas de frequência e de palavras-chave dos subcorpora: para todos os presidentes, “Brasil”, “Deus” e "governo" estão entre as cinco palavras lexicais mais frequentes e as cinco primeiras palavras-chave de cada mandato. Os dados analisados sugerem que cada presidente utiliza de maneira específica cada uma das palavras alvo. Para Lula e Dilma, "Brasil" aparece em colocados que indicam temas econômicos (“banco” e “risco” para Lula e “rico” para Dilma), políticas públicas e programas sociais (“analfabetismo” e “miséria” para Lula e “sem fronteiras” e “sem miséria” para Dilma), e termos que evocam relações exteriores (“ligação” e “Argentina” para Lula e “embaixador” e “Venezuela” para Dilma). Para Bolsonaro, "Brasil" ocorre sobretudo em contextos que expressam uma ideia de “resgate do Brasil” (como “Brasil que ressurge” e “colocar o Brasil no local de destaque”). A análise de “Deus” sugere que Bolsonaro relaciona sua fé pessoal a assuntos do governo (“agradeço”), evoca o movimento integralista (“Deus, família, Brasil”) e propaga o slogan de sua campanha eleitoral (“Brasil acima de tudo, Deus acima de todos”). Para Lula, “Deus” também é usado em contextos que apontam fé pessoal (“queira”, “peço”, “graças”). Para Dilma, a palavra “Deus” é pouco frequente, não tendo sido analisada detalhadamente. Quanto a “governo”, o primeiro colocado de Lula e Dilma é “federal”, ao passo que “meu” aparece em terceiro lugar, sugerindo que se referem ao próprio governo sobretudo de maneira institucional, mas também pessoal. Os demais colocados relacionam-se a temas como economia, relações exteriores e

[3] Bacharel em Letras-Estudos Linguísticos/ Ênfase em Linguística do Texto e do Discurso pela Universidade Federal de Minas Gerais (UFMG), membro do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da UFMG.

[4] Professor efetivo no curso de Letras da UFMG. Doutor em Estudos Linguísticos pela Faculdade de Letras da UFMG.

programas sociais (“sociais” e “programa” para Lula e “compromisso” e “bolsas de estudo” para Dilma). Para Bolsonaro, o primeiro colocado de “governo” é “nosso”, evidenciando uma postura pessoal, enquanto “federal” não figura entre primeiras vinte posições. Além disso, Bolsonaro co-relaciona "governo" a governos militares ditatoriais (“Médici” e “Figueiredo”) e civis não ditatoriais (“Sarney” e “anterior”, referindo-se a Michel Temer) e também a expressões como “respeita a família”, "adora a Deus" e “honra os militares”.

Palavras-chave: Linguística de Corpus; Discursos Presidenciais; Compilação de Corpus; Luiz Inácio Lula da Silva; Dilma Vana Rousseff; Jair Messias Bolsonaro.

IDENTIFICAÇÃO SEMIAUTOMÁTICA DE EQUIVALENTES ENTRE EXPRESSÕES IDIOMÁTICAS COM FRUTAS EM PORTUGUÊS-INGLÊS: DESCASCANDO ESSE ABACAXI

Julia de Souza BATISTA[5]
Isabela MOREIRA DE OLIVEIRA[6]
Stella Esther Ortweiler TAGNIN[7]
Elisa Duarte TEIXEIRA[8]
Rozane Rodrigues REBECHI[9]

As expressões idiomáticas (EIs) apresentam grande dificuldade para a tradução e a aprendizagem de línguas, seja por sua opacidade, seja pelo fato de que seu aprendizado só se dá por meio de repetidas exposições (MATTOS, 2003). Levando em conta as dificuldades enfrentadas por pesquisadores que trabalham com essas unidades (p. ex. XATARA, 2001; XATARA et al., 2001; SAG et al., 2002; RIVA, 2009; TAGNIN, 2013; PINNAVAIA, 2018; REBECHI e TRINDADE, 2021; SILVA e TEIXEIRA, 2021; ADEWUMI et al., 2022; ORTIZ ALVAREZ 2022), pensou-se em um projeto que visa, inicialmente, coletar o maior número possível de EIs contendo palavras relacionadas à grande área da alimentação, em português, inglês, italiano, chinês e espanhol, por enquanto. Planeja-se, em seguida, desenvolver estratégias de classificação e correlação das EIs entre as línguas, de modo a permitir sua identificação (semi-)automática em corpora, bem como a de possíveis equivalentes para diferentes contextos de tradução e de aprendizado de língua estrangeira. O objetivo do projeto é criar um sistema informatizado online para a identificação, coleta e consulta de EIs utilizando uma abordagem que permita uma busca tanto pela expressão em si, quanto pelos seus sentidos, ou seja, onomasiológica. O presente trabalho relata o planejamento e testagem das abordagens metodológicas utilizadas até o momento e os resultados obtidos no subconjunto de EIs contendo palavras do campo semântico FRUTAS, no par de línguas português-inglês. Esse primeiro recorte teve como objetivo principal chegar a uma lista de classificadores que pareça razoável e que permita a posterior identificação de EIs equivalentes em duas ou mais línguas de forma (semi-)automática. Tendo como base uma lista criada a partir da junção de: i) classificadores resultantes de um trabalho de mestrado sobre EIs da área da alimentação (MOREIRA DE OLIVEIRA, 2022); ii) o conjunto de dados para classificação refinada de emoções do Go Emotions

---

[5] Graduada em Línguas Estrangeiras Aplicadas (LEA-MSI) pela Universidade de Brasília (UnB); Graduanda em Letras Tradução Inglês também pela UnB.

[6] Mestra em Estudos da Tradução pela Universidade de Brasília (UnB), docente temporária do Departamento de Línguas Estrangeiras e Tradução (LET) da UnB e doutoranda do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS).

[7] Docente associada da Universidade de São Paulo (USP), atua nos Programas de pós-graduação em Estudos Linguísticos e Literários em Inglês e LETRA (USP).

[8] Docente da área de Tradução - Inglês do Departamento de Línguas Estrangeiras e Tradução (LET) da Universidade de Brasília (UnB), membro do Programa de Pós-Graduação em Estudos da Tradução - POSTRAD da UnB.

[9] Docente do Departamento de Línguas Modernas e Professora Permanente do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS).

(DEMSZKY et al., 2020); iii) uma lista de emoções humanas elaborada pelo Chat GPT; iv) e os classificadores listados no Themes of Oxford Dictionary of Idioms Index (AYTO, 2020), utilizados no trabalho de Rafatbakhsh e Ahmadi (2019), cinco pesquisadoras da equipe classificaram uma lista em português e outra em inglês contendo cada mais de 60 EIs do referido campo semântico. Os resultados individuais foram comparados e a lista, refinada. As EIs foram, então, classificadas novamente por cada pesquisadora, utilizando-se essa versão, contendo cerca de 90 classificadores. Depois de chegarem a uma classificação majoritariamente consensual das EIs em ambas as línguas, foi feito o cruzamento dos dados para verificar o quanto essa metodologia se mostraria eficaz na identificação (semi-)automática de possíveis equivalentes interlinguais. Observamos, por exemplo, que do total de 90 classificadores, 60 foram usados e "facilidade" nos permitiu identificar uma possível equivalência entre "mamão com açúcar" e "as easy as apple pie", "easy peasy lemon squeezy" e "low hanging fruit"; e entre "a cereja do bolo" e "the cherry on top", ambas classificadas com "excelência". Concluído este balão de ensaio, nosso próximo passo será expandir a classificação para outras categorias, procurando refinar ainda mais a lista de classificadores, sempre pensando em outras formas de correlacionar esses dados.

Palavras-chave: Expressões idiomáticas; classificadores semânticos; fraseologia português-inglês; PLN; Linguística de Corpus.

NOT EVERY B!7CH IS A B1TC|-|: TESTANDO A MODERAÇÃO DE DISCURSO DE ÓDIO EM ESPAÇOS VIRTUAIS COM PADRÕES TEXTUAIS

Priscila TAVARES DA SILVA[10]

Em espaços virtuais de interação, é comum que usuários perpetradores de discursos de ódio driblem a moderação (Cristani, 2022) utilizando Leetspeak, que são escritas alternativas com caracteres especiais (Froud, 2021), e Algospeak, neologismos ou combinações de palavras que fazem alusão ao que se pretende dizer (Huyghe, 2022). A presente pesquisa descreve um projeto piloto para propor e testar a efetividade de moderação de espaços virtuais a partir da análise de combinações de palavras. Espera-se demonstrar que existem padrões textuais recorrentes nesses discursos de ódio e que a moderação por agrupamentos de palavras pode ser mais efetiva do que por palavras isoladas. O objetivo é chegar a uma metodologia que possa ser aplicada com eficácia para diminuir os casos em que os usuários conseguem driblar a moderação. Utilizando ferramentas da Linguística de Corpus, pode-se analisar a co-ocorrência entre itens lexicais em um contexto próximo (Teixeira, 2008, p. 170). A análise é feita com um software capaz de comparar textos organizados em um corpus, que é um conjunto extenso o suficiente para ser representativo do fenômeno que se busca avaliar (Berber Sardinha apud Teixeira, 2008, p. 159). Para a análise proposta, o primeiro passo é identificar em quais contextos os itens lexicais são utilizados como discurso de ódio, entendido como um discurso disciplinador que busca manter ou reforçar o status quo de poder de um grupo sobre outro (Foucault apud Silva, 2016, p. 40). O objeto da análise é o corpus TwitterHateSpeech , que contêm tweets do X. Os dados serão analisados com o software AntConc. Pretende-se levantar e classificar as palavras-chave desse corpus em duas categorias: “sim” e “não”, em que “sim” corresponde a “palavra utilizada em contexto de discurso de ódio”. Em seguida, observaremos quais correlações surgem da análise de clusters e n-grams das palavras levantadas na pesquisa. Caso existam padrões reconhecíveis em seu uso, checaremos se buscas com esses padrões retornam somente tweets classificados como discurso de ódio, mesmo sem conter as palavras-chave, o que permitiria combater os artifícios utilizados pelos usuários. Nossa hipótese é de que a análise dos padrões permitiria chegar a uma lista de clusters e n-grams cujo uso, em um corpus não rotulado como discurso de ódio, permita identificar sua ocorrência mesmo sem a presença das palavras-chave. A título de ilustração, escolhemos a palavra bitch, que ocorre 8.353 vezes no corpus. Nas 500 primeiras ocorrências somente 92 foram classificados como discurso de ódio. Dentre os padrões mais comuns estão as expressões look like a bitch e be a bitch, usados como sinônimos de “covarde” ou “difícil”, como em “@BarackObama looks like a bitch in foreign policy” e “Life is a bitch”. Esses casos são mais numerosos do que o uso em discursos de ódio, como em “Every nigga can make a female act like a bitch #ThatsAFact”.

[10] Mestranda em Estudos da Tradução na Universidade de Brasília

Palavras-chave: Linguística de Corpus; discurso de ódio; moderação de plataformas; padronização textual; AntConc.

ABSTRACT REGISTER VARIATION BETWEEN HUMAN AND AI

João FEKETE[11]
Deise Prina DUTRA[12]

In the little time chatbots powered by Large Language Models (LLMs) and Artificial Intelligence (AI) have been available to the general public, different uses of them have been created (Holmes & Tuomi, 2022; Islam et al., 2020). The use of these technologies in writing has brought concerns to many areas, including academic writing (Thorp, 2023; Nature Editorials, 2023). In order to understand the real effect of such usages, studies have addressed AI detectors effectiveness (Hu et al., 2023; Lu et al., 2023; Walters, 2023), human judgment for detecting AI generated texts (Ma et al., 2023), and linguistic analysis of these texts (Berber Sardinha, 2024). In our research, we focus on expanding the linguistic analysis on the similarities between human-authored and AI-generated texts, taking into account lexicogrammatical features and their communicative functions in abstracts, which are understood as an academic register .The experiment was undertaken using two different corpora, 450 abstracts from high impact journals from 3 different areas (Reference corpus) and the second corpus 900 abstracts created by ChatGPT3.5, using as resource the content from the 450 articles of the reference corpus. The AI-generated corpus is split into two different subcorpora: Simple Query and Complex Query. The two corpora were created using LangChain (LangChain, 2024) and ChatGPT3.5-turbo (OpenAI, 2024). First, we gathered all the sections (except the abstract) from the articles from which the reference corpus was extracted, then, we used a splitter to segment the text into chunks and embedded the information to be used through Chroma db. From Langchain's retrieval chain, we provide all chunks to the AI, which then selects the ones which will be sent to ChatGPT at the moment of answering the query. For the Simple Query corpus we generated the abstracts with a simple prompt, which required only an abstract to be created based on the provided information (Simple Query) and, for the Complex Query, we created a prompt with the persona approach (White et al., 2023), which asked the AI to behave as an experienced article publisher and professor. To analyze the data, we make use of the Additive Multidimensional Analysis (AMDA) (Berber Sardinha et al., 2019) with the Dimensions 1, 2, and 5 from Biber (1988). Results have shown statistical differences between the AI-generated abstracts and human-authored for Dimensions 1, 2, and 5, for both AI corpora. From the output data, we can conclude that: human-authored texts display a greater variety of features; using the persona prompt does not guarantee a more human-like output from the AI model; and some communicative purposes are better defined for artificial intelligence than others, such as understanding abstracts as an information production type of text rather than an involved one. Furthermore, we encourage researchers to take into account the creation of testing data using authentic scenarios of AI use for text creation.

---

[11] Bacharel em Inglês pela Universidade Federal de Minas Gerais

[12] Professor titular da Universidade Federal de Minas Gerais

Palavras-chave: Artificial Intelligence; Multi-dimensional Analysis; Corpus Linguistics; Computational Linguistics; Academic Writing

CRIAÇÃO DE UM GLOSSÁRIO LIBRAS-PORTUGUÊS DO BRASIL: UMA INICIATIVA DE ACESSIBILIDADE COMUNICACIONAL

Gabriela WICK-PEDRO[13]

Esta pesquisa tem como objetivo a criação de um glossário Libras-Português do Brasil, visando promover a acessibilidade comunicacional, especialmente na divulgação científica. A proposta foca no desenvolvimento de conteúdos e infraestrutura voltados para a editoração e publicação de materiais científicos acessíveis, com especial atenção à comunidade surda. A comunicação de resultados de pesquisa é parte essencial do ciclo de vida da atividade científica, conectando a pesquisa à aplicação dos resultados, com potencial para melhorar a qualidade de vida e fomentar novas investigações. A divulgação científica é considerada um meio informal de comunicação científica, utilizando uma linguagem acessível a diversos públicos e incorporando novos elementos ao processo de circulação de informações científicas e tecnológicas (BJÖRK, 2005; BUENO, 2010). A linguagem pública na divulgação científica deve alcançar o maior número possível de pessoas, ultrapassando o círculo restrito dos especialistas acadêmicos (SILVA; LAZZAROTTI FILHO; SILVA, 2011). A acessibilidade comunicacional, dentro desse contexto, é essencial. A divulgação científica, entendida como um meio informal de comunicação, utiliza uma linguagem acessível e visa ampliar o público alcançado, superando o círculo de especialistas acadêmicos. Neste contexto, a acessibilidade comunicacional, especialmente para usuários da Língua Brasileira de Sinais (Libras), é crucial, considerando a diversidade da comunidade surda. O processo de criação do glossário envolve: i) a preparação e limpeza de corpus, ii) a extração automática de termos por meio de ferramentas de Processamento de Linguagem Natural (PLN) e Inteligência Artificial (IA), iii) a validação com especialistas da área e iv) a categorização dos termos. Além disso, há uma ênfase em como o projeto se ancora espacialmente em instituições e eventos, como indicam estudos sobre a relação entre território e produção científica (CERTEAU, 1994). A pesquisa, em andamento, visa criar um glossário bilíngue com o intuito de incluir mais efetivamente a comunidade surda na comunicação científica, promovendo uma inclusão mais abrangente. Resultados preliminares apontam para a viabilidade do uso de técnicas de Linguística de Corpus e PLN para a criação desse recurso, oferecendo um importante avanço no campo da acessibilidade comunicacional.

Palavras-chave: acessibilidade comunicacional; divulgação científica; glossário bilíngue; processamento de linguagem natural; inclusão surda.

[13] Pesquisadora do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília/DF, bolsista FINATEC.

RODA VIVA CORPUS: STRUCTURING A MULTIMODAL LINGUISTIC RESOURCE FROM BRAZILIAN TELEVISION INTERVIEWS

Gabriela WICK-PEDRO[14]
Cláudia Dias de BARROS[15]
Oto Araújo VALE[16]

The Roda Viva Corpus aims to formalize and structure a multimodal linguistic resource derived from interviews broadcasted on Roda Viva, a long-running interview show on TV Cultura, a staple of Brazilian television since 1986. The corpus includes both textual transcriptions and corresponding video recordings, with the initial dataset consisting of 713 interviews conducted between 1986 and 2009. These interviews feature prominent figures such as politicians, artists, scientists, and intellectuals, and were made publicly available through the Memória Roda Viva portal, a project initiated by FAPESP in 2007 (FAPESP, 2024). This study seeks to transform the raw data available on the portal into a linguistically structured corpus for use in Corpus Linguistics (CL) and Natural Language Processing (NLP). The primary aim is to provide a resource that enables detailed linguistic analyses, such as the investigation of discourse markers, interactional features, and pragmatic phenomena in Brazilian Portuguese as it is spoken in formal interview contexts. In particular, this corpus offers a unique opportunity to analyze multimodal data, combining textual transcriptions with corresponding video recordings, which are essential for understanding non-verbal cues in communication (Botin, 2016; Pacheco, 2020). The construction of the Roda Viva Corpus involved extensive data cleaning, and two preliminary versions of the corpus are currently available. Version 0.1 maintains the original transcriptions with minimal cleaning, such as the removal of hyperlinks and special characters, while Version 0.2 offers a more refined dataset that excludes non-verbal transcriber interventions (e.g., "coughing," "sighing"). Both versions are available in CSV and JSON formats, allowing for flexibility in computational processing. The corpus is also annotated with metadata, including the date of the interview, the name of the interviewee, the speaker's identity, and the order of each utterance. Although the Roda Viva material has been academically cited in only a few studies , this project aims to fill the gap by offering a formally structured resource that can support a wide range of linguistic and computational research. The project represents a step forward in the integration of CL and NLP, providing a rich, multimodal dataset that bridges the gap between traditional linguistic analysis and modern computational methods. The corpus is designed to promote interdisciplinary research at the intersection of Corpus Linguistics, Artificial Intelligence, and media studies. This contribution aims to not only enhance the availability of Brazilian Portuguese

[14] Pesquisadora em Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília/DF, bolsista FINATEC

[15] Docente Efetiva de Português/Inglês no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Campus Sertãozinho, Sertãozinho/SP

[16] Docente Associado do Departamento de Letras (DL) na Universidade Federal de São Carlos (UFSCAR), São Carlos/SP

linguistic resources but also provide a robust tool for exploring multimodal communication in high-stakes, formal interview settings.

Palavras-chave: Roda Viva corpus; multimodal corpus; corpus linguistics; brazilian television; interview analysis.

CONSTRUÇÃO DO DOMÍNIO DA ACESSIBILIDADE POR MEIO DE FRAMES SEMÂNTICOS: UMA CONTRIBUIÇÃO PARA A FRAMENET BRASIL

Iasmin Valéria Miranda RABELO[17]
Maucha Andrade GAMONAL[18]
Adriana Silvina PAGANO[19]

A teoria linguístico-cognitiva da Semântica de Frames (FILLMORE, 1982) tem por objetivo compreender a construção de significado através de contextossócio-culturais. Para isso, a teoria se fundamenta nos frames, uma estrutura designificados relacionados que evidencia estruturas cognitivas para a representaçãode conceitos específicos. Nesse contexto, a FrameNet Brasil, desenvolvida naUniversidade Federal de Juiz de Fora (UFJF), é um projeto de lexicografiacomputacional que busca desenvolver tecnologias linguísticas com base nSemântica de Frames e na Linguística de Corpus. Dessa forma, este trabalho busca expandir a base de dados da FrameNet Brasil a partir do léxico especializado da área da acessibilidade, através da construção de um domínio específico. Esta pesquisa se ancora, primeiramente, na teoria da Semântica de Frames (FILLMORE, 1982), suas aplicações léxicas e sintáticas (GAWRON, 2008) e o projeto FrameNet Brasil (SALOMÃO, 2009). As diretrizes para construção de um domínio e criação de frames seguem os trabalhos de Dutra (2024) e Gamonal (2013), assim como os critérios de compilação de corpus definidos por Sardinha (2004). Os conceitos de semântica lexical abordados têm como base o trabalho de L'Homme (2020) e a relevância da construção de frames em domínios especializados (L'HOMME ET AL, 2014). Ademais, esta pesquisa também abrange o histórico da acessibilidade noBrasil (COSTA ET AL, 2005) e o desenvolvimento da Tecnologia Assistiva (RODRIGUES e ALVES, 2013). Este projeto se desenvolve a partir de uma pesquisa quantitativa de abordagem bottom-up, partindo dos dados para a criação de frames. A primeira etapa, já em desenvolvimento, consiste na investigação do léxico especializado através da compilação de um pequeno corpus sobre a área da acessibilidade (cartilhas, glossários e outros recursos) com, até então, 37774 tokens e 6454 types. O corpus foi analisado através do software concordanceador AntConc, identificando as terminologias específicas mais frequentes nos documentos e seus contextos de uso. A partir disso, as possíveis unidades lexicais foram extraídas. O próximo passo, também em progresso, compreende o mapeamento dessas unidades lexicais candidatas com os dados já presentes na FrameNet Brasil. Após isso, é possível sugerir a criação de novos frames para a construção do domínio da acessibilidade, com definições mais específicas e abrangentes dos conceitos e terminologias da área. Por fim, através das relações entre frames

[17] Graduanda da Faculdade de Letras da Universidade Federal de Minas Gerais. Universidade Federal de Minas Gerais, Minas Gerais.

[18] Residente de pós-doutoramento no Programa de Pós-Graduação em Linguística na Universidade Federal de Minas Gerais, Minas Gerais, atualmente é bolsista Capes.

[19] Professora Titular de Linguística Aplicada da Universidade Federal de Minas Gerais, Minas Gerais, bolsista de produtividade em Pesquisa IC do CNPq.

é possível conectar o domínio da acessibilidade à rede léxico-semântica de FrameNet Brasil. Com as ULs e frames criados, a última etapa na modelagem de um domínio é a anotação de unidades lexicais realizada de acordo com a metodologia de anotação lexicográfica da FrameNet Brasil. As etapas já em andamento desta pesquisa revelam a riqueza lexical da área da acessibilidade. As investigações e análises terminológicas indicam um domínio vasto e com diversas possibilidades de aplicação dentro da FrameNet Brasil. Expandir a base de dados com esse léxico especializado permite uma compreensão aprofundada dos termos e conceitos relevantes para uma parcela discriminada da população brasileira. Além disso, a amplificação das unidades lexicais proporciona mais informações para o desenvolvimento de tecnologias linguísticas e possibilita futuras pesquisas na área.

Palavras-chave: Semântica de Frames; criação de frames; domínio especializado; acessibilidade; deficiência.

ENSINO DE ITALIANO EM CONTEXTO UNIVERSITÁRIO: POR UMA REFLEXÃO LINGUÍSTICA DIRECIONADA POR CORPUS

Graziele Altino FRANGIOTTI[20]

Esta proposta tem como objetivo contribuir para a discussão sobre possíveis caminhos para um ensino de línguas estrangeiras direcionado por corpus, focalizando de maneira especial a aprendizagem do italiano em contexto universitário. Parte-se do pressuposto de que estudos que dialoguem com a Linguística de Corpus são pouco numerosos na área do ensino de italiano no Brasil (FRANGIOTTI, 2019; FRANGIOTTI, 2021), lacuna essa que leva à indagação sobre quais seriam os efeitos do ensino direcionado por dados no aprendizado dessa língua por brasileiros. De modo mais específico, pretende-se compreender se e de que maneira um percurso didático formulado a partir de uma obra literária escrita na Idade Média e de um conjunto de suas traduções pode fomentar a compreensão sobre mudanças linguísticas na língua italiana. Como organização geral da apresentação, pretende-se, em um primeiro momento, descrever o caminho teórico-metodológico que orienta a formulação e a aplicação de uma sequência didática em uma disciplina oferecida no curso de graduação em Letras Italiano na Universidade Federal de Santa Catarina (UFSC). Para tanto, parte-se dos trabalhos de Berber Sardinha (2010), Morales (2008), Condi de Souza (2005), Barbosa (2004), entre outros, para a apresentação das potencialidades do data-driven learning. Já com base em Antunes (2014, 2009) e Carter e McCarthy (1995) expõe-se uma concepção de ensino de línguas estrangeiras centrada na análise textual e na reflexão metalinguística indutiva como catalisadores do desenvolvimento de competência comunicativa. Finalmente, levando em conta a taxonomia de Bloom e sua revisão (respectivamente BLOOM [1972] e KRATHWOHL [2002]), sustenta-se uma perspectiva de ensino que diversifique as atividades propostas em sala de aula e que coloque a comparação entre as línguas como uma técnica didática relevante para a promoção do noticing (SCHMIDT, 1990). Após essa etapa teórica, será caracterizada a sequência didática propriamente dita, que se ancora, sobretudo, na obra italiana Il Principe (1532) de Nicolau Maquiavel e em algumas de suas principais traduções para o português. Posteriormente, serão discutidos os resultados dos participantes da pesquisa em atividades didáticas onde a identificação de semelhanças e diferenças entre as línguas foi estimulada e ponderados os efeitos desse procedimento para a sensibilização quanto à variação linguística diacrônica.

Palavras-chave: Linguística de corpus; ensino de línguas estrangeiras; língua italiana; variação diacrônica.

[20] Docente do Departamento de Língua e Literatura Estrangeiras da Universidade Federal de Santa Catarina, Florianópolis/SC, bolsista ADC-1C do CNPq.

AS CARACTERÍSTICAS PROSÓDICAS TEMPORAIS DA UNIDADE INFORMACIONAL DE INTRODUTOR LOCUTIVO

Gabriela Dias TOLEDO[21]
Marcelo Augusto VIEIRA[22]
Tommaso RASO[23]

Segundo a Language into Act Theory (L-Act), teoria corpus driven que tem como foco a organização da fala espontânea, as unidades informacionais podem ser identificadas a partir da sua funcionalidade, do seu perfil prosódico e da sua distribuição em relação ao Comentário (unidade com força ilocucionária) e agrupadas a partir de sua macro-funcionalidade textual ou dialógica (CRESTI, 2000). O Introdutor Locutivo (INT) é uma unidade informacional textual cuja função é sinalizar que os elementos que o sucedem devem ser interpretados pelo ouvinte em um plano pragmaticamente diferente do resto do enunciado, evidenciando um salto para outro nível hierárquico. Seu perfil prosódico não possui foco ou forma definida, mas é descendente, com frequência fundamental e intensidade menor que o elemento seguinte e taxa de articulação maior que o resto do enunciado (CRESTI, 2000; MAIA ROCHA; RASO, 2011). O objetivo da pesquisa é investigar as características prosódicas temporais do INT seguido do discurso reportado (metailocução mais realizada após a unidade), buscando descrever um aspecto formal que pode ser crucial para veicular a função dessa unidade informacional e diferenciá-la das demais com maior acurácia. Metodologia desenvolvida para a pesquisa: seleção de 58 INTs de minicorpora compilados e tratados no projeto C-ORAL-BRASIL (RASO; MELLO, 2012), agrupados de acordo com o seu tamanho (número de palavras prosódicas); segmentação e anotação manual dos INTs e dos contextos adjacentes a eles (unidade que precede o INT e o discurso reportado) pelo Praat (BOERSMA; WEENINK, 2023); extração automática das medidas de duração normalizada, taxa de articulação e proporção de apagamento silábico do INT e dos seus contextos adjacentes a partir da adaptação do script SGdetector (BARBOSA, 2006); tratamento estatístico e modelagem linear de efeitos mistos com o auxílio do R (R CORE TEAM, 2023) das características prosódicas temporais do INT. Resultados das análises estatísticas: o INT possui menor duração, maior taxa de articulação e maior proporção de apagamento silábico que seus contextos adjacentes; quanto maior o INT, menor sua taxa de articulação e proporção de apagamento silábico; o INT possui maior proporção de apagamento silábico em oxítonas que seus contextos adjacentes, enquanto que em paroxítonas não há diferença significativa entre as estruturas; as palavras mais frequentemente realizadas no INT são as oxítonas, principalmente o verbo 'falar' e o advérbio 'assim', enquanto que as mais frequentemente realizadas nos contextos

[21] Aluna de doutorado do Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Minas Gerais, Belo Horizonte/Minas Gerais.

[22] Aluno de doutorado da School of Communication Sciences and Disorders da Universidade McGill, Montreal/Quebec.

[23] Docente da Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte/Minas Gerais.

adjacentes são as paroxítonas. Ainda, análises preliminares (sem tratamento estatístico) sugerem que INTs com duas ou mais palavras prosódicas tendem a acelerar do começo da unidade até o alongamento pré-fronteiriço ao discurso reportado, sendo a maior parte de sua estrutura acelerada.

Palavras-chave: L-Act; fala espontânea; prosódia; Introdutor Locutivo; Linguística de Corpus.

DESAFIOS METODOLÓGICOS NA COMPILAÇÃO DO CORPUS DE TEXTOS ACADÊMICOS DAS CIÊNCIAS AGRÁRIAS

Deise Prina DUTRA[24]
Ana Eliza Pereira BOCORNY[25]
Danilo Duarte COSTA[26]
Gustavo Leal TEIXEIRA[27]
Carolina Godoi de Faria MARQUES[28]

No campo dos estudos em Inglês para fins Acadêmicos (IFA), corpora especializados desempenham um importante papel em investigações linguísticas de base empírica. No que se refere à escrita acadêmica, atenção tem sido dada para o fato de que os textos das diferentes áreas do conhecimento apresentam especificidades linguísticas (HYLAND, 2004; 2006). No campo das ciências agrárias, uma área estratégica para o desenvolvimento nacional, observou-se uma carência de estudos que se proponham a descrever a escrita acadêmica de brasileiros produzida em inglês como língua adicional. Para preencher esta lacuna e, a fim de compreender a linguagem utilizada nesta área no Brasil, visando propor instrumentos de apoio ao desenvolvimento linguístico da comunidade acadêmica, identificamos a necessidade da compilação de dois corpora. O primeiro contém textos produzidos por autores pouco experientes - dissertações e teses - e o segundo por artigos publicados em revistas de alto impacto internacionais para posterior comparação. Neste trabalho descrevemos os procedimentos adotados para a compilação do corpus de teses e dissertações (Corpus de Textos Acadêmicos das Ciências Agrárias) arquitetado de maneira a ser representativo da comunidade acadêmica de autores pouco experientes das ciências agrárias. Os procedimentos metodológicos adotados para a compilação seguem os pressupostos trazidos pela Linguística de Corpus, sendo embasados sobretudo nos conceitos de representatividade (BIBER, 1993) e balanceamento (SINCLAIR, 2004). Os textos foram obtidos a partir de buscas em repositórios institucionais de universidades brasileiras das cinco regiões do país. O corpus foi dividido em subcorpora, definidos a partir dos cursos ofertados por um instituto de ciências agrárias de uma universidade federal: Agronomia, Engenharia Agrícola, Engenharia de Alimentos, Engenharia Florestal e Zootecnia. Selecionados os textos, procedeu-se à etapa de limpeza, realizada manualmente com o objetivo de removerem-se os caracteres especiais, gráficos, tabelas, imagens e números de páginas. Dividiu-se os textos em seções: resumo, introdução, metodologia, resultados e conclusão, uma vez que possuem funções específicas, com estrutura e traços linguísticos distintos. Por fim, aos textos foi atribuído um código para facilitar a organização do corpus e a localização das

[24] Docente do Programa de Pós-graduação em Estudos Linguísticos (POSLIN) da Faculdade de Letras da UFMG

[25] Professora - Universidade Federal do Rio Grande do Sul, Porto Alegre/RS

[26] Professor - Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina/MG

[27] Professor - Universidade Federal de Minas Gerais, Montes Claros, Minas Gerais/MG

[28] Doutoranda - Programa de Pós-grafuação em Estudos Linguísticos, Universidade Federal de Minas Gerais, Belo Horizonte/MG

informações. Três fatores principais impactaram a compilação do Corpus de Textos Acadêmicos das Ciências Agrárias. O primeiro foi a ausência de Trabalhos de Conclusão de Curso escritos em inglês por brasileiros, resultando na exclusão dessa categoria do corpus. O segundo fator foi a disparidade no número de trabalhos disponíveis nos repositórios de cada região do país, especialmente na região Centro-Oeste, onde foram encontradas apenas duas teses. Esse dado contrasta com os resultados das outras regiões, destacando-se a região Sudeste, onde a maioria dos textos foi coletada. O terceiro fator foi a identificação de teses e dissertações compostas por artigos científicos em sua estrutura, o que parece ser uma prática comum nas ciências agrárias. Portanto, foi necessário criar dois subcorpora: um com textos no formato canônico (organizados por seções, como introdução e metodologia) e outro com textos não-canônicos (compostos por artigos). Foram obtidos um total de 180 trabalhos acadêmicos, dos quais 153 são do formato não-canônico e apenas 27 no formato canônico, totalizando aproximadamente 2.300.000 palavras.

Palavras-chave: Compilação de corpus; Escrita acadêmica; Ciências Agrárias.

DO FRASEOLOGISMO À METÁFORA NA EXPLORAÇÃO DO CORPUS JORNALÍSTICO DE O JOIO E O TRIGO

Bianca Mara GUEDES DE SOUZA[29]

Neste pôster, apresentamos uma exploração inicial de um corpus, parcialmente coletado, composto por textos de diferentes gêneros, extraídos do jornal O joio e o trigo. Portanto, estas são as primeiras análises realizadas para uma tese de doutorado em andamento. O joio e o trigo é um projeto de jornalismo investigativo que defende o papel central da prática jornalística como ferramenta de mudança social, especialmente tratando-se de temas como o combate às grandes corporações, com destaque para as do ramo alimentício e do agronegócio (O JOIO E O TRIGO, 2017). A busca exploratória realizada no corpus teve como objetivo a identificação e descrição de unidades fraseológicas e/ou unidades fraseológicas especializadas, seguida da identificação e descrição de expressões metafóricas com detalhamento da metáfora conceptual, relações de domínios (fonte e alvo), mapeamentos e desdobramentos. As unidades fraseológicas são definidas como combinações estáveis de pelo menos duas palavras, cujo limite superior é a oração composta, são caracterizadas pela fixação e/ou idiomaticidade (CORPAS PASTOR, 2010). Já as unidades fraseológicas especializadas são unidades sintáticas (não lexicais) de um domínio especializado, compostas por mais de um lexema sendo altamente frequentes (CABRÉ et al., 1996). Para o estudo, coletamos os primeiros seis meses de publicações do jornal, a saber de outubro de 2017 a março 2018, em português brasileiro. Empreendemos uma análise fundamentada teórico-metodologicamente na Linguística de Corpus (LC) (PARODI, 2010), com a qual articulamos os estudos de Fraseologia (CORPAS PASTOR, 2010) e Metáfora (LAKOFF; JOHNSON, 2002; BERBER SARDINHA, 2009). Na análise utilizamos o Sketch Engine (2016), um programa pago que permite a análise de textos online. Realizamos a análise em duas partes, primeiro, por meio de uma análise impressionística (BERBER SARDINHA, 2004) para a qual selecionamos cinco textos aleatórios para leitura completa, a partir dela notamos a presença de: metáforas do futebol; metáforas da guerra; e a construção da indústria ligada ao léxico das emoções. Para a segunda parte da análise, retornamos ao corpus total, com o auxílio das ferramentas Wordlist, Keywords, Concordance e Word Sketch do SE. Os principais resultados estão relacionados à metáfora conceptual ALIMENTAÇÃO É GUERRA, mapeada a partir do uso de unidades fraseológicas especializadas como conflito de interesses, sinais de advertência, ligar o alerta vermelho, sair em defesa e fazer uma defesa. Ademais, entre os resultados importantes identificamos, na leitura de linhas de concordância geradas com indústria, como essa é caracterizada a partir de emoções e sentimentos, nessa esteira, inferimos a metáfora conceptual INDÚSTRIA É ENTE HUMANO.

---

[29] Doutoranda em Estudos Linguísticos no Programa de Pós-graduação em Estudos Linguísticos da Universidade Federal de Uberlândia (PPGEL/UFU), bolsista CAPES.

Palavras-chave: Linguística de Corpus; Fraseologia; Metáfora; O joio e o trigo

AS PRÁTICAS TRADUTÓRIAS PRESENTES NO LIVRO *HARRY POTTER AND THE CHAMBER OF SECRETS*: UMA COMPARAÇÃO ENTRE AS VERSÕES BRASILEIRA E JAPONESA

Júlia Cristina VALVERDE DA SILVA[30]

A série de livros “Harry Potter”, de autoria da escritora inglesa J.K Rowling, foi publicada entre 1997 e 2007 e foi traduzida para ao menos 79 línguas. Apesar de terem como ponto inicial o mesmo texto de partida, são as normas tradutórias existentes no país e cultura de recepção que determinam as estratégias de tradução empregadas. Essas normas são restrições socioculturais particulares a uma dada cultura e período que regulam quais obras são traduzidas e de que maneira o são (Munday, 2008, p.112). Tendo isso em consideração, ao analisar comparativamente, a tradução da obra *Harry Potter and the chamber of secrets* para a língua japonesa e para o português do Brasil, buscou-se identificar as tendências tradutórias de cada obra e criar hipóteses em relação à posição ocupada por obras traduzidas e pelos tradutores nos respectivos países para poder, então, realizar generalizações acerca de estratégias de tradução predominantes em um dado gênero literário nos diferentes países. Partindo dos pressupostos de Gideon Toury (2012, p.63), para quem normas são valores ou ideias gerais compartilhadas por uma dada comunidade acerca do que é adequado ou inadequado ao se traduzir, objetivou-se depreender quais são as normas presentes nas traduções de *Harry Potter and the chamber of secrets* por meio da análise de fatores linguísticos (com o uso de um corpus paralelo) e extralinguísticos (investigação de paratextos com base no modelo de descrição de Lambert e van Gorp, 2014). Dessa maneira, a partir do estabelecimento de categorias de análise, selecionadas após a investigação das palavras mais representativas (*keywords*) do texto de partida e de suas respectivas traduções, observou-se como os itens lexicais em crivo foram traduzidos nas línguas em estudo e como essas estratégias potencialmente apontavam para comportamentos de tradução predominante nas duas culturas. As categorias analisadas foram as de “Onomásticos/Antropônimos”; “Tradução de itens lexicais relacionados ao mundo da magia” e “Nível de formalidade/*Yakuwarigo.* As estratégias de tradução encontradas em fase inicial apontam para a tendência estrangeirizante da versão japonesa que, em muitas passagens, apenas reproduz as palavras em língua inglesa com a transliteração para o silabário katakana, empregando adaptação fonológica. Na contramão dessa tendência, a versão brasileira apresentou como estratégia a criação de neologismos para itens lexicais relacionados à magia e a tradução de nomes próprios em português. A terceira categoria de análise revela como o *yakuwarigo*—conjunto de marcadores linguísticos usados para acentuar características de determinados personagens e criar estereótipos—foi utilizado como recurso narrativo e tradutória na versão japonês, adicionando, inclusive, nuances ausentes no texto de partida.

---

[30] Doutoranda do Programa de Pós-Graduação em Linguística (PPGL) da UnB.

Palavras-chave: Estudos Descritivos da Tradução; corpus paralelo; literatura estrangeira; normas de tradução.

PRONOUNS ON SOCIAL MEDIA: PRACTICES AND OTHERING

Ziyang LIU[31]

Personal pronouns and gender identities in recent years have gained considerable attention as the visibility of nonbinary gender identities is increasing especially on social media. Among these personal pronouns, singular they has been an interesting one which has led to wide-ranging discussions because of its nonbinary use that personal pronoun they is increasingly used to refer to a person with a nonbinary gender identity. The establishment of the nonbinary use of personal pronoun they has been confirmed by both American Dialect Society (ADS) and Merriam-Webster, as singular they has been voted as the word of the Decade 2010-2019 by ADS (2020) and the Word of the Year for 2019 by Merriam-Webster (2019). Both the studies of Richards and Barker (2013) and of Brown et al. (2020) have confirmed the positive effect of using correct pronouns. Referring to someone by the wrong pronouns that does not correctly reflect their gender identity, also called misgendering (Yarbrough, 2018; Brown et al., 2020; Cordoba, 2022), can be harmful to the mental health of the referee (Brown et al., 2020). Given the paucity of quantitative research on use of pronouns on social media, this study attempts to fill the gap by using a combination of corpus linguistics techniques and critical discourse analysis to investigate the innovative use of personal pronouns on social media. Especially, Twitter (now X) is selected as the source of data. This study constructs a small-sized Twitter-specific corpus and uses the twitter scraping tool named Twint to collect posts that are relevant to the discussion of pronouns and importing the preprocessed data into AntConc. By utilizing the quantitative tools in AntConc—n-gram tools, word frequency list, concordance analysis—this study finds two significant practices of personal pronouns disclosure and two discursive strategies of Othering, supported with concordances lines and discourse analysis. Two practices to express pronouns include active pronouns articulation when one takes the initiative to share one's pronouns and pronouns display when one puts one's pronouns on bio. Negative uses of pronouns include that pronouns as a identifying factor to judge and ostracize an individual and that stereotypes about nonbinary people and those who use pronouns lead to Othering. Although it is clear that the discourse on social media is constantly changing, it is of great significance to capture the dynamics of different groups when it comes to personal pronouns, and to confirm the use of pronoun practices and the existence of ostracism and stereotyping towards pronoun users. The data and findings will raise people's attention of the current situation of social media discourse and hopefully will lead to discussions on how to build pronouns-inclusive communities online.

Palavras-chave: personal pronouns; social media; corpus linguistics.

[31] Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

MULHERES E LÉXICO TABU: UMA ANÁLISE DE FRASEOLOGISMOS BASEADA EM CORPUS

Mayra Natanne Alves Marra[32]

Este trabalho é parte de uma pesquisa de doutorado em andamento, de base empírico-descritiva. Neste recorte, são apresentadas as primeiras análises que resultaram da exploração inicial de uma parte do corpus da pesquisa. O objetivo deste estudo é identificar e descrever diferentes tipos de fraseologismos em torno dos vocábulos mulher/es e do léxico tabu vagina e sinônimos, buscando investigar como foram utilizados no contexto de uma plataforma de vídeos, na internet. Assim, o corpus de estudo deste trabalho é composto por transcrições de episódios do videocast Mini Saia, publicados no canal da emissora GNT, no Youtube, sendo este um produto derivado do programa de TV Saia Justa. O corpus analisado é composto por episódios publicados entre os anos de 2020-2021 e seus respectivos comentários e abordam temas sobre os corpos femininos, feminilidades e feminismos. As participantes do programa buscam dialogar, expressando diferentes opiniões, compartilhando informações e relatando experiências. As transcrições foram realizadas com o auxílio do software Transkriptor. Neste estudo, realizamos uma análise fundamentada teórico-metodologicamente na Linguística de Corpus (LC) adotando seus princípios para a compilação do corpus, identificação, extração e análise dos dados (BERBER SARDINHA, 2009; PARODI, 2010; NOVODVORSKI, 2008). Foi utilizado o programa computacional WordSmith Tools, versão 6.0 (SCOTT, 2015) para identificação de padrões e análises, especialmente, as ferramentas “lista de palavras” e “lista de concordâncias”, assim como recursos em corpora disponíveis online para verificação de diferentes aspectos linguísticos. Esta análise também está ancorada à fundamentação teórica existente na área de Lexicologia (TAGNIN, 2013), Fraseologia (CORPAS PASTOR, 1996; 2010) e tabuísmos linguísticos (GUÉRIOS, 1979; PRETI, 2010; NOVODVORSKI E LIMA, 2020). Os resultados encontrados demonstram a criatividade lexical e apontam para um contexto propicio à utilização de tabus linguísticos, fraseologias da língua comum, inclusive aquelas do registro coloquial e vulgar e, também, demonstram que nesses contextos de utilização da língua, é frequente o uso de manipulações fraseológicas, disfemismos e eufemismos.

Palavras-chave: Mulheres; Léxico Tabu; Fraseologismos; Linguística de Corpus.

[32] Professora de Ensino Básico, Técnico e Tecnológico no Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro (IFTM), campus Ituiutaba, MG. Estudante de doutorado do programa de pós-graduação em Estudos Linguísticos (PPGEL) do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU), Uberlândia, MG. mayra@iftm.edu.br

O USO DE N-GRAMAS DE CLASSE SEMÂNTICA EM UM CORPUS DE APRENDIZ

Cristina Borges GIL[33]

O objetivo deste trabalho é analisar o uso de n-gramas de classe semântica (BERBER SARDINHA, 2023; RIBEIRO, 2023) na produção escrita e oral de aprendizes de inglês como língua estrangeira. Um n-grama de classe semântica (NGCS) é uma sequência contínua de classes semânticas, sendo que cada classe semântica expressa uma ideia ou conceito e representa uma palavra no texto (BERBER SARDINHA, 2023). Dito de outra forma, um NGCS reúne sequências de palavras que compartilham as mesmas categorias semânticas. Ou seja, as sequências de palavras abarcadas em um mesmo NGCS possuem sentido semelhante. Tal fato nos possibilita investigar uma proporção maior de agrupamentos lexicais no corpus (idem, ibidem). Com esta pesquisa, avaliamos se variação no uso dos n-gramas de classe semântica pode ser explicada pelo fato de o texto ser escrito ou falado, pela tarefa atribuída ao aprendiz, pelo seu nível de proficiência, pela sua língua materna, pela sua idade ou pelos anos de estudo do idioma inglês. O corpus empregado neste estudo foi o COREFL, cujo acrônimo significa Corpus de Inglês como Língua Estrangeira (Corpus of English as a Foreign Language), disponibilizado para a comunidade de modo gratuito por pesquisadores da Universidade de Granada, sob os termos da Creative Commons (LOZANO; DÍAZ NEGRILLO; CALLIES, 2020). Primeiramente, o corpus foi dividido em subgrupos organizados de acordo com a língua materna espanhol, alemão ou inglês. e então etiquetado com o USAS, um etiquetador semântico. Em seguida, foram extraídos e selecionados os n-gramas de classe semântica e calculada a sua chavicidade. Com essas variáveis foi feita uma análise fatorial, procedimento padrão da Análise Multidimensional (BIBER, 1988), e os fatores interpretados. Identificamos três dimensões: Dimensão 1. Cuidado, movimento, idade e interações sociais, Dimensão 2. Localização, deslocamento, idade, autoridade e emoção, Dimensão 3. Narrativa oral, marcadores de discurso, pausas preenchidas. Observamos que a tarefa e o modo desempenharam um papel importante na variação dos n-gramas de classe semântica utilizados pelos aprendizes.

Palavras-chave: Linguística de Corpus; Linguística de Corpus de Aprendiz; chavicidade; Análise Multidimensional; n-grama de classe semântica.

[33] Aluna de doutorado do Programa de Linguística Aplicada e Estudos da Linguagem PUC - SP, bolsista CAPES

CALIEMT: CORPUS DE APRENDIZES DA LÍNGUA INGLESA DO ENSINO MÉDIO TÉCNICO

Shirlene Bemfica de OLIVEIRA[34]
Lucas Renato dos Santos MURTA[35]
Jasper Vilan BRAGA[36]

O CALIEMT - Corpus de Aprendizes da Língua Inglesa do Ensino Médio Técnico é um corpus de amostras da produção oral e escrita de alunos do Ensino Médio Técnico que vem sendo coletado e compilado em aulas de inglês de um Instituto Federal no Estado de Minas Gerais. A proposta pedagógica traz uma abordagem heurística para o ensino de Inglês no âmbito da escola técnica integral onde os alunos discutem temas transversais em uma atmosfera de discussão e descoberta acadêmica e escrevem textos na língua inglesa, individualmente ou em coautoria com o uso de recursos tecnológicos (STORCH, 2005). A opção de propor a escrita em coautoria, especificamente em contextos de língua inglesa, baseia-se em aportes teóricos e pedagógicos focados em uma visão de aprendizagem socioconstrutivista e de letramentos críticos calcada nos trabalhos de Vygotsky (1978), Storch (2005) e Street (2014). Esta abordagem de Multiletramentos é promovida por meio de aulas que focam no desenvolvimento das habilidades integradas (reading, writing, listening, speaking), e na análise e produção de textos escritos de diversos registros multimodais que demonstram o posicionamento crítico dos alunos sobre temas transversais. A pesquisa visa tornar o ensino de idiomas mais orientado pelos dados (data-driven) e o processo de aprendizagem mais centrado no aluno que pode aprender por descoberta de padrões linguísticos, pelo uso do computador e suas ferramentas (Computer-Assisted Language Learning - CALL), enfatizando o pensamento crítico e a metacognição dos alunos. A compilação e organização do CALIEMT é embasada teórico e metodologicamente pela Linguística de Corpus (LC) que é a área do conhecimento que estuda a linguagem por meio da utilização do computador (BERBER-SARDINHA, 2000). Ela é definida como uma maneira de se chegar à linguagem por meio da análise dos padrões probabilísticos que se constroem nos contextos em que os falantes os empregam (BIBER et al., 1998; BERBER-SARDINHA, 2004). O projeto foi aprovado pelo Comitê Nacional de Ética na Pesquisa e todas as produções foram autorizadas pelos pais dos alunos por meio de termos de assentimento e de consentimento. A pesquisa tem grande impacto acadêmico, social, econômico e tecnológico na vida dos alunos participantes e o CALIEMT, resultado dessa pesquisa tem grande potencial inovador, uma vez que não existe no Brasil nenhum corpus eletrônico com

[34] Professora titular da Coordenadoria de Línguas Estrangeiras do IFMG - Campus Ouro Preto na área de Língua Inglesa, Pós-doutoranda no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais.

[35] Discente do Ensino Médio Técnico em administração do IFMG - Campus Ouro Preto, Ouro Preto, Minas Gerais. Bolsista PIBIF Jr. IFMG.

[36] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista FAPEMIG.

amostras de alunos do Ensino Médio Técnico de escolas federais. Este vídeo tem como objetivo apresentar a organização do corpus, a interface de algumas plataformas e os recursos online disponíveis para a compilação do CALIEMT, demonstrando a escolha dos metadados, as possibilidades de busca, os processos e as ferramentas envolvidas para a limpeza, preparação e anotação para tornar o corpus adequado para consultas linguísticas (ALUÍSIO, et. al., 2006; GONZÁLES, 2007). Os resultados podem contribuir para as pesquisas nas áreas de Educação, Linguística de Corpus e Linguística Aplicada, pois o corpus em um servidor próprio divulgado a comunidade científica externa tem grande potencial social e impacto tecnológico.

Palavras-chave: corpus de aprendizes; ensino médio técnico; língua inglesa; compilação; ensino de línguas.

NOMINALIZATION: CORPUS-BASED STUDY OF DISCUSSION SECTIONS IN FORESTRY RESEARCH ARTICLES

Shirlene Bemfica de OLIVEIRA[37]
Deise Prina DUTRA[38]
Jasper Vilan BRAGA[39]
Camila Alves RAMOS[40]
Ana Clara Taborda de Paula VICTOR[41]

The dissemination of scientific knowledge in Brazilian academic journals is predominantly conveyed through research articles (RAs) in Agricultural Sciences. However, there is a tendency for internal publishing which may influence the impact factor of national research. The low rate of international publishing may be related to the proficiency level of researchers when writing in English. Academic writing poses a challenge for Brazilian investigators and it is characterized by specific sections (abstract, introduction, methods, results, discussion, references) and linguistic features (lexical sophistication, syntactic complexity, text cohesion) which can be indicators of text quality and publishing acceptance (CROSSLEY, 2020). There has been extensive interest in RA investigations from different perspectives of form and function due to their value in generating and distributing new knowledge in the academic community (JALILIFAR, 2017). Nevertheless, few studies emphasize the separate sections of RAs even though they have shown important section variation. Such investigations generally focus on specialized corpora and terminologies, lexical combinations, and rhetorical strategies (CROSSLEY, 2020). Data discussion is the most demanding section for writers in academic writing, and variation in the noun phrase structures in this particular portion of RAs, across different disciplines, is productive. Furthermore, few corpus-based studies have examined the noun phrase structure of RAs' discussion sections, especially ones that describe the nominalization process. This study, therefore, focuses on the linguistic features of RAs, particularly the noun phrase structure in discussion sections, which are crucial for conveying complex information. It provides an overview of the noun phrase structure of discussion sections of Forestry RAs in

[37] Professora titular da Coordenadoria de Línguas Estrangeiras do IFMG - Campus Ouro Preto na área de Língua Inglesa, Pós-doutoranda no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais.

[38] Professora titular na Faculdade de Letras da UFMG na área de Língua Inglesa e no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais.

[39] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista FAPEMIG.

[40] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista CNPq.

[41] Discente do curso de Graduação da Faculdade de Letras da UFMG, Belo Horizonte, Minas Gerais, Bolsista CNPq.

high-impact journals and analyses the nominalization process in those sections. 90 RAs from nine high-impact Forestry journals were analysed and Sketch Engine was used for data exploration. Findings revealed that discussion sections in Agrarian Science RAs are predominantly written with the use of embedded noun phrases centered by nominalization suffixes which convert an action expressed by a verb into a noun, typically to refer to general statements and processes as well as to explain them or to treat actions and processes as objects separated from human participants, increasing articles impersonality. Moreover, data shows that the high frequency of this pattern in the discussion sections indicates a high level of information density, abstraction, and grammatical complexity through a specialized pattern of information packaging (JALILIFAR, 2017). The latter suggests abstraction and grammatical complexity, which may be text quality and acceptance indicators for publication. This study is part of ongoing research that aims to improve academic writing produced in English by Brazilian researchers in Agrarian Sciences and assist in the international dissemination of Brazilian scientific production. The data gathered from this research will be instrumental in enhancing the authors' academic writing skills in Agrarian Sciences. By identifying effective noun phrase structures and the nominalization processes prevalent in successful RAs, we can develop targeted workshops and resources that guide researchers in improving their writing proficiency in English. These insights will make authors aware of their writing process, and promote greater acceptance and visibility of Brazilian research. Ultimately, this initiative aims to foster a stronger global presence for Brazilian scientific contributions.

Palavras-chave: English language; academic writing; research articles; nominalization; Agrarian Science corpus.

DICIPLINARY DIFFERENCES IN FORMULATION AND PRESENTATION OS RESERACH QUESTION, HYPOTHESES AND OBJECTIVES IN INTRODUCTIONS: INSIGHTS FROM SOCIAL SCIENCES AND HUMANITIES

Anna Clara TABORDA de Paula Victor[42]
Ana Eliza Pereira BOCORNY[43]
Deise Prina DUTRA[44]
Gustavo Leal TEIXEIRA[45]
Shirlene BEMFICA de Oliveira[46]

Biber (2006) demonstrates that there is language variation across different registers in academic settings, showing some distinct linguistic features that characterize the writing of various disciplines. On the genre analysis perspective, Swales' (1990) work explores the academic writing conventions. The differences in the formulation and presentation of research questions, hypotheses, and objectives across disciplines can be understood through the light of genre and register analysis based on corpus linguistics (Biber & Conrad, 2009). This study focuses on gathering information about how different disciplines present their own research in articles. This poster, specifically, presents the investigation of the linguistic and structural differences in the presentation of the terms “research questions”, “hypotheses”, and “objectives” in article introductions across disciplines within the fields of Social Sciences (SSci), Human Sciences (HSci), and Language, Linguistics, and Arts (LLArts). We used Antconc (ANTHONY, 2024) to explore our 1600 text corpus (1,573,549 tokens) with 100 texts sampled from each of the sixteen disciplines: Law and Legal Sciences, Communication, Demography, Economy (SSci), Archeology, Anthropology, Education, Geography, Public Policies, Psychology, Political Science, Religious Faiths, Sociology, Philosophy (HSci), and Linguistics and Languages (LLArts). The frequency counts of the targeted terms were normalized per thousand words for comparison across texts (BIBER, 1988). AntConc generated the concordance lines and identified typical lexical structures used to express research questions, hypotheses, and objectives. Our findings reveal distinct disciplinary conventions, indicating how research questions, hypotheses, and objectives are articulated in introductions. For example, in the Human Sciences (HSci), the term "Hypothesis" showed a high frequency, with ≈0.68 per thousand, especially in Political Science (≈0.47 in Religion and ≈0.34 per thousand in Psychology). On the other hand, in

[42] Graduanda – Bolsista CNPQ (420180/2022-2) - Universidade Federal de Minas Gerais, Belo Horizonte – MG

[43] Professora - Universidade Federal do Rio Grande do Sul, Porto Alegre - RS

[44] Professora - Universidade Federal de Minas Gerais, Belo Horizonte - MG

[45] Professor, Universidade Federal de Minas Gerais - Montes Claros, Minas Gerais – MG

[46] Professora titular do IFMG - Campus Ouro Preto, Pós-doutoranda no Programa de Pós-Graduação em Estudos Linguísticos (PosLin/UFMG), Belo Horizonte, Minas Gerais-MG.

disciplines such as Education and Philosophy, the frequency of the term "Objective" stood out more with ≈0.24 and ≈0.21. In the Social Sciences (SSci), "Hypothesis" appeared at ≈0.41 per thousand in Economics. Meanwhile, in Linguistics, Literature, and Arts (LLArts), the term "Hypothesis" had a significant presence in Linguistics, with ≈0.8 per thousand, indicating a different methodological emphasis compared to other fields. In conclusion, the empirical nature of SSci research favors more structured and repetitive phrasing while LLArts tend to employ a more implicit use of these expressions, reflecting the diverse nature of writing conventions in these fields. Our results have pedagogical implications and will inform MOOC activities as part of the CNPq project this paper is part of. For instance, in disciplines, such as Political Science, Religious Faiths, and Psychology, the pedagogical focus should be on how to formulate hypotheses. In disciplines such as Education and Philosophy, the teaching might focus on how to define clear and achievable objectives (e.g. the use of the word “objective”), while for LLArts, presenting the use of the word "hypothesis" in various articles would be more relevant. These variations between disciplines reflect distinct methodological traditions and emphasize the need for customized pedagogical approaches to support the formulation of research questions, hypotheses, and objectives in each discipline.

Palavras-chave: Research Questions, Hypotheses, Objectives, Social Sciences, Humanities.

EXPLORING MODAL VERB USAGE IN AGRARIAN SCIENCES RESEARCH ARTICLES: A CORPUS-BASED ANALYSIS

Camila Alves RAMOS[47]
Deise Prina DUTRA[48]
Gustavo Leal TEIXEIRA[49]
Shirlene Bemfica de OLIVEIRA[50]
Carolina Godoi de Faria MARQUES[51]

Modal verbs play an important role in academic writing, offering a way to convey stance and nuance (Biber et al., 2021). However, their specific usage in Agrarian Science Research Articles (RAs) remains underexplored. This study seeks to uncover how modals are employed by researchers in Agrarian Sciences to express stance in their papers, recognizing that modals are “by far the most common grammatical device used to mark stance in university registers” (Biber, 2006, p. 103). Following the Longman Grammar of Spoken and Written English (Biber et al., 2021), modal verbs are categorized by: possibility/permission/ability, necessity/obligation, and prediction/volition. These categories show logical possibility or predictions, rarely indicating personal agency and, from this perspective, this study compared abstract, introduction, method, results, discussion and conclusion RA sections. The CorAgrarian corpus, selected for analysis, consists of 447 academic articles from five sub-areas within Agrarian Sciences. The five areas are Agriculture Engineering, Agronomy, Animal Sciences, Food Engineering, and Forestry. These research articles were compiled and chosen from specialized journals to represent a wide range of research topics within the field. All sub-corpora were uploaded on Sketch Engine. The tag for modals (MD) was counted in each sub-corpus to determine their frequency and section prevalence. Due to large differences in word count among sections, all extracted data was normalized by 1000 words. The analysis revealed distinct patterns in modal use across RA sections. Abstract sections presented a low-frequency use of modals with a normalized frequency of 4.90. Introduction sections showed a normalized frequency of 7.58. Method sections exhibited the lowest frequency of modals with a normalized frequency of 2.02. Results sections also showed a lower normalized frequency of 3.32. In contrast, the Discussion sections displayed a prominent modal use, with a normalized frequency of 10,77. Finally, the conclusion section had the highest modal frequency, with a normalized frequency of 13,31. These findings show that modals are most frequent in the conclusion section (e.g. “... such policies should be continuously promoted and extended …”), followed by the discussion (e.g. “the comprehension rates could not be considered as sufficient”) and introduction sections (e.g. “...,

[47] Aluna de graduação, UFMG, Belo Horizonte - MG, bolsista institucional PIBIC/CNPq

[48] Professora - Universidade Federal de Minas Gerais, Belo Horizonte - MG

[49] Professor, Universidade Federal de Minas Gerais - Montes Claros, Minas Gerais - MG

[50] Professora - Pós-Doutoranda. Filiação: Instituto Federal de Minas Gerais Ouro Preto - MG e Universidade Federal de Minas Gerais, Belo Horizonte - MG

[51] Doutoranda, Universidade Federal de Minas Gerais, Belo Horizonte/MG. Bolsista CAPES (n. 88887.939578/2024-00)

the varied optimistic and pessimistic versions must be contrasted.”). Following, each modal was also examined and categorized according to its function, revealing that over 50% of the modals fell into the "possibility/permission/ability" category, making it the most prevalent. According to Liu and Xiao (2022, p. 47), conclusions enable authors to emphasize their research results, highlight contributions, and suggest future directions. These communicative purposes align with modal functions, explaining their high frequency. By exploring modal frequency and distribution across different RA sections, this study deepens the understanding of stance in Agrarian Sciences academic writing. This research aims to utilize these insights to develop teaching materials that enhance academic writing skills, particularly in using modals to convey stance and engage with research findings.

Palavras-chave: modal verbs; academic writing; Agrarian Sciences; research articles; Corpus Linguistics.

O USO DE *PRESENT SIMPLE, PRESENT PERFECT, PAST SIMPLE E PAST PERFECT* NAS INTRODUÇÕES DE ARTIGOS CIENTÍFICOS, TESES E DISSERTAÇÕES ESCRITOS EM INGLÊS NA ÁREA DE CIÊNCIAS AGRÁRIAS

Jasper Vilan BRAGA[52]
Carolina Godoi de Faria Marques[53]
Deise Prina Dutra[54]
Gustavo Leal Teixeira[55]
Shirlene Bemfica de Oliveira[56]

As Ciências Agrárias são uma área importante para o desenvolvimento nacional, com um grande volume de produção acadêmica. Entretanto, essas pesquisas apresentam pouco alcance internacional, como consequência de um déficit de publicações em inglês. Para auxiliar os pesquisadores brasileiros da área a dominarem a escrita acadêmica em inglês, aumentando suas chances de publicação internacional e obtenção de um maior alcance das suas pesquisas, são necessários estudos que descrevam a escrita acadêmica em inglês dessa área. No entanto, conforme foi descrito por Shi e Wannaruk (2014) são poucos os estudos a esse respeito. De forma a contribuir com esse cenário este trabalho se propõe a analisar as estruturas verbais de past e present tense utilizadas na introdução de produções acadêmicas dessa área. Segundo Swales e Feak (2012), a introdução de artigos acadêmicos visa estabelecer o espaço da pesquisa, apresentando e contextualizando o estudo realizado, assim como seu tema, objetivos e motivações. Na introdução essas funções são geralmente assistidas pelo present simple, o present perfect e o simple past (Swales e Feak, 2012; Biber et al. 2021). Ademais, pesquisas constataram que as formas verbais: present simple, present perfect, past simple e past perfect apresentam maior frequência de uso na introdução quando comparada com as demais seções dos artigos acadêmicos, quais sejam: resumo, metodologia, resultados e discussão e conclusão (Berber Sardinha et al., no prelo). Diante do exposto, hipotetiza-se que elas tenham uma frequência significativa nessa seção também nas produções acadêmicas de Ciências Agrárias. Para realização deste trabalho, utilizamos os corpora CorAgrarian e CorAgrSc. O primeiro é um corpus de artigos científicos publicados em inglês em revistas de alto fator de impacto (A1) com 447 textos , totalizando 2.532.420 palavras, representativo das seguintes subáreas das Ciências Agrárias: Engenharia Agrícola, Agronomia, Zootecnia, Engenharia de Alimentos e Engenharia Florestal. O segundo, por sua vez, é um corpus de teses e dissertações de programas de pós-graduação brasileiros escritos em inglês, com 26 textos, totalizando 89.036 palavras, contendo textos das mesmas subáreas que o primeiro. Para a realização deste estudo foi utilizado o Sketch Engine para anotar os corpora e, para realizar as análises

[52] Aluno da Graduação, UFMG, Bolsista FAPEMIG (APQ-01173-22).
[53] UFMG
[54] UFMG
[55] UFMG
[56] UFMG

linguísticas, sua ferramenta Colocate. Visando identificar quais formas verbais são mais frequentes e seu uso nas introduções das produções acadêmicas de Ciências Agrárias, foi realizada uma a busca, por CQL, das ocorrências de present perfect, present simple, simple past e past perfect em cada corpora. Os resultados preliminares indicam que nas introduções, seja tanto dos artigos científicos quanto das teses e dissertações, as formas verbais simples tanto no passado quanto no presente apresentam uma frequência elevada em relação às demais seções. O present perfect também ocorre significativamente na introdução em todos os corpora, entretanto de forma pontual, referenciando pesquisas anteriormente realizadas. Já, o past perfect apresenta poucas ocorrências na introdução quando comparado com as demais formas verbais analisadas. Trata-se de uma pesquisa em andamento com o objetivo de auxiliar a difusão internacional da pesquisa brasileira da área de Ciências Agrárias.

Palavras-chave: linguística de corpus; Ciências Agrárias; inglês acadêmico; formas verbais; introdução.

AUTOMATIZAÇÃO COM INTELIGÊNCIA ARTIFICIAL DA EXTRAÇÃO E CLASSIFICAÇÃO DE LEXICAL FRAMES E LEXICAL BUNDLES PARA ANÁLISE DE ARTIGOS ACADÊMICOS

Simone OLIVEIRA[57]
Ana Eliza Pereira BOCORNY[58]
Júlia TAMAGNO[59]
Pedro FERNANDES[60]
Tony Berber SARDINHA[61]

A pesquisa proposta se insere no contexto do projeto geral intitulado “A internacionalização da produção científica brasileira em Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes fomentada por recursos pedagógicos on-line baseados em corpus”, coordenado pelo Professor Dr Tony Berber Sardinha. A sua investigação tem o intuito analisar artigos científicos brasileiros do período de 2013 a 2023 de produções publicados na língua inglesa em revistas internacionais de alto impacto. Para colaborar com as investigações do projeto geral, essa pesquisa visa automatizar os processos de extrair, agrupar e categorizar dados linguísticos dos artigos utilizando técnicas de Processamento de Linguagem Natural (PLN) e Inteligência Generativa. O intuito é desenvolver um sistema automatizado para agilizar a análise de dados linguísticos a partir de corpora selecionados. Com isso, será necessário buscar técnicas de PLN adequadas, construir um modelo, criar um Produto Viável Mínimo (MVP) para automatizar os processos de extração, limpeza, categorização e armazenamento de dados em escala, com menor tempo e assertividade no processo por meio de padronizações. A fundamentação teórica está estruturada em Biber (2009) e Gray e Biber (2013) que propõe duas metodologias para extração de Estruturas Lexicais (ELs). O estudo de ELs permite identificar padrões na linguagem e compreender como os autores constroem seus textos. Esse processo com uso de PLN contribuirá significativamente para a análise de corpus a partir das classificações de textos e dos modelos de similaridade, oportunizando expandir a análise da Linguística de Corpus, mesmo com corpora extremamente grandes (DUNN, 2022). O método de investigação científica mais apropriado para esse contexto é o método de pesquisa aplicada, utilizando uma abordagem quantitativa com técnicas de análise computacional e experimentação. Será necessária a organização do corpora divididos por áreas, contendo subcorpora de 1 milhão de palavras cada. A primeira ação será a extração, depois o agrupamento de pacotes lexicais por similaridade e sentido. No processo de agrupamento de nGramas, utilizaremos o modelo Sentence-BERT (BIRD, 2024). Em seguida, aplicaremos a técnica de clusterização DBSCAN, que é capaz de agrupar frases com base em sua similaridade sem exigir um número predefinido de clusters.

[57] Bolsista CNPQ (Chamada CNPq/MCTI/FNDCT No 40/2022)
[58] UFRGS
[59] UFRGS
[60] UFRGS
[61] PUCSP

Utilizando a métrica de similaridade de cosseno para calcular a proximidade entre os embeddings, o algoritmo DBSCAN formará clusters de nGramas com significado semântico similar. Esta abordagem nos permite identificar grupos de expressões semelhantes, enquanto separa os outliers (ruído), que serão excluídos da análise principal. Alguns resultados preliminares podem ser observados, como um estudo comparativo entre processos manuais e automatizados de dados já coletados no ano anterior. A relevância do estudo reside na criação de um sistema inovador de Inteligência Artificial para agilizar o tempo e melhorar os resultados no trabalho em escala de grandes grupos de corpora para extrair, agrupar e categorizar. O projeto buscará não apenas avançar na análise linguística computacional, mas também contribuir significativamente para área, permitindo uma compreensão mais profunda da produção científica brasileira em menos tempo, oferecendo insights e ferramentas que possam apoiar na elaboração e publicação de futuros trabalhos acadêmicos.

Palavras-chave: corpora; quadros e pacotes Lexicais; automação com inteligência artificial; extração; agrupamentos.

ANÁLISE MULTIDIMENSIONAL LEXICAL EM CORPORA DE RESENHAS E VIDEORRESENHAS ONLINE: UMA ABORDAGEM DA LINGUÍSTICA DE CORPUS COMO ÁREA AUTÔNOMA DE PESQUISA CIENTÍFICA

Mauricio José Ferreira LOPES[62]

Este estudo investiga as variações léxico-discursivas em corpora de resenhas escritas e videorresenhas literárias, produzidas por influenciadores digitais literários (IDLs) nas plataformas Instagram e YouTube. A pesquisa posiciona a Linguística de Corpus (LC) como uma área de investigação científica autônoma, utilizando a Análise Multidimensional (AMD) para identificar padrões linguísticos em registros distintos, com base nos métodos de Biber (1988) e Berber Sardinha (2000). Para além da análise quantitativa oferecida pela AMD, a Análise do Discurso (AD) de Pêcheux é incorporada ao estudo, a fim de interpretar as práticas discursivas, observando-se suas formações ideológicas e sociais, conforme abordado por Pêcheux (2010). A intersecção entre LC e AD permite uma análise integrada dos registros, revelando como as práticas discursivas refletem contextos sociais e como as dimensões discursivas emergentes mostram formações ideológicas subjacentes aos discursos dos influenciadores. Além disso, o uso de técnicas de Inteligência Artificial (IA) possibilita uma análise mais sofisticada de grandes volumes de dados linguísticos, oferecendo novas oportunidades para investigar os discursos produzidos em plataformas digitais, como observado por Silva (2019). O estudo examina como influenciadores literários configuram suas práticas discursivas de acordo com o público-alvo e as características das diferentes plataformas. Resenhas publicadas no Instagram tendem a apresentar uma abordagem mais introspectiva e analítica, enquanto as videorresenhas no YouTube enfatizam a comunicação direta e a interação com o público. A combinação entre LC e IA permite uma análise mais precisa de gêneros e subgêneros literários, oferecendo insights valiosos sobre as dinâmicas discursivas em plataformas digitais. A pesquisa destaca o papel fundamental da LC como ciência autônoma e interdisciplinar, que fornece uma compreensão crítica das práticas discursivas contemporâneas e suas implicações sociais e ideológicas. O estudo, assim, contribui para a consolidação da LC como uma área científica que dialoga com outras disciplinas, ampliando o escopo da análise linguística em contextos digitais e colaborando para a formação de comunidades discursivas online e a disseminação do conhecimento literário.

Palavras-chave: Palavras-chave: Linguística de Corpus; análise multidimensional; práticas discursivas; redes sociais; resenhas literárias.

[62] Professor de Língua Estrangeira na rede pública municipal de São Paulo. Mestre e doutorando em Linguística Aplicada e Estudos da Linguagem pela PUC-SP, bolsista CAPES. Email: mauricio.lopes@sme.prefeitura.sp.gov.br

EAT THE FROG: USING GENERATIVE MODELS TO AID IN THE CORPUS-BASED IDENTIFICATION OF METAPHORS IN MULTILINGUAL TWEETS

Anna Beatriz Dimas FURTADO[63]
Anne O'CONNOR[64]

Recent technological advancements gave rise to new means of communication especially valuable and profitable in the Information Society: social media. First proposed to bridge the distance between people, social media became essential for many institutions as a means of bringing their followers close. The Catholic Church is not different; the Pope has been using Twitter since 2012 to discuss a wide range of topics, from climate change to daily religious practices. The @pontifex accounts are a case of tremendous success, reaching more than 1 million followers in daily basis. Such multilingual practice is underpinned by a large-scale translation endeavour in more than thirty languages. Indeed, Corpus Linguistics has revolutionized the study of language, especially translation, enabling the identification and description of patterns across multiple languages, textual features, and several domains (O'KEEFFE and MCCARTHY, 2022). An interesting feature of human language is the pervasive use of metaphorical language, especially on the religious domain (DORST, 2021). Notably, corpus-linguistics techniques have been efficiently employed in the identification and description of metaphors (STEFANOWITSCH 2006, 2020; TISSARI, 2017). However, even using corpus exploration tools, the identification of metaphors in a big-size corpus in multiple languages can be rather time-consuming and labour-intensive.The automatic treatment of metaphors with natural language processing is not a new task. It has been investigated through several subtasks: metaphor identification (MAO et al., 2019), metaphor interpretation (SHUTOVA, 2010), conceptual mappings (ROSEN, 2018). While several models have been tested as shown in the survey by Tong et al. (2019), results show that this task remains quite challenging. One of the reasons for this is the notion of metaphor itself. Therefore, we seek to investigate whether the use of recent generative chatbot models can aid in the identification of metaphors (defined as the result of the mapping between conceptual domains as in Lakoff and Johnson (1980)) on a ten-year corpus (2012-2022) comprising 35,684 tweets (619,984 words) in seven languages (Arabic, English, Italian, French, Spanish, Portuguese). For this ongoing case study, we subsampled the corpus into a 500-tweet sample to facilitate manual analysis. We employed ChatGPT (OPENAI, 2023) and Gemini (GOOGLE, 2023) to perform metaphor detection and source-and-target domain identification in English, Portuguese, Spanish, French, Italian and Arabic. We compare automatic results with corpus-based results by employing WMatrix (RAYSON, 2008) to extract key semantic domains and their corresponding keywords so that source-and-target domains can also be identified and recorded.

---

63 Research Assistant in the Institute for Creative Technologies, University of Galway, Ireland, funded by PIETRA Project Consolidator Grant No. 101001478, European Research Council

64 Full Professor in the School of Languages, Literatures, and Cultures, University of Galway, Ireland

Our results show that metaphor detection is far from solved either by chatbots or corpus-based methods. While detecting key domains with corpus-based methods is more reliable, the task depends heavily on the quality of the reference corpus tagged for semantic fields. Although Gemini and ChatGPT can both be used to identify crystalised metaphors (65% in English and 40% in Portuguese), hallucinations are still pervasive in the source-and-target domain identification. The best approach is then, to combine both methods to facilitate the identification of metaphors.

Palavras-chave: metaphor identification; conceptual domain identification; corpus-based metaphor studies;

LINGUÍSTICA DE CORPUS E ACESSIBILIDADE: INTERFACES ENTRE CORPORA E SIMPLIFICAÇÃO TEXTUAL

Bruna Rodrigues da SILVA[65]

Este trabalho apresenta recorte, sob o viés da Linguística de Corpus, de pesquisa de Doutorado, que se insere nos estudos de Acessibilidade Textual e Terminológica (ATT). A pesquisa como um todo busca a união da experiência docente com a pesquisa acadêmica, por meio da investigação da leitura e da compreensão de materiais, em tese, adaptados para um público com doze anos ou mais, por jovens e adolescentes do Ensino Fundamental II de escola pública de Porto Alegre-RS. O objetivo principal do trabalho como um todo é descrever e analisar se um livro da área da saúde, disponível on-line, adaptado para um público leitor jovem, é compreendido por esse público e de que forma. Inicialmente, o foco será a publicação digital Aprendendo sobre vírus e vacinas, da Editora da UFCSPA. Essa editora lançou várias publicações, todas na área da saúde, adaptadas para diferentes públicos. A única dessas obras direcionada para público jovem, com doze anos ou mais, foi escolhida para análise neste estudo porque essa é a faixa etária que corresponde aos alunos da pesquisadora responsável, com os quais será possível dar continuidade à pesquisa, num próximo momento, por meio de testes de compreensão leitora. O recorte que se apresenta neste resumo faz parte do momento inicial do estudo, em que, com apoio da estatística linguística (BIDERMAN, 1978, 1998) e da Linguística de Corpus (BERBER SARDINHA, 2004), serão realizados contrastes do corpus de estudo com outros corpora. A fim de constatar possíveis diferenças de vocabulário escrito, o corpus selecionado para contraste foi a publicação digital Somos Heróis – os cuidados para o coronavírus ir embora, de Pedro Leite. Essa publicação foi seleciona porque tem vários pontos em comum com o corpus de estudo: o fato de ser adaptado; a faixa etária destinada; o acesso livre, disponível para download e a gratuidade; o formato digital; e a temática do COVID-19. A comparação será feita com o auxílio do software AntConc (ANTHONY, 2019). Esse é um software de acesso livre que contém ferramentas para gerar dados estatísticos, a partir de um texto em formato digital. Os contrastes iniciais indicam que cerca de 22% do vocabulário do corpus de estudo coincide com o vocabulário do corpus de contraste em questão. Porém, palavras como, por exemplo, analgésico, adsorver, ancorada, atenuado, papiloma, partículas, pneumocócica e proliferam, entre outras, fazem parte das diferenças entre os corpora. Assim como essas palavras, outros pontos de divergência surgirão dessas comparações, merecendo atenção, pois podem servir de base para os testes de compreensão leitora a serem realizados com os alunos nas etapas subsequentes da pesquisa. Além disso, tais contrastes também vão enriquecer a análise e a discussão sobre a acessibilidade desse material para esse público, servindo de base para o estudo como um todo.

---

[65] Doutoranda pelo PPG-Letras/UFRGS, professora da rede pública de ensino, Porto Alegre – RS

Palavras-chave: Acessibilidade; Linguística de Corpus; corpus; contraste; Linguística Computacional.

DESENVOLVIMENTO DE UMA METODOLOGIA E APRIMORAMENTOS DE RECURSOS LEXICOGRÁFICOS PARA UMA PLATAFORMA DE DICIONÁRIOS DE COLOCAÇÕES ACADÊMICAS EM PORTUGUÊS E INGLÊS

Adriane ORENHA-OTTAIANO[66]
Tanara Zingano KUHN[67]
Stella Esther Ortweiller TAGNIN[68]
Giseli Aparecida CECÍLIO[69]
Cristiane Krause KILIAN[70]

O objetivo do presente trabalho é apresentar a metodologia usada para identificar colocações acadêmicas em um corpus acadêmico de português no âmbito do projeto Dicionários Online de Colocações Acadêmicas. Embora as colocações acadêmicas tenham recebido considerável atenção nos últimos anos, revisão da literatura indica a existência de diferentes formas de entendimento acerca do que estas são (por exemplo, DURRANT, 2009; PAQUOT, 2010; ACKERMANN; CHEN, 2013). Com base nessas abordagens, e tendo em vista um posicionamento crítico, que considera também reflexões sobre vocabulário acadêmico, nosso entendimento sobre colocações acadêmicas considera dois níveis, quais seja, estatístico e fraseológico. Sob uma abordagem estatisticamente orientada, vemos as colocações acadêmicas como combinações frequentes de palavras em textos acadêmicos, cuja coocorrência é estatisticamente maior do que o esperado em comparação a quaisquer outras palavras combinadas aleatoriamente em uma língua específica e em um campo específico de conhecimento. Sob uma abordagem fraseológica, as colocações acadêmicas são combinações de palavras que são recorrentes e convencionalizadas em textos acadêmicos, que podem ter assumido um significado diferente ou novo daqueles usados na linguagem não acadêmica, podendo variar entre disciplinas. Uma vez definido nosso entendimento acerca das colocações acadêmicas, a metodologia adotada neste trabalho está estruturada nos seguintes passos. Primeiramente, criaremos uma lista de 50 palavras lexicais que ocorram com uma frequência mínima de 3em, no mínimo, 4 grandes áreas do subcorpus Brasil do CoPEP (KUHN; FERREIRA, 2020). O ponto de corte para a frequência mínima de ocorrência e o número mínimo de dispersão ainda serão definidos, uma vez que não há consenso nos estudos revisados. A seguir, serão extraídos automaticamente candidatos a colocações acadêmicas do corpus anotado com o UDPIPE. As colocações serão aquelas automaticamente calculadas pela função Word Sketch do Sketch Engine. Por fim, os candidatos serão importados para o Dictionary Writing System (ORENHA-

[66] Professora Associada do Programa de Pós-Graduação em Estudos Linguísticos, da Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo

[67] CELGA-ILTEC, Universidade de Coimbra, Portugal

[68] Universidade de São Paulo

[69] Aluna de Doutorado do Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP)

[70] Instituto Superior de Educação Ivoti

OTTAIANO ET AL., 2021; ORENHA-OTTAIANO et al., 2023), para que as colocações que integrarão o dicionário sejam selecionadas. Para tanto, os lexicógrafos deverão seguir critérios discutidos e definidos pelos membros do projeto, sob uma perspectiva fraseológica, tendo em vista ainda o usuário final do dicionário.

Palavras-chave: colocações acadêmicas; dicionário de colocações; plataforma de dicionários; recursos lexicográficos

ANÁLISE COMPARATIVA DE FERRAMENTAS DE EXTRAÇÃO TERMINOLÓGICA AUTOMÁTICAS E SEMIAUTOMÁTICAS

Helena Cid TELES OLIVEIRA[71]
Elisa Duarte TEIXEIRA[72]

A prática da tradução especializada está diretamente associada ao uso de terminologia. Teixeira (2008) propõe que unidades linguísticas que são copiadas/repetidas/imitadas/mimetizadas no texto, sejam denominadas "Unidades de Tradução Especializada" (UTEs). Mas como identificar UTEs nos textos de partida e chegada, em grandes coletâneas de textos sobre o mesmo tema de forma rápida e minimamente confiável, principalmente, com a ampliação dos meios de tecnologia e o grande volume de informações? O avanço acelerado da tecnologia ocorrido nas últimas décadas impactou veementemente o trabalho terminológico (SARDINHA, 2004), o que desencadeou a oferta de ferramentas informatizadas que facilitem o processamento de tamanho volume de dados digitalizados e de corpora eletrônicos. Ferramentas de extração automática e semiautomática identificam candidatos a termo – trabalho antes feito manualmente – com busca de equivalentes em seus textos de partida e em corpora de apoio à tradução, processo que auxilia na criação de dicionários e glossários especializados (BOWKER, 2015). No entanto, há poucas pesquisas sobre quais ferramentas de extração terminológica estão disponíveis atualmente para uso por tradutores e terminógrafos, bem como sobre seus custos, a facilidade de uso, o tipo de input requerido, o nível de eficiência dos resultados, entre outros fatores que poderiam auxiliar na decisão de usá-las ou não, e qual delas escolher. O objetivo deste trabalho foi realizar o levantamento de algumas destas ferramentas e testá-las a fim de contribuir com a comunidade científica de terminologia e o trabalho tradutório. As duas que obtiveram os melhores resultados, de acordo com os critérios definidos, foram a Termostat (DROUIN, 2010) e a Sketch Engine (KILGARRIFF et al, 2003). A depender da demanda, as ferramentas podem vir a atender o tradutor, e mais ainda se o profissional dispuser de recursos para investir nessas ferramentas.

Palavras-chave: tradução especializada; terminologia para tradução; extratores automáticos e semiautomáticos de terminologia; ferramentas de auxílio à tradução

[71] Aluna de graduação em Letras Tradução Inglês da UnB

[72] Docente do Departamento de Línguas Estrangeiras e Tradução (LET) da UnB

ANÁLISE DE ATRIBUTOS-CHAVE FOR DUMMIES: O INÍCIO DE UM MANUAL

Carolina BOHORQUEZ[73]

A pesquisa em LC acerca da escrita acadêmica é capaz de produzir resultados que podem auxiliar no ensino dessa habilidade, principalmente através da comparação das características de diferentes registros (BIBER; CONRAD, 2009). Três principais metodologias podem cumprir esse objetivo: a Análise Multidimensional (AMD) (BIBER, 1988), a Análise de palavras-chave (SCOTT, 1997) e a Análise de atributos-chave (EGBERT; BIBER, 2023). Essa última, mais recente e menos complexa, envolve itens funcionalmente relevantes a um registro e conta com cálculos estatísticos refinados. Alunos das áreas de Ciências Humanas, ao se depararem com trabalhos ricos em cálculos e análises estatísticas, sentem-se receosos e muitas vezes optam por não utilizar aquela metodologia. Um estudo realizado em 2002 concluiu que esses alunos apresentam atitudes mais negativas em relação às disciplinas de matemática e estatística (SILVA et al., 2002). O estudo enfatiza também que quanto mais o aluno compreende os conceitos básicos dessas áreas, maior será a tentativa de aproximação das mesmas. Uma vez que a LC debruça-se sobre dados numéricos, a estatística é essencial para que se possa trabalhar eficazmente com informação quantitativa (BREZINA, 2018). Este trabalho pretende, portanto, desenvolver um manual que possa auxiliar alunos de Linguística a aprenderem a utilizar a metodologia de análise de atributos-chave em suas pesquisas. O manual tem o objetivo de detalhar o passo a passo da metodologia apresentada no trabalho de Biber & Egbert (2023); apresentar exemplos de programas que possam realizar as tarefas envolvidas; demonstrar uma aplicação manual da metodologia para que ela possa ser compreendida e para que, futuramente, um script possa ser desenvolvido com o intuito de automatizar o processo; correlacionar literatura relevante; descrever um exemplo de análise de atributos-chave aplicada no âmbito da escrita acadêmica contrastando introduções de artigos científicos e introduções de teses e, por fim; explicar os cálculos estatísticos presentes na metodologia. Neste estudo, um atributo escolhido no exemplo de análise foi o tamanho das palavras. Percebeu-se que ele se dá predominantemente nas introduções de artigos. Biber (1988) argumenta que quanto maior o tamanho da palavra, maior é o peso ou densidade informacional. O uso de palavras maiores são empregadas para expressar que o texto se caracteriza por ser um foco na informação (KITJAROENPAIBOON, W. et al., 2023). Notou-se que palavras como productivity, consolidation, agricultural, effectiveness, production e security estão presentes nas introduções de artigos, enquanto que as introduções de teses apresentaram menos palavras desta natureza . Esse resultado confirma a hipótese de que introduções de artigos, por serem menores, precisam condensar as informações de maneira eficaz para que sua função de incorporar o tema principal ao estudo realizado seja executada. Diante dos resultados, acredita-se que um manual que contivesse os detalhes da metodologia de análise de atributos-chave seria extremamente útil para

[73] Mestre em Linguistica Aplicada, Universidade Federal de Minas Gerais.

alunos que pretendem adotá-la. Aplicando-se a metodologia manualmente, detalhes importantes de busca foram revelados e pretende-se listar todas as soluções e casos problemáticos na versão definitiva. Os resultados do exemplo de análise deste trabalho podem ser desdobrados e auxiliar na produção de atividades didáticas para o aluno de escrita acadêmica.

Palavras-chave: análise de registro; escrita acadêmica; manual para análise de atributos-chave; estatística para alunos de humanas; seções de introdução

CONSTRUÇÃO DE CORPORA LINGUÍSTICOS COM PYTHON E IA: EXTRAÇÃO DE DADOS DE POSTS JORNALÍSTICOS, YOUTUBE E X (TWITTER) VIA WEB SCRAPING E APIS

Wagner da Cunha NUNES[74]

Este trabalho explora a metodologia para a obtenção de um corpus linguístico utilizando ferramentas de W*eb Sscraping* em *Python* (MITCHELL, 2018, p.47), com foco em dados provenientes do YouTube, X (anteriormente conhecido como Twitter) e posts jornalísticos de opinião e política. A construção de um *corpus* é essencial para diversas pesquisas em linguística, processamento de linguagem natural (PLN) e áreas afins. De acordo com Jurafsky e Martin (2021, p. 123), "o processamento de linguagem natural envolve a interação entre computadores e linguagem humana". As plataformas de redes sociais e sites de notícias são fontes ricas de dados textuais que refletem o uso cotidiano da linguagem, sendo, portanto, ideais para esse propósito. Um *corpus* linguístico é uma coleção estruturada de textos utilizados para conduzir análises e estudos linguísticos. A relevância de um *corpus* reside na sua capacidade de oferecer uma representação ampla e diversificada do uso da linguagem em contextos reais. Redes sociais como YouTube e X, bem como sites de notícias, fornecem uma abundância de dados textuais espontâneos e variados, fundamentais para análises aprofundadas em linguística e PLN. Foram escolhidas as plataformas YouTube, X e sites de notícias devido à diversidade e volume de comentários, postagens e artigos de opinião e política. Para a coleta de dados, utilizou-se a *API* do *YouTube Data* para acessar comentários de vídeos públicos e a *API* do *Twitter* para extrair *tweets* baseados em *hashtags* e palavras-chave. A extração de textos jornalísticos foi realizada por meio de *Web Scraping* em sites de notícias, focando em artigos de opinião e política. A implementação do *Web Scraping* foi realizada utilizando bibliotecas específicas do *Python*, como *BeautifulSoup*, *Selenium* e *Scrapy.* No caso da *API* do YouTube, foi empregada a biblioteca *google-api-python-client*, que facilita a interação com os serviços do Google. Para a *API* do *X*, utilizou-se a biblioteca *Tweepy*, amplamente utilizada para interagir com a *API* do Twitter usando *Python* (BROWN, 2017, p. 89). A integração dessas bibliotecas permitiu a construção de *scripts* automatizados para a extração de grandes volumes de dados textuais. Uma vez coletados, os dados passaram por um processo de limpeza, que envolveu a remoção de duplicatas, normalização de texto, eliminação de *emojis* e caracteres especiais. Foram utilizadas bibliotecas como *Pandas* e *Re* (expressões regulares) para a manipulação e limpeza dos dados. A criação de corpora por meio da linguagem *Python*, utilizando ferramentas de *Web Scraping* e *APIs*, juntamente com a integração de técnicas de inteligência artificial (*IA*) oferecidas pelo *ChatGPT*, desenvolvido pela *OpenAI* (BROWN et al., 2020, p. 30), tornou-se mais eficiente devido à facilidade de uso dessas ferramentas, mesmo sem a necessidade de conhecimento aprofundado em programação. O *corpus* linguístico, composto por comentários de vídeos do YouTube, *tweets* e textos de posts jornalísticos de

[74] Pesquisador Independente, Uberlândia – MG wagner.nunes@ufu.br

opinião e política, proporciona uma rica base de dados para uma ampla gama de análises linguísticas e estudos de processamento de linguagem natural (PLN).

Palavras-chave: *IA*; Linguística de *Corpus*; *Web Scraping; Python* e *APIs.*

UM ETIQUETADOR PARA SINTAGMAS VERBAIS DA LÍNGUA ASURINÍ DO TOCANTIS

Luan Daniel dos Santos Sousa[75]
Thiago Blanch Pires[76]

É perceptível a necessidade de mais estudos e novas ferramentas que auxiliem nas pesquisas de línguas minorizadas, como grande parte das línguas indígenas brasileiras. Uma dessas línguas é o Asuriní do Tocantins, também conhecido como Asuriní do Trocará, do povo homônimo. Os Asurnís do Tocantins estão localizados no município de Tucuruí, no Pará, e são cerca de 500 habitantes da mesma etnia na região. Levando isso em consideração, como forma de contribuir para a revitalização linguística, este estudo visa criar um etiquetador morfossintático automático para sintagmas verbais na língua Asuriní do Tocantins a partir do corpus extraído do "Livro de Relatos Asuriní 2" utilizando-se de conhecimento do Processamento de Linguagem Natural (PLN) e diversas outras pesquisas que analisam a estrutura gramatical da língua. Manipulado computacionalmente por humanos, a linguagem de programação Python com o auxílio de três de suas bibliotecas, NLTK, spaCy e pandas, foi a ferramenta escolhida para criação do etiquetador morfossintático. Durante a criação do algoritmo para realizar a etiquetagem, houve uma tentativa de realizar o trabalho sem o uso da NLTK, usando apenas a spaCy para processamento de linguagem natural e a pandas para a análise de dados. Porém, o processo de criar as etiquetas customizadas usando a biblioteca spaCy se tornou inviável levando em consideração o tempo restante para realizar a pesquisa. Grande parte do trabalho feito com a spaCy foi aproveitado e a etiquetagem se resumiu usando NLTK e as ferramentas do pacote de Expressões Regulares do Python. Os resultados obtidos foram satisfatórios e possíveis de serem replicados e complementados por futuros pesquisadores.

Palavras-chave: Processamento de Linguagem Natural; Asuriní do Tocantins; etiquetador morfossintático; Python; sintagmas verbais.

---

[75] Graduando de Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação pela Universidade de Brasília. Artigo apresentado como Relatório Final do PIBIC Biênio 2022/2023.

[76] Professor adjunto de Línguas Estrangeiras Aplicadas, Doutor em Gestão da Informação e Idealizador do GeLinC.

COMPILAÇÃO DE CORPUS DE APRENDIZES DE ITALIANO: COLIB-Aprendizes

Angela Maria Tenório ZUCCHI[77]

Este trabalho apresenta premissas à compilação de um corpus de aprendizes de italiano, graduandos em Letras. Os corpora de aprendizes de língua estrangeira (LE) foram tema de pesquisas em vários países na primeira década dos anos 2000. Naquele período era comum a transcrição de manuscritos, apesar do uso de textos digitais produzidos por editores de textos e troca de mensagens e arquivos por e-mails na rede ADSL. Ainda não havia a facilidade do compartilhamento textual, como quando começou a se difundir o uso *drives* compartilhados em nuvem, além da comunicação em grupos *WhatsApp* e redes sociais. As pesquisas com a Linguística de Corpus (LC) sobre a produção escrita de aprendizes de LE podem se beneficiar do momento atual com o número de textos escritos disponíveis. Porém, o acúmulo de textos digitais não configura um corpus na concepção de 'corpus' para a LC, mas talvez dados que podem ser 'minerados'. Para ser um corpus, pela definição de Tagnin (2010; 2013), é necessário que, além do formato eletrônico, a coletânea de textos deva ser reunida segundo critérios específicos e adequados ao estudo pretendido. Dado o uso de textos digitais por estudantes de LE e os atuais ambientes virtuais de aprendizagem, o momento atual oferece vantagem na compilação de corpus de aprendizes de italiano seguindo critérios da LC: textos autênticos em formato eletrônico, com tratamento uniforme, balanceado e representativo (BARBERA, ONESTI, CORINO, 2007). Como corpus de aprendizes de italiano, há em rede o consultável e etiquetado VALICO, da Universidade de Turim, para o qual houve contribuição de aprendizes brasileiros, alunos desta pesquisadora. O *input* para a produção escrita eram anedotas ilustradas sobre as quais o aluno escrevia. Com a descrição da ilustração, é possível observar, quantificar e comparar fluência textual, diversidade lexical e morfossintática dos aprendizes de italiano. Para comparar a produção de aprendizes com as de italófonos, foi criado o corpus VINCA (*Varietà di Italiano di Nativi Corpus Appaiato*) com os mesmos *inputs*, as ilustrações. Desses dois corpora, Valico e Vinca, foram produzidos vários estudos (CORINO;ONESTI, 2017; CORINO,2012; CORINO;MARELLO, 2009). Na compilação do corpus de aprendizes COLIB-A (Corpora de Língua Italiana do Brasil - Aprendizes), pretende-se oferecer *inputs* específicos para a produção escrita, visando determinados gêneros textuais e circunstâncias de comunicação específicas, com ou sem auxílio de recursos como TDICs, dicionários digitais ou IA (*Chat GPT* ou *Gemini*), de forma a criar parâmetros de comparação longitudinal para a observação individual e coletiva da interlíngua dos aprendizes de italiano graduandos em Letras. Neste evento, pretende-se

[77] Docente DLM/FFLCH/USP

mostrar o exemplo de VALICO e VINCA, os possíveis modelos de *input* para a compilação do COLIB-A e organização metodológica para a coleta de textos. O COLIB-A fará parte de um projeto maior já estabelecido na universidade, o Projeto COMET, especificamente no CoMAprend (TAGNIN, 2008), e poderá ser utilizado por professores e pesquisadores de ensino de língua italiana ou de outras línguas românicas com interesse contrastivo. Espera-se que a compilação do corpus de aprendizes gere interesse entre alunos aprendizes de italiano e docentes na universidade.

Palavras-chave: corpus de aprendizes; língua italiana; interlíngua; léxico; ensino de línguas

ARTIGOS CURTOS EBRALC-2024

REVISÃO E AMPLIAÇÃO DE ÁRVORES DE DOMÍNIO A PARTIR DA ANÁLISE DE CORPUS

Amanda Letícia Valadares dos SANTOS[78]
Flávia de Oliveira MAIA-PIRES[79]

RESUMO: Árvores de domínio são ferramentas de contextualização, servindo como estrutura norteadora de quais unidades linguísticas de fato constituem termos. Nesse sentido, não basta criar uma versão inicial da árvore, também se faz necessário revisá-la e ampliá-la conforme a pesquisa avança, de modo a incluir novos termos identificados. Para tanto, convém que o terminológo utilize a Linguística de Corpus (LC) para analisar relações de frequência e coocorrência nos textos da área.

Palavras-chave: Terminologia; Árvore de domínio; Linguística de Corpus; *Sketch Engine*; Lei Geral de Proteção de Dados Pessoais.

INTRODUÇÃO

A Terminologia é a área da Linguística responsável pelo estudo dos **termos**, isto é, das unidades linguísticas utilizadas em discursos especializados, cujo significado não é conhecido por leigos. Devido a esse recorte investigativo, entende-se que a Terminologia é interdisciplinar e envolve processos de familiarização com áreas distintas das Ciências da Linguagem.

Existem muitas formas de se aproximar e de compreender conceitualmente a estrutura de um novo tema de pesquisa. Nos estudos terminológicos, cabe destacar a elaboração de árvores de domínio, definidas como “a representação, em uma forma piramidal, dos conceitos-chave de um domínio e das relações que eles mantêm entre si” (ZAFIO, 1985, p. 161, tradução nossa).

Sendo assim, a Linguística de Corpus (LC) pode ser adotada como metodologia capaz de tornar essa elaboração mais precisa e eficiente. A partir do processamento dos textos, a LC produz um resultado numérico de quais termos são mais frequentes em textos da área de especialidade, bem como com quais substantivos, verbos e adjetivos esses vocábulos estabelecem relações significativas.

Portanto, este trabalho tem o objetivo geral de investigar como a LC pode auxiliar os terminólogos na elaboração das árvores de domínio. Notadamente, como objetivo específico, busca-se promover uma análise prática de aplicação da LC nos processos de revisão e ampliação das árvores de domínio inicialmente elaboradas pelo pesquisador — de modo a checar a pertinência dos termos

[78] Mestranda do Programa de Pós-Graduação em Linguística, da Universidade de Brasília (UnB), Brasília/DF. E-mail de contato: <linguista.amandavaladares@gmail.com>.

[79] Docente do Departamento de Linguística, Português e Línguas Clássicas (LIP), da Universidade de Brasília (UnB), Brasília/DF, e pesquisadora do grupo de pesquisa da UnB/CNPq: LexiC: Ciência, projetos e pesquisa sobre léxico: <http://lexic.com.br/>.

escolhidos para integrá-la, além de quais podem ser adicionados ao diagrama inicial.

Cabe destacar, ainda, que não se intencionou uma busca exaustiva de termos a partir da análise do *corpus*. Em vez disso, este estudo buscou mapear quais técnicas e recursos da ferramenta de LC *Sketch Engine* podem ser utilizados para os fins de revisão e ampliação das árvores de domínio — com destaque para a *Wordlist*, as *Keywords*, o *Word Sketch* e a *Concordance*.

FUNDAMENTAÇÃO TEÓRICA

Existem muitos modelos de árvores de domínio possíveis, todas com base epistemológica — algumas versões são mais ontológicas, outras taxionômicas e outras próximas de mapas mentais. Devido essa estrutura, também podem ser entendidas como uma forma de arquitetar informações. Nesse sentido, contribuem para o treinamento de Inteligências Artificiais (IA), ao constituir bases simplificadas e computacionalmente interpretáveis de conhecimento (KNIGHT, 2017).

Em geral, seguindo recomendações da própria ISO 704 (2000), as árvores de domínio, vistas como sistemas conceituais, fazem parte de uma etapa prévia de aproximação do terminológo com relação ao seu objeto de pesquisa (BARROS, 2004). Entretanto, cabe destacar que a árvore de domínio não é uma ferramenta estática. No decorrer dos estudos, o terminológo encontrará novos termos e novas relações epistemológicas, de modo que é prudente revisar e atualizar constantemente a primeira versão do diagrama.

> ...a árvore de domínio permite enquadrar cada termo em algum de seus ramos ou subáreas e desse modo garante tanto a existência do termo quanto seu pertencimento ao domínio (...) cada um dos termos deve se situar em algum lugar da estrutura básica que a árvore proporciona. Se isso não for possível, o candidato a termo deverá ser considerado um caso duvidoso ou inclusive ser excluído (BARITÉ, 2016, p. 97, tradução nossa).

Diante disso, ferramentas tecnológicas, utilizadas nos estudos terminólogicos que incluem a abordagem da LC, contribuem significativamente na identificação e na alocação de termos relevantes para a árvore de domínio da área estudada. Nesse sentido, recursos que identificam frequência e colocações auxiliam na construção de uma versão mais completa e verossímil do diagrama inicial, bem como garantem maior fidedignidade com os termos usados na prática. Com isso, é possível mitigar o que Aubert (2001) chama de “risco de ruído” e “risco de silêncio”. Isto é, respectivamente, o risco de incluir termos não relacionados ao campo estudado e o risco de não se incluírem termos importantes para a análise.

Do ponto de vista da estrutura, Barité (2016) explica que árvores de domínio são formadas por um anel nuclear e outro de termos afins. Por um lado, o anel nuclear é composto por termos que indicam conceitos diretamente e unicamente associados à área de especialidade em questão. No caso desse estudo, podem-se citar “titular”, “controlador”, “operador” e “dado pessoal” como exemplos. Por outro lado, os termos do anel afim seriam aqueles que se relacionam de forma mais circunstancial com aquela área do saber — no contexto da LGPD, é possível citar os termos “segurança da informação”, “termo de uso”, “terceiros”, “fornecedores” e “prestadores de serviço”.

METODOLOGIA

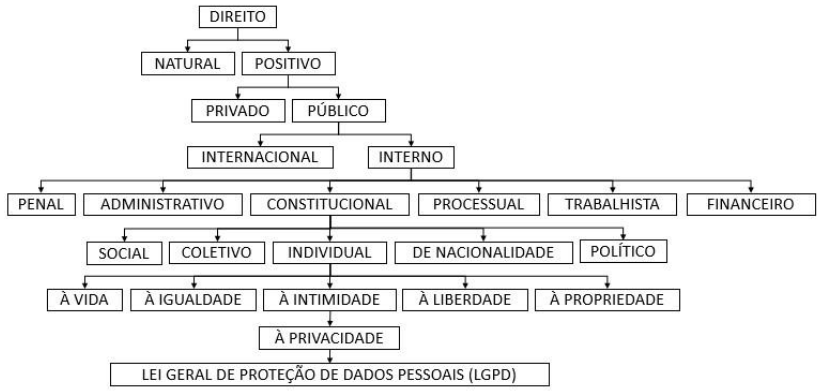
Para comprovar a necessidade de revisão e atualização de árvores conceituais em uma pesquisa terminológica, foram elaboradas árvores associadas a uma pesquisa em *corpus* com Políticas de Privacidade, leis, guias e normativos associados à Lei Geral de Proteção de Dados Pessoais – LGPD (BRASIL, 2018). Trata-se de uma lei que rege sobre a privacidade dos brasileiros, garantindo-lhes o direito de decidir o que fazer com suas informações, com quem as compartilha, com qual finalidade e por quanto tempo.

Para integrar o *corpus* dessa legislação, foram selecionados nove textos: (1) LGPD, com redação final de 2019; (2) Guia de Elaboração de Termo de Uso e Política de Privacidade para Serviços Públicos, na Versão 1.3; (3) Norma ABNT/NBR ISO 27701; (4) Política de Privacidade do BRB; (5) Política de Privacidade da Caixa; (6) Política de Privacidade do Banco do Brasil; (7) Política de Privacidade do Itaú; (8) Política de Privacidade do Bradesco; (9) Política de Privacidade do Santander. Esse *corpus* foi submetido ao *Sketch Engine* e, então, os recursos *Wordlist*, *Keywords*, *Word Sketch* e *Concordance* foram consultados para identificar possíveis candidatos a termos que se encaixariam na árvore de domínio, visando torná-la o mais completa possível.

DISCUSSÃO DOS DADOS

Em uma pesquisa da área do Direito, a primeira árvore de domínio elaborada buscou localizar como a LGPD se encaixa nessa área de especialidade, resultando em uma representação taxionômica da árvore disposta na Figura 1.

Figura 1 – Primeira árvore de domínio da LGPD

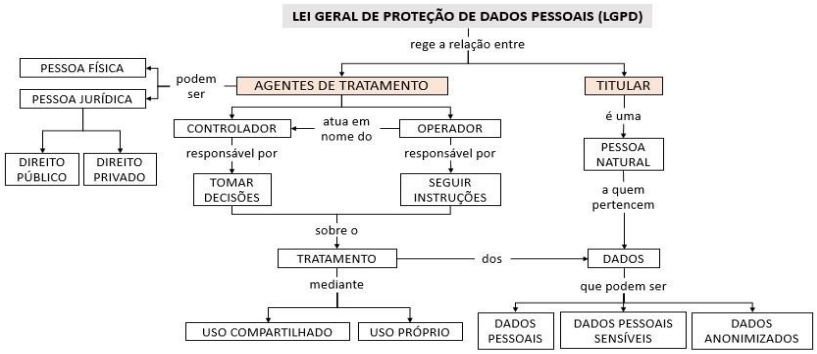


Fonte: Elaboração própria.

Uma segunda versão dessa árvore de domínio incluiu as relações específicas da própria LGPD, situando os termos da legislação em uma rede semântica. Dessa vez, a representação aproximou-se mais de estruturas ontológicas de organização das informações, conforme disposto, abaixo, na

Figura 2:

Figura 2 – Segunda árvore de domínio da LGPD



Fonte: Elaboração própria.

Após essa aproximação inicial, o *corpus* da LGPD foi submetido à ferramenta *Sketch Engine*. Em seguida, efetuou-se uma busca pela categoria "noun" no recurso *Wordlist*. Essa categoria gramatical foi a escolhida para investigação, pois a maioria dos termos são sintagmas nominais (KRIEGER, 2006). Em adição a isso, foi consultada a lista de *Keywords* da ferramenta. Trata-se de um recurso que identifica possíveis **candidatos a termo**, a partir de um cálculo comparativo de frequência — entre o *corpus* de estudo e um *corpus* de língua geral.

A Figura 3 ilustra os 15 primeiros substantivos e *keywords* (simples e complexos) identificados pelo *Sketch Engine*, de modo a demonstrar termos importantes ausentes nas árvores de domínio anteriores.

Figura 3 – Substantivos mais frequentes e keywords do *corpus* da LGPD

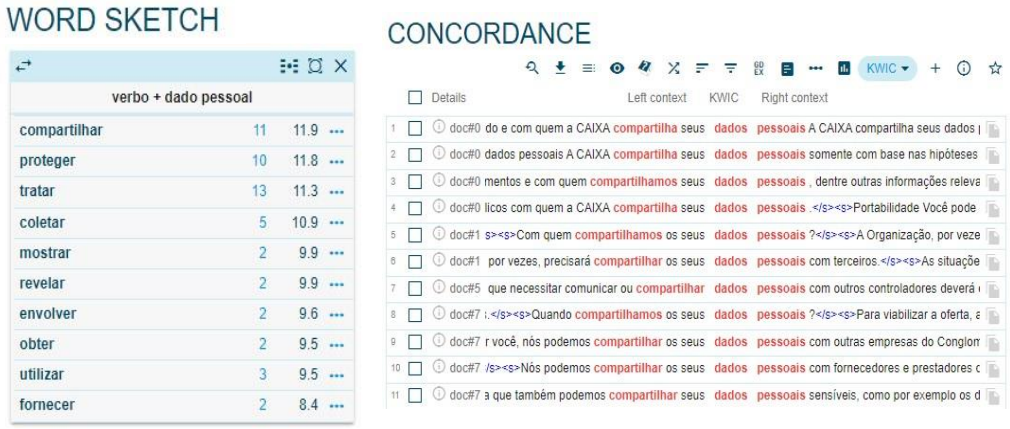
Fonte: Sketch Engine, 2024.

A partir dessa listagem, é possível notar que "dp", "informação", "organização", "serviço", "lei", "segurança", "direito", "controle", "anonimização", "subcontratado", "ANPD", "termo de uso" e "segurança da informação" são exemplos de termos que convém incluir na árvore de domínio.

Em seguida, executou-se uma busca pelos contextos de ocorrência de um dos termos relevantes da área: "dado pessoal", por meio do recurso *Word Sketch*. Essa busca revelou que "dado pessoal" estabelece relação com diversos verbos, sendo que

"compartilhar" é o mais frequente e se associa com o termo "uso compartilhado", presente na árvore da Figura 2.

Figura 4 – Verbos mais frequentes de “dado pessoal” e colocações de “compartilhar dado pessoal”



Fonte: Sketch Engine, 2024.

Uma verificação detalhada dessas ocorrências de “dado pessoal” associado a “compartilhar”, por meio do recurso *Concordance*, revelou a presença de entes de compartilhamento, como “terceiros”, “fornecedores” e “prestadores de serviço” — os quais também convém incluir na árvore de domínio elaborada.

CONCLUSÃO

Diante do exposto, comprova-se que a Linguística de *Corpus* pode ser uma abordagem metodológica muito importante na construção, revisão e ampliação de árvores de domínio durante a pesquisa terminológica. No *Sketch Engine*, ressaltam-se os recursos *Wordlist*, *Keywords*, *Word Sketch* e *Concordance*, como auxiliadores da identificação de possíveis candidatos a termos relevantes para o diagrama de representação conceitual da área.

Tais resultados contribuem para a elaboração de árvores de domínio mais completas e verossímeis com o uso real dos termos na comunicação especializada. Assim, o terminológo dispõe de mais ferramentas, além de sua leitura dos textos, para construir a esquematização conceitual de como os conceitos-chave se relacionam.

Além dos benefícios terminológicos de norteamento para a seleção e compreensão dos candidatos a termos, tal abordagem para revisão e atualização das árvores de domínio pode oferecer outras vantagens. No contexto das IAs, por exemplo, tais configurações cada vez mais densas de arquitetura da informação auxiliam no treinamento direcionado e completo de tecnologias relacionadas à linguagem em áreas de especialidade.

Agradecimentos: Agradecemos aos docentes e discentes da Universidade de Brasília (UnB) e do Grupo de pesquisa da UnB/CNPq: LexiC: Ciência, projetos e pesquisa sobre léxico, com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERÊNCIAS BIBLIOGRÁFICAS

AUBERT, Francis H. **Introdução à Metodologia da Pesquisa Terminológica Bilíngüe**. São Paulo, Humanitas Publicações-FFLCH/USP, 2001.

BARITÉ, Mario. Los árboles de dominio. *In*: CATALÁ, Sara A.; BARITÉ, Mario. (Org.). **Teoría y praxis en terminología**. 1. ed. Montevidéu: Ediciones Universitarias, Unidad de Comunicación de la Universidad de la República, 2016, v. 1, p. 91-102.

BARROS, Lídia A. **Curso Básico de Terminologia**. São Paulo: EdUSP, 2004.

BRASIL. Lei n° 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). **Diário Oficial da União**: seção 1, Brasília, DF, ano 155, n. 157, p. 59-64, 15 ago. 2018.

INTERNATIONAL STANDARD ORGANIZATION (ISO). **ISO 704**: Terminology work: principles and methods. Genebra: ISO, 2000.

KNIGHT, Michelle. Taxonomy vs Ontology: Machine Learning Breakthroughs. **Dataversity**, Los Angeles, 17 de out. de 2017. Disponível em: <https://www.dataversity.net/taxonomy-vs-ontology-machine-learningbreakthroughs/>. Acesso em: 07 jun. 2024.

KRIEGER, Maria G. Do ensino da terminologia para tradutores: diretrizes básicas. **Cadernos de tradução**, v. 1, n. 17, p. 189-206, 2006. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=4925478. Acesso em: 07 jun. 2024.

ZAFIO, Massiva N. L’arbre de domaine en terminologie. **Meta**: Journal des traducteurs, Montreal, vol. 20, n. 2, p. 161-168, jun. 1985. Disponível em: <https://www.erudit.org/fr/revues/meta/1985-v30-n2-meta308/004635ar.pdf>. Acesso em: 07 jun. 2024.

DISFLUÊNCIAS NA FALA ESPONTÂNEA DE PACIENTES COM ESQUIZOFRENIA: UMA ANÁLISE BASEADA NO CORPUS C-ORAL-ESQ

Átila Augusto Soares VITAL[80]
Bruno Neves Rati de Melo ROCHA[81]

ABSTRACT: This study compares disfluencies in the speech of patients with and without schizophrenia.The analysis uses the C-ORAL-ESQ corpus (ROCHA et al., 2020) for schizophrenic patients and CORAL-BRASIL I (RASO & MELLO, 2012). Disfluencies were examined in prosodic and pragmatic units. Results show higher occurrences in complex informational structures for both groups, with fewer reformulations by patients in sequences with more than one illocution.

Palavras-chave: disfluências; esquizofrenia; ilocuções; Prosódia; estrutura informacional.

INTRODUÇÃO

Este trabalho tem como objetivo apresentar os primeiros resultados de uma pesquisa em curso sobre disfluências em dois corpora de fala espontânea – o CORAL-ESQ (ROCHA et al., 2020), corpus que documenta a fala de pessoas com esquizofrenia durante consultas psiquiátricas, e o C-ORAL-BRASIL I (RASO & MELLO, 2012), corpus de referência do português brasileiro falado informal.

A esquizofrenia é uma doença mental com prevalência de cerca de 0,2% a 1% na população em geral, a depender de critérios diagnósticos e de grupo analisado. Os sintomas são subdivididos em positivos (como alucinações, delírios e baixo controle motor) e negativos (como anedonia, bloqueio afetivo e pobreza da fala). Muitas vezes, os sintomas linguísticos da esquizofrenia são descritos a partir da chamada desordem formal do pensamento (*formal thought disorder*), que inclui pobreza de conteúdo, falhas na expressão das informações, perda de objetivos, distração por sílabas e palavras e discurso incoerente. Nos últimos anos, há trabalhos que correlacionam os sintomas da patologia com a diminuição da variedade de unidades prosódicas produzidas pelos pacientes (disprosódia) (COVINGTON et al., 2004).

O corpus C-ORAL-ESQ tem sido compilado pelo Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL/UFMG) em cooperação com a equipe de psiquiatria do Instituto Raul Soares (IRS, Fundação Hospitalar do Estado de Minas Gerais - FHEMIG), em Belo Horizonte. Em momento oportuno, os dados serão disponibilizados eletronicamente para consultas e contarão com alinhamento texto-som (ROCHA et al, 2020).

LANGUAGE INTO ACT THEORY (L-ACT)

Assim como o C-ORAL-BRASIL I, o C-ORAL-ESQ tem sido segmentado e etiquetado em unidades informacionais segundo a *Language into Act Theory*

[80] Mestrando em Linguística Teórica e Descritiva pela Faculdade de Letras da Universidade Federal de Minas Gerais (FALE/UFMG), Belo Horizonte/MG. Contato: 4tilavital@gmail.com

[81] Professor da Faculdade de Letras da Universidade Federal de Minas Gerais (FALE/UFMG), Belo Horizonte/MG.

(L-AcT) (CRESTI, 2000; MONEGLIA & RASO, 2014; CAVALCANTE, 2016), teoria *corpusdriven* que considera que o enunciado e a *stanza* – sequências terminadas que veiculam duas ou mais ilocuções – são as unidades básicas da fala. Tanto o enunciado quanto a *stanza* veiculam atos de fala (AUSTIN, 1962) e representam unidades linguísticas com autonomia pragmático-prosódica. Nessa perspectiva, a fala espontânea implica uma constante execução de ações através da troca de ilocuções que compõem o contínuo do sinal sonoro. Segundo a L-AcT, a interpretação das ilocuções realizadas em enunciados e *stanzas* é guiada sobretudo por características do seu perfil prosódico.

As sequências terminadas podem ser de dois tipos: (i) simples, constituídas por uma única unidade tonal ou (ii) complexas, com duas ou mais unidades tonais, separadas por quebras prosódicas não terminais, que fazem com que a sequência linguística não seja autônoma pragmática e prosodicamente. Quando complexos, os enunciados são constituídos por estruturas internas que veiculam diferentes unidades informacionais, que também são distinguíveis através de formas prosódicas, função e posição em relação a outras unidades. Dentre elas, as unidades que constituem o texto da sequência e veiculam ilocuções são Comentário (COM), Comentário Ligado (COB) e Comentário Múltiplo (CMM). Há aquelas que constituem o texto, fornecendo informações sobre como interpretá-lo, sem veicular ilocuções: Apêndice de Comentário (APC), Tópico (TOP), Apêndice de Tópico (APT), Parentético (PAR) e Introdutor locutivo (INT). Por fim, há também as unidades dialógicas, que não participam da constituição do texto, mas que possuem a função de regular a interação: Alocutivo (ALL), Conector Discursivo (DCT), Incipitário (INP) e Expressivo (EXP).

As *stanzas,* que representam sequências prosodicamente terminadas de diferentes níveis de complexidade informacional, contêm mais de uma ilocução (CRESTI, 2009). Entre uma unidade ilocucionária e outra, podem haver unidades textuais e dialógicas que enriquecem os padrões informacionais.

Os fenômenos de disfluência – foco deste trabalho – são distinguidos, a princípio, em três tipos: enunciados interrompidos (marcados com “+”), quando há quebra prosódica e reformulação completa do planejamento da sequência; retractings ([/n], em que “n” é o número de palavras reformuladas), definidos como uma espécie de borracha prosódica e utilizados para reformulação de trechos do enunciado, sem que haja a interrupção completa (CAVALCANTE, 2020); e escansões (SCA), quando há a quebra do isomorfismo, e uma unidade informacional passa a ser realizada em duas ou mais unidades tonais. No exemplo 1, há um enunciado interrompido, já que, após a quebra não terminal anotada como “+”, há a reformulação completa do texto do enunciado. A barra simples (“/”) corresponde à quebra prosódica não terminal, enquanto que a barra dupla (“//”) corresponde à quebra terminal.

Exemplo 1 (bpubmn01)
*SHE: **então eu já levo meu som / eu já levo +** graças a Deus / né //

Nos exemplos 2 e 3, a seguir, há dois fenômenos de retractings, anotados com “[/2]” e “[/4]”, respectivamente. Em (2), há o cancelamento das duas palavras que constituem a primeira unidade tonal, o que faz com que elas não integrem o texto do enunciado. Nesse caso, a unidade é anotada com uma etiqueta vazia

(EMP). Em (3), por outro lado, há o cancelamento de 4 das 8 palavras que constituem a primeira unidade tonal, já que algumas delas são utilizadas para a constituição do sentido no restante do enunciado. Nesse caso, a quebra prosódica produz o retracting e uma unidade de escansão (SCA).

Exemplo 2 (bfamcv02)

*RUT: **e a [/2]=EMP=** e o meninim /=COM= né //=AUX=

Exemplo 3 (bfamcv08)

*REN: então quanto que dá **os que nũ &so [/4]=SCA=** os que sobraram //=COM=

METODOLOGIA

Através dos corpora etiquetados informacionalmente, foram realizadas buscas automáticas pelos fenômenos de disfluência descritos na seção anterior. Para a comparação entre a fala patológica e a fala não patológica, foi utilizada a metodologia apresentada por Raso et al. (2023), a partir dos estudos com os corpora C-ORALBRASIL I e C-ORAL-ESQ.

Os enunciados passaram por um pré-processamento para retirada de elementos que seriam desconsiderados para a análise. No caso do C-ORAL-ESQ, por exemplo, foram retirados os turnos de fala dos psiquiatras e dos acompanhantes, mantendo apenas a fala dos pacientes.

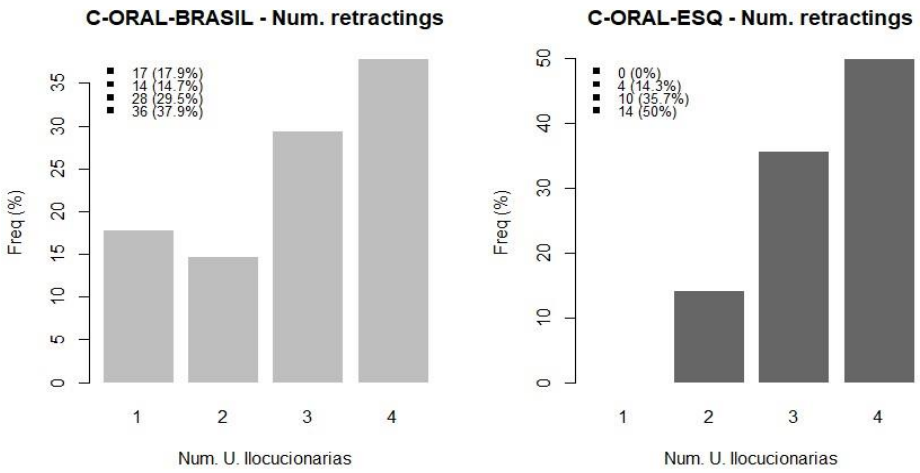
Os textos foram segmentados em sequências terminadas, que foram agrupadas em listas a partir do número de ilocuções. Tais listas comportam as stanzas de diferentes complexidades (de uma a quatro unidades ilocucionárias). Para cada um dos níveis de complexidade, foi realizada amostragem aleatória de quarenta e uma sequências terminadas, dentro das quais foram realizadas as medidas de quantidade e tipo de disfluências, posição na sequência e quantidade de palavras envolvidas.

A comparação entre os corpora se dá através da comparação da distribuição de disfluências entre os diferentes níveis de complexidade de *stanzas*, de modo que as amostras que contenham uma unidade ilocucionária no corpus de fala patológica sejam comparadas com aquelas que também contenham uma unidade ilocucionária no de fala não patológica. O mesmo procedimento é realizado para cada um dos quatro níveis de complexidade.

RESULTADOS E DISCUSSÕES

Após a coleta das amostras, foram realizadas as medidas das disfluências. Nas 41 amostras do C-ORAL-ESQ, foram encontradas 28 ocorrências de retractings, sendo que 82,9% cancelam de uma a duas palavras. No caso do C-ORAL-BRASIL, para o mesmo número de enunciados, foram encontrados 95 retractings, sendo 94,7% de uma ou duas palavras. Conforme a figura 1, ambos os corpora concentram maior quantidade de retractings em enunciados com maiores níveis de complexidade (com três e quatro unidades ilocucionárias). Aparentemente, a fala patológica conta com menor número de disfluências em enunciados menos complexos.

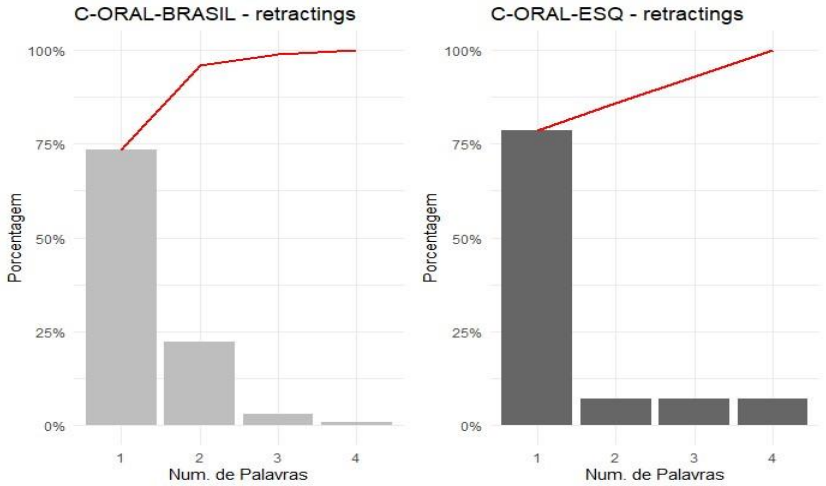
Figura 1: Gráficos sobre a quantidade de retractings presentes em cada nível de complexidade das amostras dos corpora.



Fonte: elaboração própria.

No caso do número de palavras canceladas, há diferenças entre a fala patológica e a não patológica. Na figura 2, é possível perceber que, no discurso dos pacientes, são menos frequentes retractings que cancelem mais de uma palavra.

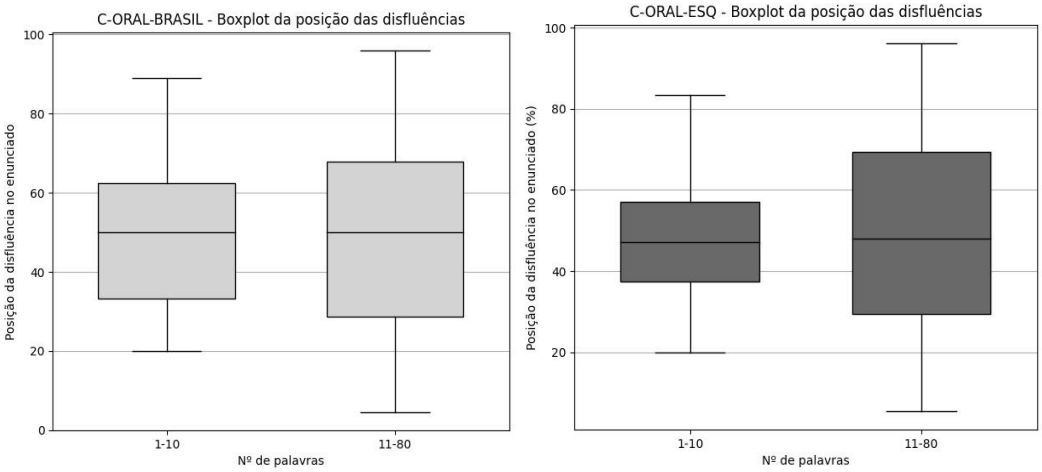
Figura 2: Gráficos sobre o número de palavras canceladas nas amostras dos corpora.



Fonte: elaboração própria.

Em relação à medida das posições das disfluências, para os dois corpora, quanto maior é o número de palavras no enunciado, mais variadas tendem a ser as posições para a ocorrência de reformulações, interrupções e escansões, conforme a figura 3.

Figura 3: Boxplots da posição das disfluências nos enunciados dos corpora. A posição é medida em termos da porcentagem do enunciado (ex.: uma escansão ocorreu em 30% de um enunciado de 10 palavras; a escansão ocorreu entre a segunda e a terceira palavra).



Fonte: elaboração própria.

Quanto ao tipo de retractings, no C-ORAL-BRASIL, são mais frequentes aqueles em que o falante cancela todas as palavras da unidade tonal. No C-ORALESQ, por outro lado, são mais frequentes os cancelamentos parciais.

Os resultados apresentados são preliminares e serão revisados à medida em que mais consultas psiquiátricas forem gravadas, transcritas e revisadas para constituição do C-ORAL-ESQ. Os próximos passos da pesquisa são a análise das unidades informacionais em que mais ocorrem as disfluências e dos enunciados em que retractings e escansões são realizados em sequências. Com a finalização do corpus e a disponibilização de mais ferramentas para o estudo da fala de pacientes com esquizofrenia, novos caminhos serão abertos para a descrição da linguagem.

REFERÊNCIAS

AUSTIN, J. L. *How to do things with words*. Oxford University Press, Oxford 1962.

CAVALCANTE, F. A., *The informational unit of topic: a crosslinguistic, statistical study based on spontaneous speech corpora*. PhD dissertation in Theoretical and Descriptive Linguistics, Universidade Federal de Minas Gerais, 2020.

CAVALCANTE, F. A., *The topic unit in spontaneous american English: a corpus-based study*. Master 's dissertation in Linguistics, Universidade Federal de Minas Gerais, 2016.

COVINGTON, M.A. et al. Schizophrenia and the structure of language: the linguist's view. *Schizophrenia research*, v. 77, n. 1, p. 85-98, 2005.

CRESTI, E., *Corpus di Italiano parlato*, Accademia della Crusca, Firenze 2000.

CRESTI, E. La Stanza: un'unità di costruzione testuale del parlato. In: *Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana*, SILFI 2008. Basilea, 30.06-03.07 2008. 2009, p. 1-25.

MONEGLIA, M., RASO, T., Notes on Language into Act Theory (L– AcT), in T. Raso, H. Mello (eds), *Spoken Corpora and Linguistic Studies*, John Benjamins Publishing Company, Amsterdam – Philadelphia 2014, pp. 468–495.

RASO, T., DE MELO ROCHA, B.N.R., SALGADO, J.V. et al. The C-ORAL-ESQ project: a corpus for the study of spontaneous speech of individuals with schizophrenia. *Language Resources and Evaluation*, p. 1-21, 2023.

RASO, T. MELLO, H. (eds), C–ORAL–BRASIL I: *Corpus de referência do Português Brasileiro falado informal*, Editora UFMG, Belo Horizonte 2012.

ROCHA, B., FERRARI, L. A., MANTOVANI, L. M., RASO, T., SALGADO, J. V., A corpus of Brazilian Portuguese speech by schizophrenic patients: preliminary observations. *Lingua e patologia: i sistemi instabili*, 2020, pp. 307-333.

RODA VIVA: UM CORPUS ORAL E A UNIVERSAL DEPENDENCIES

Cláudia Dias de BARROS[82]
Oto Araújo VALE[83]

Resumo: Neste artigo é apresentado o trabalho sobre a construção de um subcorpus composto por quatro entrevistas extraídas do Corpus Roda Viva (MIRANDA JR. et al., 2024), o qual é composto por 713 entrevistas do programa Roda Viva da TV Cultura. As quatro entrevistas trabalhadas foram transcritas automaticamente por meio da ferramenta Whisper (RADFORD et al., 2023), anotadas e revisadas com as etiquetas de Universal Dependencies.

Palavras-chave: Universal Dependencies; sintaxe; Linguística de Corpus; PLN; corpus oral.

INTRODUÇÃO

Os trabalhos na área de Processamento de Línguas Naturais (PLN) se utilizam muito de corpora a fim de comprovarem hipóteses linguísticas sobre algum fenômeno ou pesquisar novos fenômenos, por exemplo.

Neste artigo é apresentada a pesquisa realizada sobre a construção de um subcorpus do Corpus Roda Viva (MIRANDA JR. et al., 2024), que é formado por 713 entrevistas de vários anos do programa Roda Viva da TV Cultura, transcritas por jornalistas de forma textualizada, nas quais há complementações das falas, por meio de inserções textuais, informações enciclopédicas, entre outros, o que faz com que percam o status de língua oral, passando a língua escrita.

Dessa forma, a pesquisa aqui retratada tomou a decisão de construir o subcorpus com quatro entrevistas, totalizando 4024 sentenças, e, a fim de manter o status de língua oral, decidiu-se realizar a transcrição automática das entrevistas trabalhadas por meio de um ASR (Sistema de Reconhecimento Automático de Fala) chamado Whisper (RADFORD et al., 2023). Os textos transcritos apresentaram alguns problemas como transcrição equivocada de algumas palavras e erro de segmentação das sentenças, que precisaram ser corrigidos manualmente posteriormente.

A escolha das quatro entrevistas se deu baseada na possível diversidade sintática apresentada pelos quatro entrevistados, sendo eles: uma governadora, um desenhista de história em quadrinhos, um jogador de futebol e um rapper.

A partir dos textos transcritos revisados foi realizada a anotação automática com o formalismo da Universal Dependencies (DE MARNEFFE et al.,

[82] Docente do Curso de Licenciatura em Letras, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Câmpus Sertãozinho, claudias84@gmail.com.

[83] Docente do Curso de Licenciatura em Letras e Bacharelado em Linguística, Universidade Federal de São Carlos – UFSCar.

2021). Essa anotação foi realizada pelo parser PortParser (LOPES et al., 2024). Após a anotação automática, foi feita uma revisão manual por meio da ferramenta Arborator-Grew ElizIA (GUIBON et al., 2020) e foram identificados alguns fenômenos característicos da língua falada, como a presença de vocativos e marcas discursivas.

O objetivo dessa anotação é fornecer um corpus de língua oral ao projeto Porttinari (PARDO et al., 2021), um grande corpus multigênero do Português do Brasil, composto por textos escritos, como artigos de jornal, tweets do mercado financeiro brasileiro, revisões de consumidores de ecommerce e revisões de livros.

Inicialmente, serão adicionadas ao Porttinari as quatro entrevistas trabalhadas, porém, posteriormente, após o treinamento do parser com essas entrevistas, serão também anotadas automaticamente e adicionadas ao projeto as outras 709 entrevistas do Corpus Roda Viva (MIRANDA JR. et al., 2024).

FUNDAMENTAÇÃO TEÓRICA

A pesquisa apresentada neste artigo teve como arcabouço teórico a Universal Dependencies[84] (UD) (DE MARNEFFE et al., 2021), um projeto que tem como objetivo uma anotação gramatical consistente (etiquetas morfossintáticas, características morfológicas e dependência sintática), entre línguas humanas diferentes. A UD é um esforço colaborativo de cerca de 500 colaboradores que produziram quase 200 treebanks para aproximadamente 100 línguas.

Atualmente, a UD possui dezessete etiquetas morfossintáticas ou Partof-Speech (PoS) tags, como: ADJ: adjetivo, VERB: verbo e ADV: advérbio. Ela também possui 37 etiquetas de relações de dependência – deprel (de dependency relation). Uma deprel é uma relação que liga dois a dois os elementos (tokens) de uma sentença. Um deles é chamado de *head* (núcleo), que é sempre uma palavra de conteúdo (verbo, substantivo, adjetivo, pronome, numeral e advérbio) – exceções são símbolos que podem ser expressos por palavras, como R$ (reais), % (por cento) e § (parágrafo); e o outro é chamado de dependente.

Toda sentença tem uma raiz (normalmente o predicado da oração principal), marcada como dependente da deprel *root*. A atribuição de relações de dependência deve observar o princípio da projetividade, ou seja, os arcos das relações não devem se cruzar.

Alguns exemplos de etiquetas *deprel* são: amod: modificador adjetival; appos: modificador aposicional e aux: auxiliar.

Na subseção a seguir são apresentados os passos metodológicos seguidos na pesquisa.

METODOLOGIA

[84] Disponível em: https://universaldependencies.org/

Nesta subseção são apresentados os passos metodológicos seguidos até o momento para a construção do subcorpus com as quatro entrevistas retiradas do Corpus Roda Viva (MIRANDA JR, et al., 2024):

1. Escolha das quatro entrevistas que compõem o subcorpus;
2. Transcrição automática dos vídeos presentes no Youtube das quatro entrevistas, realizada pelo ASR Whisper (RADFORD et al., 2023), a qual gera um arquivo .txt;
3. Correção manual de problemas apresentados nos textos transcritos, como transcrição equivocada de palavras (principalmente estrangeirismos e dicção ruim dos entrevistados) e segmentação errada de sentenças;
4. Anotação automática com as etiquetas da Universal Dependencies (DE MARNEFFE et al., 2021), por meio do parser PortParser (LOPES et al., 2024);
5. Revisão manual da anotação automática, por meio da ferramenta Arborator-Grew ElizIA (GUIBON et al., 2020);
6. Verificação automática da anotação e revisão, por meio da ferramenta Verifica UD (LOPES et al., 2023);
7. Observação de fenômenos linguísticos característicos de um corpus de fala.

DISCUSSÃO DOS DADOS

Nesta subseção serão apresentados e discutidos os fenômenos linguísticos observados na pesquisa.

Por meio das análises das sentenças trabalhadas, puderam ser notados alguns fenômenos típicos da fala, como o uso de vocativos, como mostra a Figura 1 e de palavras discursivas, como ‘né’, apresentado na Figura 2:

Figura 1: Exemplo do uso de um vocativo. Fonte: elaborado pela autora

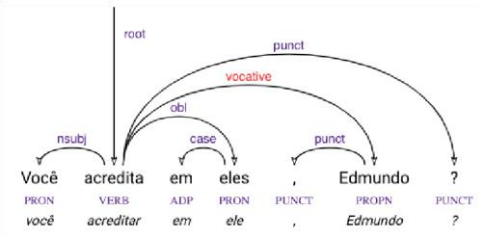
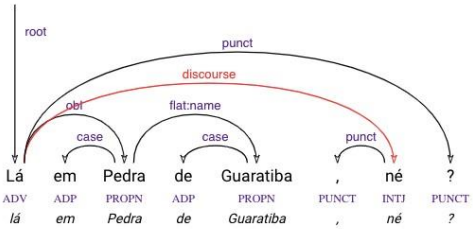
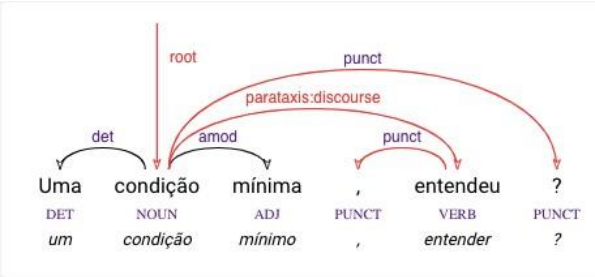


Figura 2: Exemplo do uso de uma palavra discursiva. Fonte elaborado pela autora



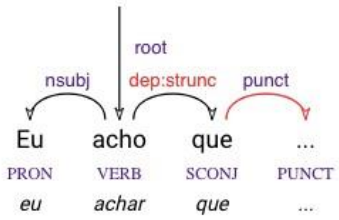
Com relação às expressões discursivas que contêm um verbo, decidiuse que seriam etiquetadas com *parataxis:discourse*, como é mostrado na Figura 3.

Figura 3: Exemplo de uso da etiqueta *parataxis:discourse.* Fonte: elaborado pela autora



Outro fenômeno bastante recorrente observado no corpus foram as sentenças quebradas, ou seja, que apresentam um truncamento, nas quais o falante não termina sua linha de raciocínio. Elas são sempre marcadas pela presença das reticências. Para anotar esse fenômeno, decidiu-se criar a subrelação *strunc* (truncamento de sentença) e adotou-se a relação *dep* (dependência não especificada), formando-se a etiqueta *dep:strunc*. Um exemplo desse tipo de anotação pode ser observado na Figura 4.

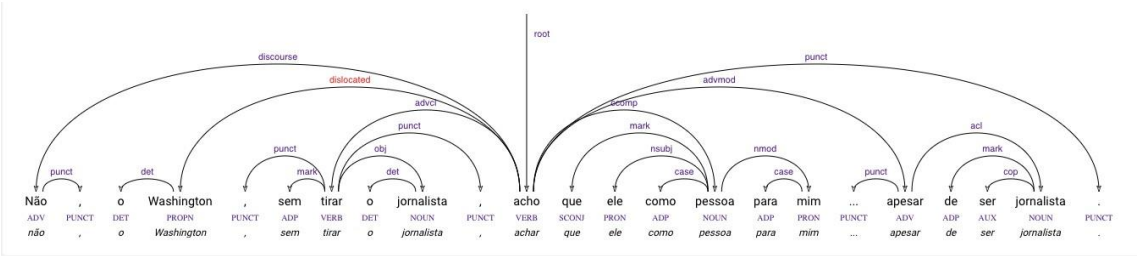
Figura 4: Exemplo de uso da etiqueta *dep:strunc.* Fonte: elaborado pela autora



Foi observada, também, a presença frequente de um outro fenômeno, o qual é relacionado à posição de sujeito das sentenças: os sujeitos deslocados, anotados com a etiqueta dislocated. Nesses casos, o falante utiliza dois sujeitos para um mesmo verbo, sendo que um deles é topificado, no início da sentença,

e o outro fica próximo ao verbo. Um exemplo desse fenômeno pode ser notado na Figura 5.

Figura 5: Exemplo de uso da etiqueta *dislocated.* Fonte: elaborado pela autora



Como se havia previsto, a entrevista com o rapper foi a que mais apresentou menor formalidade e se mostrou desafiadora para o parser anotar corretamente as relações sintáticas.

Na entrevista com a governadora do estado, observou-se uma grande quantidade de orações subordinadas e coordenadas, fruto de um discurso mais prolixo, característico de um político.

A entrevista do jogador de futebol apresentou a ocorrência de muitas etiquetas *dislocated.*

O objetivo final do trabalho é realizar o mesmo processo retratado aqui com as outras entrevistas do corpus Roda Viva, a fim de se aumentar a porção de corpus oral do projeto Porttinari (PARDO et al., 2021).

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.

REFERÊNCIAS BIBLIOGRÁFICAS

DE MARNEFFE, M. C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal dependencies. *Computational linguistics*, 47(2), pp. 255– 308, 2021.

GUIBON, G.; COURTIN, M.; GERDES, K.; GUILLAUME, B. When collaborative treebank curation meets graph grammars: arborator with a grew back-end. *Proceedings of the 12th Language Resources and Evaluation Conference,*

Marseille, France, European Language Resources Association, pp. 5293- 5302, mai. 2020.

LOPES, L.; DURAN, M. S.; PARDO, T. A. S. Verifica UD - A verifier for Universal Dependencies annotation in Portuguese. In: *Proc. of the UDFest-BR 2023*, 2023. DOI: https://doi.org/10.5753/stil.2023.25485

MIRANDA Jr., Isaac; PEDRO, Gabriela Wick; BARROS, Cláudia Dias de; VALE, Oto Araújo. Roda Viva Boundaries: an overview of an audio-transcription corpus. *Proceedings of the 16th International Conference on Computational Processing of Portuguese,* vol, 2, pp. 165-169, mar. 2024.

PARDO, Thiago Alexandre Salgueiro; DURAN, Magali Sanches;, LOPES, Lucelene; DI FELIPPO, Ariani; ROMAN, Norton T.; NUNES, Maria das Graças Volpe. Porttinari - a large multi-genre treebank for brazilian portuguese. *Proceedings of the XIII Symposium in Information and Human Language (STIL)*, pp. 1-10. nov, 29 a dez, 3. 2021.

RADFORD, A; KIM, J.W.; XU, T.; BROCKMAN, G.; MCLEAVEY, C.; SUTSKEVER, I. Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning,* PMLR 202, pp. 28492-28518, 2023.

OS (DES)ENCONTROS DA LINGUÍSTICA DE CORPUS COM A TRADUÇÃO FEMINISTA

Luciana Carvalho FONSECA[85]

Resumo: Este estudo, publicado em FONSECA (2024), explora as interseções entre a Linguística de Corpus (LC) e a Tradução Feminista, destacando a lacuna existente em pesquisas que combinem LC, tradução *e* gênero. Trata-se de uma revisão bibliográfica de 23 coletâneas e 17 números especiais de periódicos sobre gênero/feminismo/mulheres e tradução, publicados até 2022. A análise revelou uma marcada ausência de pesquisas nos Estudos Feministas da Tradução que se valem de LC. **Palavraschave:** Estudos Feministas da Tradução; Linguística de Corpus; gênero; feminismo; tradução.

Introdução

A Linguística de Corpus (LC) tem sido consistentemente empregada para analisar linguagem e gênero. Há estudos sobre gênero e diferença, gênero e linguagem, gênero e discurso, gênero e representação, gênero e terminologia etc. No que diz respeito aos estudos da tradução (ET) e gênero, a amplitude de campos e objetos de estudo é ainda maior e abrange os estudos teóricos, descritivos e aplicados da tradução. No entanto, estudos que reúnam especificamente a LC, tradução *e* gênero não têm atraído muito interesse de pesquisa.

Onde os estudos da tradução e os estudos de gênero se encontram, situa-se a tradução feminista (TF), identificada como um subcampo específico dos ET na década de 1970 no Quebec. Mais de 50 anos depois, a TF – um campo que se ocupa de gênero e tradução, mulheres e tradução etc. – tem sido praticada e teorizada por pesquisadores e pesquisadoras da tradução em todo o mundo a partir de múltiplos pontos de vista. No entanto, metodologicamente, o campo ainda se baseia predominantemente em leituras atentas (*close reading*) e métodos manuais (*hand and eye*).

Assim, para identificar, investigar e discutir como as abordagens de LC têm sido empregadas na TF, começo com um breve relato da inter-relação entre "corpus e tradução", "corpus e gênero" e "tradução e gênero". A partir de uma revisão da literatura baseada em 23 coletâneas e 17 números especiais de periódicos sobre gênero/feminismo/mulheres e tradução publicados até 2022, apresento e discuto como – e se – os ETF (Estudos da Tradução Feminista) têm se valido de métodos e ferramentas da LC.

Fundamentação teórica

O uso de corpora nos ET permite que pesquisadoras e pesquisadores testem hipóteses de tradução, identifiquem padrões e lacunas de tradução, construam e alimentem memórias de tradução e empreguem tradução

[85] Professora Doutora do Departamento de Letras Modernas, Programas de Pós-Graduação em Letras Estrangeiras e Tradução (LETRA) e Estudos Linguísticos e Literários em Inglês (ELLI), da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH/USP). E-mail: lucianacarvalhof@usp.br

automática, para citar apenas alguns pontos de entrada. Além dessa aplicação abrangente nos ET, os corpora também desempenharam um papel importante nas pesquisas sobre gênero na LC, em combinação com disciplinas como a sociolinguística e a análise do discurso. Contudo, nos estudos sobre gênero, a utilização de métodos de LC está longe de ser generalizada. Isso foi ilustrado por Paul Baker, em sua introdução ao número especial sobre LC do periódico *Gender and Language* (BAKER, 2013, p.1). O autor reuniu cinco artigos publicados anteriormente pelo mesmo periódico que empregam métodos de LC a temas de gênero. Entretanto, a tradução não foi abordada em nenhum deles.

Se pesquisas que reúnem gênero e LC têm negligenciado a tradução, as pesquisas reunindo tradução e gênero também têm negligenciado a LC. Pouco antes do número organizado por Baker, Olga Castro editou um número especial do mesmo *Gender and Language* intitulado “Gender, Language and Translation at the Crossroads of Disciplines” (CASTRO, 2013). Dos seis artigos que compõem o número de Castro, nenhum faz referência explícita aos métodos de LC, embora um deles mencione um 'corpus' de várias centenas de páginas (SANTAEMILIA, 2013).

Em suma, os números especiais organizados por Baker e Castro indicam a existência de pesquisas em gênero e LC e em gênero e ET; no entanto, sugerem uma ausência de pesquisas combinem LC, tradução e gênero. Seria de se esperar que tal combinação fosse encontrada nos ETF. Foi com o intuito de aprofundar a discussão sobre as metodologias de LC nos ETF que esta pesquisa teve início.

Metodologia

Embora apenas dois números especiais de *Gender and Language* sejam meramente indicativos de uma possível lacuna na literatura sobre o emprego de LC nos estudos sobre tradução e gênero, a revisão da literatura realizada neste estudo não só confirma esta lacuna, mas também revela as poucas instâncias em que LC, tradução e gênero foram reunidos em estudos sobre tradução e gênero/feminismo/mulheres.

A revisão da literatura baseou-se em 23 coletâneas e 17 números especiais de periódicos sobre gênero/feminismo/mulheres e tradução publicados até 2022. As 40 publicações contêm estudos que abordam o emprego de LC, em alguma medida, para estudar tradução e gênero/feminismo. Classifiquei os estudos em três grupos. O Grupo 1 é composto por 10 artigos/capítulos que adotaram uma abordagem manifestamente baseada em ou orientada por corpus. O Grupo 2 contém artigos/capítulos cuja metodologia aparentava ser informada por corpus, mas que não explicitam seus métodos. E o Grupo 3 é composto por artigos que mencionaram brevemente pesquisas de corpus em suas referências, mas não se engajam profundamente com a LC.

Discussão

Em termos de língua e edição, é interessante notar que das 10 publicações – 9 capítulos e 1 artigo – do Grupo 1, todas foram todas escritas em inglês (CAIAZZO, 2011; CALABRESE, 2011; FEDERICI; LEONARDI, 2012; FIANO; GRIMALDI, 2021; LEONARDI, 2017; PETTINI, 2020; PIZARRO SEIJAS, 2018; SALDANHA, 2003; SANTAEMILIA, 2017a; WILLIAMS CAMUS, 2018) e

publicadas em obras cujas organizadoras e organizador são principalmente da Espanha (CAMUS; CASTRO;

WILLIAMS CAMUS, 2017; CASTRO; ERGUN, 2017; FEDERICI; MACI, 2021;

GODAYOL, 2012; SANTAEMILIA, 2003; WILLIAMS CAMUS et al., 2018; ZARAGOZA

NINET et al., 2018). A distribuição por idioma e nacionalidade das organizadoras convida a discussões sob a atual tendência transnacional nos ETF (CASTRO et al., 2024), os quais têm sido produzidos predominantemente em inglês. Tal constatação nos permite questionar o papel - ou a falta do papel - do multilinguismo nas publicações e/ou colaborações transnacionais na TF.

Embora os estudos feministas da tradução tenham tradicionalmente se preocupado com a tradução literária, três textos revisados no Grupo 1 abordam a tradução especializada. Como afirmam Federici e Maci “estudos feministas recentes demonstram que a tradução especializada está se tornando um campo produtivo de estudo na tradução feminista” (FEDERICI; MACI, 2021, p. 9), e as organizadoras mencionam duas coleções de tradução feminista que incluíram estudos sobre linguagem especializada (CAMUS; CASTRO; WILLIAMS CAMUS, 2017; SANTAEMILIA, 2017b), ambas levantadas durante a presente pesquisa de revisão bibliográfica.

Um aspecto constatado sobre os estudos sobre gênero e tradução baseados em corpus que merece ser destacado é ausência de explicitação do emprego de métodos baseados em corpus e uma hesitação em nomear a LC. Apenas dois dos estudos mencionam corpus no título (PIZARRO SEIJAS, 2018; SANTAEMILIA, 2017a); dois artigos não mencionam tradução *e* gênero em seus títulos (CALABRESE, 2011; PETTINI, 2020); e, o que continua a surpreender – embora seja esperado – apenas um dos textos reivindica abertamente o título político de feminista (CASTRO; ERGUN, 2017, p. 2). O capítulo que faz essa menção é o de Santaemília (2017), que aborda especificamente o tema da terminologia no campo da TF. A falta geral de informação em títulos e palavras-chave – embora nove dos dez textos fossem capítulos e não contivessem palavras-chave individuais – explica a dificuldade inicial que tive ao reunir publicações que tratassem de corpus, tradução e gênero/feminismo/mulheres consultando bancos de dados indexados.

Assim, a partir do levantamento e análise realizados, foi constatado que os métodos e/ou ferramentas de LC têm sido empregados de forma bastante incipiente, mas que os estudos sobre gênero e tradução envolvendo linguagens de especialidade são uma promissora porta de entrada para abordagens e métodos em LC.

. Por fim, a presente pesquisa, a qual se encontra publicada na íntegra em (FONSECA, 2024) também ofereceu subsídios para se discutir as possíveis dificuldades na integração da TF com a LC, as quais podem ser resumidas da seguinte forma: a natureza quantitativa da LC e a natureza qualitativa da TF; a suposta natureza objetiva da LC e a natureza subjetiva da TF; o dilema da abordagem ‘baseada em corpus; a dependência excessiva na forma por parte da LC e a instabilidade das categorias na teoria feminista; a prevalência da

tradução literária nos ETF e a prevalência da tradução especializada na LC; e a institucionalização da teoria feminista ao lado da literatura e dos ET ao lado da linguagem especializada, onde a LC também costuma ser encontrada.

Referências

BAKER, Paul. Introduction: Virtual Special Issue of Gender and Language on corpus approaches. **Gender and Language**, *[S. l.]*, v. V1, n. Virtual Special Issue, p. 1–5, 2013. DOI: 10.1558/8psxqda5wh3d.

CAIAZZO, Luisa. Female text(ure) and science: Ada Byron's Notes and translation.
Em: PALUSCI, Oriana (org.). **Traduttrici: female voices across languages**. Trento: Tangram, 2011. p. 59–72.

CALABRESE, Rita. Living on the edge of two languages: le costruzioni possessive in In the Second Person di Smaro Kamboureli. *Em*: PALUSCI, Oriana (org.).
Traduttrici: female voices across languages. Trento: Tangram, 2011. p. 175–188.

CAMUS, Carmen Camus; CASTRO, Cristina Gómez; WILLIAMS CAMUS, Julia T. **Translation, Ideology and Gender**. Newcastle upon Tyne: Cambridge Scholars Publishing, 2017.

CASTRO, Olga (ORG.). Gender, language and translation at the crossroads of disciplines. **Gender and Language**, *[S. l.]*, v. 7, n. 1, p. 5–12, 2013. DOI: 10.1558/genl.v7il.5.

CASTRO, Olga; ERGUN, Emek (ORG.). **Feminist Translation Studies: Local and Transnational Perspectives**. New York & London: Routledge, 2017.

CASTRO, Olga; ERGUN, Emek; SPURLIN, William J.; BRACKE, Maud Anne; FONSECA, Luciana Carvalho. Transnationalizing Feminist Translation Studies? Insights from the Warwick School of Feminist Translation. **Journal of Feminist Scholarship, Special issue: Translating Transnational Feminisms**, *[S. l.]*, v. 24, n. 24, 2024.

FEDERICI, Eleonora; LEONARDI, Vanessa. Using and Abusing Gender in Translation: The Case of Virginia Woolf's A Room of One's Own Translated into Italian. **Dossier. La traducció i els estudis de gènere. Quaderns**, *[S. l.]*, v. 19, p. 183–198, 2012.

FEDERICI, Eleonora; MACI, Stefania (ORG.). **Gender Issues: Translating and** Mediating Languages, Cultures and Societies. Bern: Peter Lang, 2021. v. 281
FIANO, Carmen; GRIMALDI, Agnese Daniela. Gender Advisor, a New Role to Ensure Gender Equality Within NATO. To Translate or Not to Translate? *Em*: FEDERICI, Eleonora; MACI, Stefania (org.). **Gender Issues: Translating and Mediating Languages, Cultures and Societies**. Linguistic Insights: Studies in Language and CommunicationBern: Peter Lang, 2021. v. 281p. 199–220.

FONSECA, Luciana Carvalho. Corpora, Translation and Gender: Feminist Translation and Corpus Linguistics at the Crossroads. *Em*: LI, Defeng; CORBETT, John (org.). **The Routledge Handbook of Corpus Translation Studies**. London and New York: Routledge, 2024. p. 544–563.

GODAYOL, Pilar (ORG.). Dossier. La traducció i els estudis de gènere. **Quaderns**, *[S. l.]*, v. 19, 2012. Disponível em: https://raco.cat/index.php/QuadernsTraduccio/article/view/105017.

LEONARDI, Vanessa. Gender, Language and Translation in the Health Sciences: gender biases in medical textbooks. *Em*: CAMUS, Carmen Camus; CASTRO, Cristina Gómez; WILLIAMS CAMUS, Julia T. Williams (org.). **Translation, Ideology and Gender**. Newcastle upon Tyne: Cambridge Scholars Publishing, 2017. p. 8–31.

PETTINI, Silvia. Gender in war video games: The linguacultural representation and localization of female roles between reality and fictionality. *Em*: FLOTOW, Luise Von; KAMAL, Hala (org.). **The Routledge Handbook of Translation, Feminism and Gender**. (Eds.)London /New York: Routledge, 2020. p. 444–456.

PIZARRO SEIJAS, Paloma. Using Corpus Tools to Analyse the Rendering of Joseph Conrad's Women in the Heart of Darkness into Four Spanish Translations. *Em*: WILLIAMS CAMUS, Julia T.; GÓMEZ CASTRO, Cristina; ASSIS ROSA, Alexandra; CAMUS CAMUS, Carmen (org.). **Translation and gender. Discourse strategies to shape gender**. Santander: Cantabria University Press, 2018. p. 135–152.

SALDANHA, Gabriela. Studying Gender-Related Linguistic Features in Translated Language. *Em*: SANTAEMILIA, José (org.). **Género, lenguaje y traducción**. Valencia: Universitat de València, 2003. p. 420–432.

SANTAEMILIA, José (ORG.). **Género, Lenguaje y Traducción: actas del Primer Seminario Internacional sobre Género y Lenguaje**. Valencia: Guada Impresores, 2003.

SANTAEMILIA, José. Translating international gender-equality institutional/legal texts: The example of 'gender' in Spanish. **Gender and Language**, *[S. l.]*, v. 7, n. 1, p. 75–94, 2013. DOI: 10.1558/genl.v7i1.75.

SANTAEMILIA, José. A Corpus-Based Analysis of Terminology in Gender and Translation Research: The case of Feminist Translation. *Em*: CASTRO, Olga; ERGUN, Emek (org.). **Feminist Translation Studies: Local and Transnational Perspectives**. [s.l.] : Routledge, 2017. a. p. 15–28.

SANTAEMILIA, José (ORG.). **Traducir para la igualdad sexual: hacia una ética activa y responsable**. Granada: Editorial Comares, 2017. b.

WILLIAMS CAMUS, Julia T.; GÓMEZ CASTRO, Cristina; ASSIS ROSA, Alexandra; CAMUS CAMUS, Carmen (ORG.). **Translation and gender.**

Discourse strategies to shape gender. Santander: Cantabria University Press, 2018.

WILLIAMS CAMUS, Julia Teresa. Translation and Gender: Franco, my dear, might give damn. *Em*: ZARAGOZA NINET, Gora; MARTINEZ SIERRA, Juan José; CEREZO MERCHÁN, Beatriz; RICHART MARSET, Mabel (org.). **Traducción, género y censura en la literatura y en los medios de comunicación**. InterlinguaGranada: Editorial Comares, 2018. p. 191–204.

ZARAGOZA NINET, Gora; MARTINEZ SIERRA, Juan José; CEREZO MERCHÁN, Beatriz; RICHART MARSET, Mabel (ORG.). **Traducción, género y censura en la literatura y en los medios de comunicación**. Granada: Editorial Comares, 2018.

QUÃO CONFIÁVEIS SÃO AS FERRAMENTAS DE IA PARA A TRADUÇÃO DE RECEITAS CULINÁRIAS? ALGUMAS SURPRESAS

Stella E. O. TAGNIN[86]
Rozane R. REBECHI[87]

Introdução

Desde a proposta inovadora de Warren Weaver em 1949 (Weaver 1955 (1949)) de usar computadores para realizar traduções de forma automática, os tradutores temeram perder sua função. No entanto, ao longo dos tempos, se deram conta de como essas ferramentas podem agilizar suas tarefas, embora o produto necessite de correções e ajustes, a chamada pós-edição. A área evoluiu empregando vários sistemas até chegar à Inteligência Artificial. Em 2018 a OpenAI lançou o modelo GPT, que apresentou avanços significativos em termos de qualidade, tanto na tradução automática quanto em outras tarefas de Processamento de Linguagem Natural (PLN). Nosso objetivo é avaliar o desempenho de algumas dessas ferramentas na tradução de uma receita culinária baseando-nos em dois corpora comparáveis em inglês e português, um de Culinária Geral e outro de Culinária Brasileira.

Metodologia e aporte teórico

As ferramentas a serem analisadas são o Google Tradutor, o Microsoft Bing, o Deep-L da Microsoft e o Chat GPT da OpenAI. O objeto de investigação serão as traduções produzidas por essas ferramentas para o português e o inglês de uma das 790 receitas do livro *La Scienza in Cucina e l'Arte di Mangiar Bene* de Pellegrino Artusi (1891), considerado revolucionário na época e que pretendia popularizar as tradições culinárias das várias regiões da Itália.

Na análise proposta serão especialmente considerados termos da culinária sob a ótica das noções de adequação e aceitabilidade, conforme Toury (1995). Por adequação, entende-se uma tradução que mais se aproxima da língua de partida, enquanto a aceitabilidade privilegia uma tradução que melhor se insere na língua de chegada.

O original juntamente com as traduções publicadas em inglês (ARTUSI, 2004) e em português (ARTUSI, 2009) formam um corpus paralelo, que servirá de base para nossos comentários.

Análise

A análise das traduções será feita primeiramente para o português e em seguida para o inglês. Foi selecionada uma receita curta no formato usual do autor, neste caso com uma anedota introdutória, sem uma lista de ingredientes e incluindo um verso.

---

[86] Professora Associada, Universidade de São Paulo, São Paulo, SP
[87] Professora Adjunta, Universidade Federal do Rio Grande do Sul

A receita original, em italiano, é a seguinte:

276. PICCIONI IN UMIDO
A proposito di piccioni sentite questa che vi do per vera, benché sembri incredibile, e valga come riprova di ciò che vi dicevo sulle bizzarrie dello stomaco
Una signora prega un uomo, che le capita per caso, di ucciderle un paio di piccioni, ed egli, lei presente, li annega in un catino d'acqua. La signora ne ricevé una tale impressione che d'allora in poi non ha più potuto mangiar la carne di quel volatile.
Guarnite i piccioni con foglie di salvia intere, poneteli in un tegame o in una cazzaruola sopra a fettine di prosciutto grasso e magro e conditeli con olio, sale e pepe. Quando essi avranno preso colore, aggiungete un pezzo di burro e tirateli a cottura con brodo. Prima di ritirarli dal fuoco spremeteci sopra un limone e adoperate il loro sugo per servirli con fette di pane arrostito postevi sotto. Avvertite di salarli pochissimo a motivo del prosciutto e del brodo. Al tempo dell'agresto,potete usare quest'ultimo invece del limone, seguendo il dettato:
Quando Sol est in leone
Bonum vinum cum popone
Et agrestum cum pipione

Apesar de a introdução à receita ser bastante ilustrativa do estilo do autor, por limitação de espaço, e pelo fato de não estar diretamente ligada ao gênero ‘receita culinária’, não será aqui analisada.

Vejamos as traduções dos títulos em português:

GoogleTranslator	Microsoft Bing	Deep-L	ChatGPT
POMBOS NO MOLHADO	POMBOS COSTURADOS	POMBOS ESTUFADOS	POMBOS EM MOLHO

Todas as ferramentas traduziram o ingrediente principal de forma correta, contudo não houve concordância em relação ao modo de preparo da ave. O Google Translator propôs uma tradução inadequada para uma receita culinária, mas fazendo referência à suculência do prato; o Microsoft Bing ofereceu uma tradução equivocada do termo *in umido*; o Deep-L forneceu um equivalente adequado na variante portuguesa, ‘estufado’, apesar de ter sido selecionada a variante brasileira para a pesquisa; já o ChatGPT apresentou uma tradução adequada, do ponto de vista da culinária. Contudo, de acordo com pesquisa nos corpora citados, observamos que a fraseologia convencional se dá com ‘ao’ e ‘com’, em geral acompanhada da preposição ‘de’, para fazer referência ao ingrediente principal desse molho, como, por exemplo, ‘ao/no molho de laranja’. Quando não se faz referência ao tipo de molho, costuma-se adotar o termo ‘ensopado’, estratégia utilizada na tradução publicada – ‘pombos ensopados’. Contudo, os corpora nos mostraram que é comum utilizar a forma singular, ainda que se refira a mais de uma unidade. Assim, julgamos que uma tradução adequada para o título da receita seria ‘pombo ensopado’.

A receita em análise não apresenta uma lista de ingredientes, como as receitas contemporâneas (TAGNIN, REBECHI e TEIXEIRA, 2022). Os ingredientes são incluídos diretamente no preparo do prato, sem menção às quantidades necessárias.

As ferramentas utilizadas mantiveram o formato original de apresentação dos ingredientes na tradução para o português.

Quanto à tradução dos recipientes adequados para a cocção – *tegame* ou *cazzaruola* –, as sugestões foram:

Google Translator	Microsoft Bing	Deep-L	ChatGPT
tacho ou tacho	panela ou lixo	panela ou caçarola	panela ou caçarola

Observamos, portanto, que a primeira simplesmente repete o utensílio, possivelmente por não ter conseguido identificar uma diferença entre eles. A segunda oferece para o termo *cazzaruola* uma tradução absolutamente equivocada e inadequada para o gênero, enquanto as duas últimas, além do termo genérico 'panela', propõem também 'caçarola', tipo de panela com duas alças laterais. Vale ressaltar que essas duas traduções, apesar de fiéis ao texto de partida, desconsideram que as receitas em português brasileiro costumam ser menos detalhadas (REBECHI, 2015), apoiando-se no senso comum do leitor para decidir qual o tipo apropriado de recipiente usar. Na versão publicada, os termos *tegame* e *cazzaruola* foram traduzidos por 'frigideira' e 'caçarola', respectivamente.

Passemos às traduções dos títulos para o inglês:

Google Translator	Microsoft Bing	Deep-L
WET PIGEONS	STEWED PIGEONS	STEWED PIGEONS

Nota-se que o Google Translator fez uma tradução literal, totalmente inadequada do ponto de vista culinário. As duas outras traduções são perfeitamente adequadas.

A primeira tradução produzida pelo Chat GPT (aqui numerada como 0) causou tanta surpresa que foram pedidas outras, mas essa parte será discutida mais adiante. É apenas mencionada aqui para explicar porque há, na realidade, cinco traduções do Chat GPT.

Chat GPT 0	Chat GPT 1	Chat GPT 2	Chat GPT 3	Chat GPT 4
PICCIONI IN UMIDO (Stewed Pigeons)	276.PICCIONI IN UMIDO (STEWED SQUABS)	STEWED PIGEONS	STEWED PIGEONS	STEWED PIGEONS

Quanto ao título, observa-se que nas versões Chat GPT 0 e Chat GPT 1 ele foi mantido em italiano e a tradução dada entre parênteses, num claro esforço de adequação, por manter o título no original, mas também de aceitabilidade, para se aproximar do público leitor.

As outras versões apresentam apenas a tradução *Stewed Pigeons*. Cabe ressaltar, entretanto, que a tradução norte-americana publicada traz como título *Stewed Squabs*. Enquanto *pigeons* são mesmo 'pombos', *squabs* são pombos jovens que ainda não desenvolveram penas. No português, salvo engano, essa diferença não é lexicalizada.

O que realmente chamou nossa atenção foi a primeira tradução do Chat GPT, que apresentou os ingredientes em formato de lista:

Ingredients:
- Whole sage leaves
- Pigeons
- Slices of fatty and lean ham (prosciutto)
- Olive oil
- Salt and pepper
- Butter
- Broth
- Lemon

Na receita em italiano, como se observa acima, os ingredientes são mencionados no decorrer da explicação de como elaborar o prato. Pareceu-nos surpreendente a ferramenta 'extrair' os ingredientes do texto e apresentá-los no formato usual de uma receita contemporânea. O mesmo ocorreu com o modo de preparo, que também foi apresentado em formato de lista.

Foi essa tradução que nos levou a solicitar mais algumas ao Chat GPT para verificar se o fenômeno se repetia, mas as outras quatro versões mantiveram o texto corrido. Por outro lado, na comparação dessas versões com a versão publicada em livro, a ChatGPT 1 mostrou-se idêntica à publicada, sem qualquer indicação – nem referência – de que essa tradução seja a publicada.

Na parte referente ao modo de preparo há algumas 'traduções' que nos chamaram a atenção. Em italiano, Artusi indica uma *tegame o [...] una cazzaruola* para cozinhar a ave. As opções das respectivas ferramentas foram:

Google Tradutor	Microsoft Bing	DeepL	ChatGPT 0	ChatGPT 1	ChatGPT 2	ChatGPT 3	ChatGPT 4
saucepan or saucepan	a pan or a garbage	a pan or trough	pot or casserole dish	pot or saucepan	pot or casserole	pan or casserole dish	pan or casserole dish

Desnecessário salientar a inadequação tanto das traduções do Google Tradutor, que simplesmente repete *saucepan*, quanto do Microsoft Bing, que traduz *cazzaruola* por *garbage* ('lixo'). A tradução do Deep-L também é inaceitável, uma vez que *trough* designa um 'cocho', ou seja, onde é colocado o alimento para os animais.

Dentre as opções oferecidas pelo ChatGPT, as mais adequadas nos parecem ser as das versões 3 e 4, *pan or casserole dish*, porém todas são aceitáveis. Como a ave, depois de dourada, deve ser cozida com caldo, justifica-se *pot* por ser um recipiente mais fundo, e mesmo *saucepan*, uma panela comum.

Considerações Finais

A análise salientou a inadequação das traduções do Google Tradutor em todos os itens analisados nas duas línguas. O Microsoft Bing apresentou resultado similar, com exceção do título em inglês, que foi adequado. O Deep-L teve desempenho satisfatório, exceto no título em português, quando o produziu na variante portuguesa. As surpresas ficaram a cargo do ChatGPT que, na

versão 0, extraiu a lista de ingredientes de um texto corrido, assim como apresentou o modo de fazer em tópicos. Outra surpresa foram os títulos traduzidos para o inglês. A versão 0 manteve o título original e acrescentou a tradução Stewed Pigeons, a versão 1 trouxe o título original com a tradução Stewed Squabs, exatamente como a tradução publicada. Outras ferramentas, como o Gemini e o Copilot, serão discutidas na apresentação oral.

Referências

ARTUSI, Pellegrino. *A Ciência na Cozinha e a Arte de Comer Bem*. Tradução de Marusca Oliva Bertolozzi e Anabela Cristina Costa da Silva Ferreira. Salto e Itu, SP:
Associação Emiliano Romagnoli Bandeirante, 2009.

----- *La Scienza in Cucina e l'Arte di Mangiar Bene*. Firenze: Salvadore Landi, 1891. ----- *Science in the Kitchen and the Art of Eating Well*. Toronto, Buffalo, London: University of Toronto Press, 2004.

REBECHI, Rozane Rodrigues. *A Tradução da Culinária Típica Brasileira para o Inglês: um estudo sob o Enfoque da Linguística de Corpus*. 2015. 393 p. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo.

TAGNIN, Stella E. O.; REBECHI, Rozane R.; TEIXEIRA, Elisa D. A fraseologia das receitas culinárias – com destaque para as brasileiras. In NOVODVORSKI, A.; BEVILACQUA, C. (org.) *Fraseologia: enfoques especializados e contrastivos* Uberlândia: Universidade Federal de Uberlândia, p. 441-473, 2022.

TOURY, Gideon. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins, 1995.

WEAVER, W. Translation. In LOCKE, W. N. and BOOTH, A. D. (eds.) *Machine translation of languages: fourteen essays.* Cambridge, Mass.: Technology Press of The Massachusetts Institute, p. 15-23. 1955 (1949).

UD_NHEENGATU-COMPLIN: O CORPUS SINTATICAMENTE ANOTADO DO NHEENGATU DA COLEÇÃO *UNIVERSAL DEPENDENCIES*

Leonel Figueiredo de ALENCAR[88]

ABSTRACT: This paper introduces the latest version of UD_Nheengatu-CompLin, the treebank of the Universal Dependencies collection for Nheengatu, a Brazilian Indigenous language threatened with extinction. It is the largest and has the highest evaluation grade among the 21 Amerindian languages in this collection.

Palavras-chave: linguística computacional; nheengatu; *treebank*; tupinologia; *parsing* sintático.

A linguística de corpus e o processamento de linguagem natural (PLN) desenvolveram-se muito nos últimos 20 anos no Brasil. Essas disciplinas, contudo, têm focado entre nós quase que exclusivamente o português e outras línguas majoritárias, ignorando a enorme diversidade de línguas indígenas, salvo iniciativas mais recentes, como Galves *et al.* (2017) e Rodríguez *et al.* (2022). Este trabalho relata sobre o atual estágio de um esforço iniciado há três anos de inclusão digital do nheengatu ou Língua Geral Amazônica (LGA), por meio da construção do UD_Nheengatu-CompLin, o primeiro *corpus* sintaticamente anotado (*treebank*) dessa língua, visando tanto investigações linguísticas computacionalmente embasadas quanto a implementação de um *parser* sintático neural.

A LGA foi língua oficial do Maranhão e Grão-Pará de 1689 a 1727 e, até meados do século XIX, sobrepujava o português na região norte (RODRIGUES, 1996; CRUZ, 2011, 2015; FREIRE, 2011; MOORE, 2014; NAVARRO, 2016). Atualmente, encontra-se ameaçada de extinção, não obstante 6000 falantes no município brasileiro de São Gabriel da Cachoeira e 8000 na Colômbia (EBERHARD; SIMONS; FENNIG, 2024). Várias características a distinguem no quadro das línguas indígenas brasileiras. Em primeiro lugar, não se restringe a uma única etnia. Em São Gabriel da Cachoeira (AM) é a língua materna dos barés, uarequenas e baníuas, originalmente de línguas aruaques. Nunca foi língua tribal, tendo emergido do tupinambá, uma das variedades do tupi, pela sua utilização como língua geral por portugueses e seus descendentes mestiços e membros de inúmeras etnias incorporadas ao sistema colonial. É a principal língua adotada em iniciativas de (re)vitalização em diversas regiões do país como meio de afirmação de identidade étnica, contando ainda com um crescente número de traduções de clássicos da literatura universal realizadas por não indígenas (NAVARRO; AVILA; TREVISAN, 2017).

Todos esses fatores concorreram para tornar o nheengatu a língua indígena brasileira, segundo parece, com o maior volume de registros escritos, que permitem acompanhar seu desenvolvimento histórico desde o século XVII,

[88] Professor Titular do Departamento de Letras Estrangeiras e do Programa de Pós-graduação em Linguística da Universidade Federal do Ceará, Fortaleza, CE.
E-mail: leonel.de.alencar@ufc.br

com uma produção notável na segunda metade do século XIX e início do século XX. Não obstante isso, antes da iniciativa objeto deste trabalho, não havia nenhum *corpus* sintaticamente anotado do nheengatu. Além de um certo descaso da área de PLN pelas línguas indígenas brasileiras, desprovidas de apelo comercial, vários outros fatores contribuíram para esse estado de coisas. Em primeiro lugar, constitui um empecilho ao desenvolvimento de recursos e ferramentas de PLN a grande disparidade de ortografias com que a LGA tem sido registrada ao longo dos séculos. Além disso, a maior parte das publicações só está disponível em papel ou em arquivos que demandam transcrição manual.

Desse quadro resultou uma situação de estagnação que perdurou até recentemente: sem ferramentas de PLN, a língua não dispunha de *corpora* anotados, o que, por sua vez, impedia o treinamento de modelos por meio de aprendizado de máquina supervisionado. Para romper esse ciclo, iniciamos há cerca de quatro anos um projeto de construção de recursos computacionais para o nheengatu, que culminou na implementação, em 2022, do Yauti, um analisador sintático baseado em regras (ALENCAR, 2023), utilizado na anotação do UD_Nheengatu-CompLin.

O UD_Nheengatu-CompLin conforma-se ao modelo Dependências Universais (doravante UD) (MARNEFFE *ET AL.*, 2021), aparentemente o mais difundido para anotação sintática de *corpora*. De fato, a coleção UD cresceu de 10 *treebanks* de 10 línguas na versão 1.0, de 15.01.2015, para 283 *treebanks* de 161 línguas na versão 2.14, de 15.05.2024, que inclui 21 *treebanks* de línguas ameríndias, 14 das quais do Brasil. Uma razão da popularidade de UD é seu foco tanto no processamento computacional quanto na tipologia linguística. Outra vantagem decisiva é a disponibilidade de uma variada gama de ferramentas gratuitas para edição, visualização e manipulação de *treebanks*, treinamento de analisadores sintáticos neurais (STRAKA; STRAKOVÁ, 2017) etc.

Entre os 21 *treebanks* de línguas ameríndias da versão mais recente da coleção UD, o UD_Nheengatu-CompLin sobressai em diversos parâmetros quantitativos. Possui, por exemplo, 27,74% mais palavras do que o UD_Mbya_Guarani-Dooley, o segundo maior com base nesse critério. A versão de desenvolvimento do UD_Nheengatu-CompLin supera as versões correspondentes dos dois outros maiores *treebanks* de línguas tupis em várias estatísticas computadas pela ferramenta `conllu-stats.pl` do projeto UD (Tabela 1). O *treebank* do nheengatu é o único de língua ameríndia que tem crescido significativamente a cada versão semestral da coleção UD, desde quando estreou na versão 2.11, de 15.11.2022, com apenas 196 sentenças, perfazendo 2146 palavras (ALENCAR, 2024). Da versão 2.14 da coleção UD para a atual versão de desenvolvimento, o número de palavras e sentenças aumentou em 27,17% e 24,10%, respectivamente.

Tabela 1: Dados quantitativos das versões de desenvolvimento do UD_Nheengatu-CompLin (UNC), UD_Guajajara-TuDeT (UGT) e UD_Mbya_Guarani-Dooley (UMD) em 27.09.2024.

Treebank	**Sentenças**	**Palavras**	***POS-tags***	**Lemas**	**Formas**	**Relações de dependência**	***Features***
UNC	1824	19122	16	1447	2109	37	82

UGT	1182	9160	15	593	1314	29	72
UMD	1046	11771	16	103	114	34	44

Fonte: Elaboração própria.

Construímos o UD_Nheengatu-CompLin incrementalmente, começando com sentenças de estrutura mais simples, incorporando progressivamente fenômenos mais complexos. A ferramenta inicial disponível era apenas um analisador morfológico baseado num léxico computacional derivado do glossário de Navarro (2016). Implementamos regras em Python para projetar as árvores dependenciais a partir da análise morfológica, levando em conta características sintáticas gerais da língua, como a ordem básica SVO e a posição final de adposições, subordinadores e do núcleo nominal da construção genitiva (ALENCAR, 2023). Em seguida, expandimos progressivamente o glossário e o analisador morfológico com dados extraídos de Avila (2021), aprimorando o analisador sintático por meio da sua aplicação a sentenças representativas de um espectro cada vez mais amplo de fenômenos gramaticais.

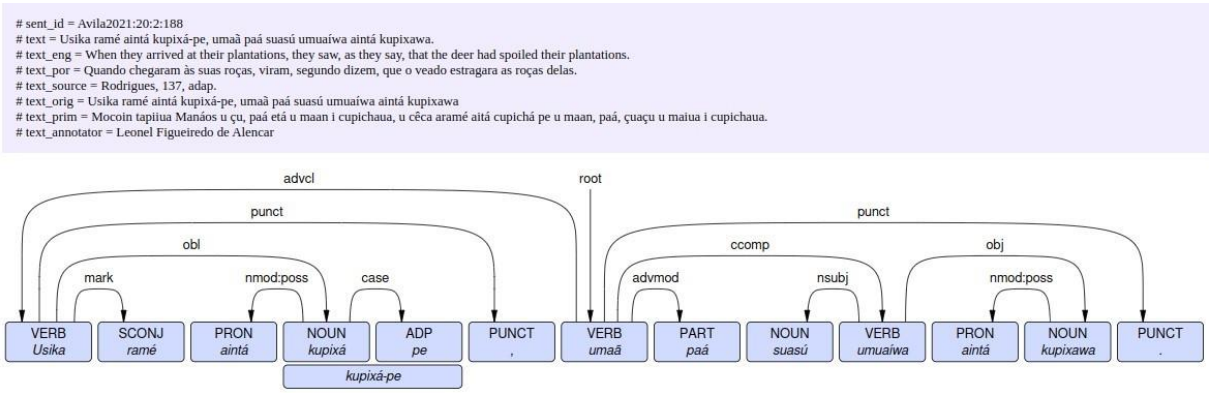
Figura 1: Aplicação do analisador Yauti a exemplo extraído de Avila (2021).

```
Python 3.8.10 (default, Sep 11 2024, 16:02:53)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
>>> import Yauti
>>> s=''' Yautí uyana piri/advg se suí? (Muniz, 84, adap.) -  O jabuti corre mais do que eu?'''
>>> Yauti.parseExample(s,'Avila2021',0,0,190,check=False)
# sent_id = Avila2021:0:0:190
# text = Yautí uyana piri se suí?
# text_eng = Does the tortoise run faster than me?
# text_por = O jabuti corre mais do que eu?
# text_source = Muniz, 84, adap.
# text_annotator = Leonel Figueiredo de Alencar
# inputline = Yautí uyana piri/advg se suí?
1       Yautí   yautí   NOUN    N       Number=Sing     2       nsubj   _       TokenRange=0:5
2       uyana   yana    VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       TokenRange=6:11
3       piri    piri    ADV     ADVG    AdvType=Deg     2       advmod  _       TokenRange=12:16
4       se      se      PRON    PRON2   Case=Gen|Number=Sing|Person=1|PronType=Prs      2       obl     _       Tok
enRange=17:19
5       suí     suí     ADP     ADP     AdpType=Post    4       case    _       SpaceAfter=No|TokenRange=20:23
6       ?       ?       PUNCT   PUNCT   _       2       punct   _       SpaceAfter=No|TokenRange=23:24
```

Fonte: Elaboração própria.

Um total de 58,9% das sentenças do *treebank* são exemplos isolados, a maioria das restantes integram blocos de duas, três ou mais sentenças que constituem trechos contínuos dos textos dos quais foram extraídas, incluindo uma lenda inteira de Magalhães (1876), as 12 lendas de Casasnovas (2006) em sua quase totalidade, a transcrição de uma conversa entre dois falantes, extraída de Moore, Facundes e Pires (1994), e quatro textos de Navarro (2016). Quase 40% dos exemplos provêm de Avila (2021). Este dicionário contém mais de 4000 abonações em ortografia normalizada, que basta copiar e colar no IDLE, ambiente de desenvolvimento de Python, para realizar a análise sintática dependencial por meio do Yauti (Figura 1). Navarro (2016) e Casanovas (2006) contribuem cada um com 11,8% e Cruz, com 6,6%.

Figura 2: Análise do UD_Nheengatu-CompLin para exemplo extraído de Avila (2021).



Fonte: Elaboração própria.

O restante das sentenças do *treebank* distribui-se entre 18 outras publicações, a maior parte do século XIX e início do século XX. Metadados informam a procedência de cada sentença e se integra um bloco (e, nesse caso, qual a sua posição relativa dentro do bloco) ou constitui exemplo isolado (Figura 2). Conquanto limite o potencial de utilização do *treebank* para investigações quantitativas ou de cunho discursivo, a atual composição do *treebank* não impede que venha a ser utilizado com proveito para investigações qualitativas de caráter lexical, morfológico ou sintático ou para treinamento de um *parser* neural (ALENCAR, 2024).

O UD_Nheengatu-CompLin atende a todos os critérios do validador `validate.py`, a que são submetidos os *treebanks* da coleção UD a cada *release*. Esta ferramenta verifica não apenas a obediência às especificações de formato, mas também aspectos da consistência com o esquema de anotação da teoria UD. Na versão 2.14 da coleção UD, o UD_Nheengatu-CompLin, o UD_Mbya_Guarani-Thomas e os dois de variedades regionais do nahuatl são os únicos de línguas ameríndias com a avaliação de duas estrelas, numa escala de zero a cinco (ALENCAR, 2024). Essa classificação, computada pela ferramenta `evaluate_treebank.pl`, leva em conta uma série de fatores, como disponibilidade, diversidade de gêneros textuais e quantidade de erros computados pela ferramenta `udapy`. Atualmente, a versão de desenvolvimento do UD_Nheengatu-CompLin é a única a obter 3,5 estrelas, representando uma melhora de 75%. As notas das versões de desenvolvimento dos demais *treebanks* de línguas ameríndias variam entre 0 e 2 estrelas. O alto desempenho na validação e avaliação automáticas, porém, não assegura que cada sentença do UD_Nheengatu-CompLin tenha sido analisada da melhor maneira conforme o modelo UD. Por enquanto, apenas 30,54% das análises foi revisada por um anotador adicional. Procuraremos sanar essa lacuna numa próxima versão do *treebank*.

Agradecimentos: FAPESP (Processo 22/09158-5).

Referências

ALENCAR, L. F. de. Yauti: A tool for morphosyntactic analysis of Nheengatu within the Universal Dependencies framework. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14, 2023, Belo Hori zonte/MG. *Anais* … Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 135- 145.

ALENCAR, L. F. de. Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo Dependências Universais. *Texto Livre* , Belo Horizonte, v. 17, p. e52653, 2024.

AVILA, M. T. *Proposta de dicionário nheengatu-português*. 2021. Tese (Doutorado em Estudos da Tradução) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2021.

CASASNOVAS, A. Noções de língua geral ou nheengatú: gramática, lendas e vocabulário. 2. ed. Manaus: Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, 2006.

CRUZ, A. *Fonologia e gramática do nheengatú*: a língua falada pelos povos Baré, Warekena e Baniwa. Utrecht: LOT, 2011.

CRUZ, A. The rise of number agreement in Nheengatu. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, Belém, v. 10, n. 2, p. 419-439, 2015.

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (Org.). *Ethnologue*: languages of the world. 27. ed. Dallas: SIL International, 2024. Disponível em: http://www.ethnologue.com. Acesso em: 28 set. 2024.

FREIRE, J. R. B. *Rio Babel*: a história das línguas na Amazônia. 2. ed. Rio de Janeiro: EdUERJ, 2011.

GALVES, C. *et al.* Annotating a polysyntheticlanguage: from Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, v. 59, n. 3, p. 631-648, 2017.

MAGALHÃES, J. V. C. de. *O selvagem*. Rio de Janeiro: Typographia da Reforma, 1876.

MARNEFFE, M.-C. de et al. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, p. 255–308, 2021.

MOORE, D.; FACUNDES, S.; PIRES, N. *Nheengatu (Língua Geral Amazônica), its history, and the effects of language contact*. UC Berkeley: Department of Linguistics, 1994. Disponível em: https://escholarship.org/uc/item/7tb981s1 Acesso em: 31 mai. 2023.

MOORE, D. Historical development of Nheengatu (Língua Geral Amazônica). In: MUFWENE, S. S. (Org.). *Iberian imperialism and language evolution in Latin America*. Chicago: University of Chicago Press, 2014. p. 108-142.

NAVARRO, E. A. *Curso de Língua Geral (nheengatu ou tupi moderno)*: a língua das origens da civilização amazônica. 2. ed. São Paulo: Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2016.

NAVARRO, E. A.; AVILA, M. T.; TREVISAN, R. G. O Nheengatu, entre a vida e a morte: a tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, Campina Grande, v. 6, n. 2, p. 9-29, 2017.

RODRIGUES, A. D. As línguas gerais sul-americanas. *Papia*, São Paulo, v. 4, n. 2, p. 6-18, 1996.

RODRÍGUEZ, L. M. et al. Tupian Language Resources: data, tools, analyses. In: MELERO, M.; SAKTI, S.; SORIA, C. (Org.). *Proceedings of the LREC 2022 Workshop of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*. Paris: European Language Resources Association, 2022. p. 48-58.

STRAKA, M.; STRAKOVÁ, J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task*: Multilingual parsing from raw text to Universal Dependencies. Vancouver: Association for Computational Linguistics, 2017. p. 88–99.

LEVANTAMENTO DE COLOCAÇÕES EM BLOGS DE COWORKING: UM COTEJO PRELIMINAR DE TEXTOS AUTÊNTICOS E TRADUZIDOS

Patrícia Helena FREITAG[89]

RESUMO: Este artigo apresenta os resultados preliminares de um estudo sobre colocações (Tagnin, 2013) que se insere nos Estudos da Tradução e utiliza Linguística de Corpus como metodologia. O objetivo é comparar artigos de blogs de coworking em português autêntico e traduzido, verificar se há diferenças nas colocações e investigar os motivos. Partiu-se de listas de n-gramas para identificar as colocações e de linhas de concordância para realizar a análise. Espera-se que os resultados contribuam para a conscientização acerca do uso de colocações consagradas em traduções.

Palavras-chave: blog de coworking; colocações; tradução; convencionalidade; Linguística de Corpus.

Introdução

No setor da tradução especializada, o controle de qualidade é uma etapa importante no fluxo de trabalho das agências de tradução. Existem diversos modelos de avaliação de tradução profissional, como o LISA e o DQF-MQM, em que um avaliador encontra e classifica erros de acordo com categorias e gravidades predefinidos (Portilho, 2019). Uma das categorias costuma ser reservada para inadequações de estilo, englobando traduções literais, estruturas que não soam naturais e combinações de palavras pouco convencionais. Por isso, é importante que os tradutores produzam textos não apenas corretos, mas que soem naturais e convencionais.

A convencionalidade pode ser entendida como o uso rotineiro da linguagem (López-Rodríguez, 2016) e ocorre em diferentes níveis, como o sintático, o semântico e o pragmático (Tagnin, 2013). No nível sintático, atua a combinabilidade, ou seja, a atração de determinadas palavras entre si por uma questão de convenção de uso.

Esse é justamente o nível de interesse neste trabalho, que investiga um tipo específico de combinação de palavras — as colocações — em artigos de blogs de coworking, cotejando as ocorrências em um corpus de português traduzido e um corpus de português autêntico. Como exemplo de colocação neste gênero, temos *trabalho remoto*. Nos corpora de estudo, não ocorrem alternativas com significado semelhante, como *trabalho distante* ou *trabalho afastado*, pois simplesmente não se convencionou usar essas combinações.

Colocações e tradução

As colocações consistem em palavras que coocorrem com maior frequência que o acaso, e Tagnin (2013) observa que pode haver hífen, artigo e/ou preposição entre elas. Existem colocações adjetivas (Adj. + S ou S + Adj.), adverbiais (Adv. + Adj. ou V + Adv.), nominais (S + S) e verbais (V + S ou V +

[89] Doutoranda na linha de pesquisa Estudos do Léxico e da Tradução do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre/RS. patriciafreitag@gmail.com

Adj.). Tagnin (2013) apresenta ainda dois outros tipos de colocações: expressões especificadoras de unidade e coletivos, que não serão abordados aqui.

Como mostram alguns estudos sobre processamento da linguagem, há evidências de que as traduções podem sofrer influência do texto-fonte (HansenSchirra; Nitzke; Oster, 2017). Partimos do pressuposto que essa influência pode acabar afetando os padrões colocacionais na língua-alvo, caso o tradutor considere as palavras individualmente em vez de tratar as combinações como unidades linguísticas convencionais. Caso isso se confirme, é provável que ocorra quebra de convencionalidade na tradução no que diz respeito a colocações. Alguns estudos já mostram evidências de que as traduções contêm padrões colocacionais diferentes daqueles de textos autênticos (Mauranen, 2007), e buscamos investigar isso no gênero blog de coworking.

Corpora do estudo

Este trabalho com base em corpus visa comparar as colocações encontradas em artigos de blogs de coworking escritos originalmente em português e artigos do mesmo gênero escritos originalmente em inglês e traduzidos para o português. Com esse fim, foi compilado um corpus bilíngue paralelo de textos autênticos em inglês e traduzidos para o português e um corpus monolíngue de textos autênticos em português do mesmo gênero. Esse gênero textual foi escolhido por ser relevante nos dias de hoje, em que o trabalho remoto é uma realidade: segundo a Woba, houve um aumento de 63% no número de espaços de coworking no Brasil (Woba, 2023). Além disso, existem empresas internacionais que oferecem espaços de escritório no país e precisam ter seus materiais traduzidos.

O corpus paralelo é formado por artigos de três empresas de coworking internacionais: Spaces, Regus e WeWork. Não foram encontradas outras empresas com publicações de blog em inglês traduzidas para português. Portanto, o corpus paralelo contém 100% do universo possível desse gênero traduzido. Esse corpus contém 335 artigos, sendo que a porção em português contém 12.795 types e 398.039 tokens.

Havia uma disponibilidade maior de blogs de empresas nacionais. Para fins de seleção para o corpus monolíngue, optou-se por desconsiderar blogs com menos de 10 publicações. Foram necessários 535 artigos de seis empresas nacionais (CWK, nex., Trust Coworking, Vip Office, Club Coworking e Coworking Town) para chegar a um número balanceado em termos de tokens (8.437 types e 396.296 tokens) em relação ao corpus de textos traduzidos.

Todos os blogs eram de livre acesso. Os artigos foram consultados manualmente e salvos em arquivos .txt individuais com um nome exclusivo. Os metadados (URL, data de coleta e data de publicação, quando disponível) foram registrados em uma planilha de Excel. Os arquivos .txt dos corpora em português foram carregados na ferramenta Sketch Engine (Kilgarriff *et al*. 2014). Além disso, os arquivos do corpus paralelo foram alinhados com o LF Aligner (Farkas, 2017) e carregados no AntPConc (Anthony, 2017) para servir de apoio na análise.

Extração e organização de colocações

Foram geradas as listas de n-gramas (2 a 4 palavras) para os dois corpora de português. Em seguida, as listas foram exportadas para Excel e reorganizadas em três planilhas: 1) n-gramas que apareciam nos dois corpora; 2) n-gramas que apareciam exclusivamente no corpus de português autêntico; e 3) n-gramas que apareciam exclusivamente no corpus de português traduzido. Como ponto de corte para este trabalho, foram excluídos os itens que apareciam em menos de 5% dos textos dos corpora. Em seguida, foi feita a leitura atenta, linha a linha, identificando as colocações e excluindo os demais itens.

Durante a análise, para entender as semelhanças e diferenças entre o corpus de textos traduzidos e o de textos autênticos, consultamos linhas de concordância no Sketch Engine (para ver o contexto de uso em português) e no AntPConc (para ver o texto em inglês e investigar uma possível influência da língua-fonte); também usamos o recurso Word Sketch do Sketch Engine para visualizar outros possíveis colocados para uma palavra de busca.

Resultados e discussão

Muitas das colocações que ocorrem tanto no corpus de textos traduzidos quanto no de textos autênticos destacam os pontos em comum abordados por empresas nacionais e internacionais de coworking. Estes são alguns exemplos: a) colocações adjetivas: *grandes empresas*, *home office*, *trabalho híbrido*, *trabalho remoto*; b) colocações nominais: *ambiente de trabalho*, *espaço de coworking*, *espaço de trabalho*, *local de trabalho*, *modelo de trabalho*, *sala de reunião*; e c) colocações verbais: *trabalhar em casa*. Nota-se que as empresas de coworking, tanto as nacionais quanto as internacionais, abordam bastante os espaços de trabalho e os modos de trabalhar.

Desses dados, é interessante observar que o estrangeirismo *home office* aparece nos dois corpora. Com base em nosso conhecimento de mundo e de língua, pensamos em outras opções que poderiam expressar esse conceito e as buscamos no corpus de textos autênticos: encontramos 5 ocorrências de *escritório em casa* e nenhuma de *escritório doméstico* ou *escritório residencial*. Os números foram parecidos no corpus de textos traduzidos: 1 ocorrência apenas de *escritório em casa* e nenhuma das outras duas opções. Ou seja, parece que apenas *home office* é amplamente aceito para esse conceito. É importante para o tradutor ter esse conhecimento para que aposte no uso do estrangeirismo, em vez produzir uma tradução pouco convencional simplesmente com a finalidade de evitá-lo.

Quanto às diferenças, um exemplo interessante é *sala privativa*, colocação adjetiva encontrada apenas no corpus de textos autênticos. Para identificar se e como esse conceito aparece no corpus de textos traduzidos, buscamos o lema *sala* no recurso Word Sketch, que mostra palavras que coocorrem com a palavra de busca. Os resultados não foram reveladores, uma vez que as colocações de *sala* + adjetivo foram: *sala comercial*, *sala espaçosa*, *sala pequena*, *sala grande* e *sala interna*. Ou seja, nenhuma dessas colocações se concentra no caráter privativo do espaço. Partimos para mais uma busca com o Word Sketch, dessa vez no corpus de textos traduzidos. Utilizamos a palavra de busca *privativo* e encontramos 50 ocorrências de *escritório privativo*. Isso sugere que, enquanto em textos autênticos se fala em *sala privativa*, nos textos

traduzidos se usa *escritório privativo*. Com auxílio do AntPConc, buscamos *escritório privativo* no corpus traduzido e constatamos que parece haver influência do texto em inglês na tradução, visto que o texto-fonte usava *private office* nesses casos, e é de amplo conhecimento que uma tradução direta de *office* costuma ser *escritório*. Para garantir uma tradução que soe natural, o tradutor poderia optar por uma tradução menos direta de *office* no contexto de *private office*, resultando na colocação *sala privativa*, visto que é a combinação consagrada em textos autênticos em português.

Mais um exemplo de diferença é a colocação adverbial *bem localizado*, com 34 ocorrências no corpus de textos autênticos contra apenas 1 ocorrência no de textos traduzidos. Isso poderia indicar que os coworkings brasileiros estão preocupados com a localização de seus escritórios e que as empresas internacionais não têm essa mesma preocupação. No entanto, isso não parece plausível, pois é de conhecimento geral que as empresas devem levar a localização em conta para serem bemsucedidas. Então, usamos o AntPConc para consultar o texto-fonte em inglês dessa única ocorrência de *bem localizado* no corpus traduzido e investigar as possíveis formas de expressar uma boa localização em inglês. O texto em inglês para *bem localizado* era *well-positioned*. Imaginamos que poderia haver outras ocorrências de *well-positioned* na porção em inglês do corpus paralelo que poderiam ter sido traduzidas de outra forma. Procuramos, então, *well-positioned* na porção em inglês do corpus paralelo, mas não foram encontradas ocorrências. Em mais uma tentativa, buscamos parte dessa colocação: *position**. Nessa última busca, encontramos diferentes estruturas para falar sobre a localização dos escritórios, como *occupies a prime position*, *enjoys a premium position*, *is perfectly positioned*, *it's ideally positioned*, entre outras. Em todas elas, a tradução ficou com o verbo *posicionar*, conjugado conforme adequado gramaticalmente. Voltando para a ideia original de *bem localizado*, continuamos as buscas na porção em inglês do corpus bilíngue, dessa vez buscando por *located*, imaginando que *localizado* poderia ser uma opção que ocorre nas traduções devido à semelhança das palavras em inglês e português. De fato, encontramos três ocorrências de *conveniently located* traduzidas como *convenientemente localizado*. Essas buscas mostram que, quando existem os termos *located* ou *positioned*, a tradução parece ser influenciada pelo texto-fonte e utiliza os cognatos *localizado* e *posicionado*. Porém, no primeiro caso (*localizado*), a influência é positiva, já que gera uma forma convencional no gênero na língua de chegada (como apresentamos no início do parágrafo, *bem localizado* é uma colocação frequente no corpus de textos autênticos). Por outro lado, o segundo caso (*posicionado*), consiste em uma influência negativa, pois trata-se de uma forma atípica no gênero em português. Dessa forma, convém que o tradutor considere a opção *bem localizado*, mesmo que o texto-fonte faça menção a *position*, como em *occupies a prime position* e *is perfectly positioned*.

Considerações finais

O estudo ainda está em andamento. A análise das colocações continuará até o final das listas de n-gramas, sempre buscando as semelhanças e diferenças e procurando entender se o texto-fonte em inglês parece influenciar as traduções.

Com esta análise preliminar, observa-se que este estudo descritivo pode ajudar na conscientização de tradutores profissionais e em formação sobre a importância de se produzir combinações de palavras convencionais na língua-alvo e no gênero trabalhado.

As reflexões aqui levantadas e os processos de extração e análise de colocações podem ser aplicados a outros gêneros, evidenciando a utilidade de corpora para a identificação de combinações convencionais para fins de tradução.

REFERÊNCIAS

ANTHONY, Laurence. **AntPConc**. Versão 1.2.1. Tóquio, 2017. Programa. Download disponível em https://www.laurenceanthony.net/software. Acesso em: 05 maio 2022.

FARKAS, András. **LF Aligner**. Versão 4.21. 2019. Disponível em: https://sourceforge.net/projects/aligner/. Acesso em: 20 jun 2022.

KILGARRIFF, Adam *et al*. The Sketch Engine: ten years on. **Lexicography**, [S.l.], v. 1, n. 1, p. 7-36, jul. 2014. Equinox Publishing. DOI: http://dx.doi.org/10.1007/s40607-014-0009-9. Acesso em: 05 maio 2022

LÓPEZ-RODRÍGUEZ, Clara Inés. Using corpora in scientific and technical translation training: resources to identify conventionality and promote creativity. **Cadernos de Tradução**, [S.l.], v. 36, n. 1, p. 88-120, 26 abr. 2016. Universidade Federal de Santa Catarina (UFSC). DOI: http://dx.doi.org/10.5007/2175-7968.2016v36nesp1p88. Acesso em: 07 set. 2020.

MAURANEN, Anna. Universal Tendencies in Translation. *In*: ANDERMAN, Gunilla; ROGERS, Margaret (ed.). **Incorporating Corpora**: the linguist and the translator. Clevedon: Multilingual Matters Ltd, 2007. Cap. 3, p. 32-48.

PORTILHO, Talita. **Avaliação de tradução nos contextos profissional e pedagógico:** proposta de unidade didática para revisão e avaliação por pares. 2019. Dissertação (Mestrado em Estudos da Tradução) — Universidade Federal de Santa Catarina, Florianópolis, 2019.

TAGNIN, Stella E. O. **O jeito que a gente diz:** combinações consagradas em inglês e português. São Paulo: Disal, 2013.

WOBA. **Censo Coworking**: uma análise Woba do mercado brasileiro. S.L, 2023. Disponível em: https://21669165.fs1.hubspotusercontentna1.net/hubfs/21669165/censo-coworking-woba-

2023%20(1).pdf?utm_medium=email&_hsmi=254376693&_hsenc=p2ANqtz8O5nFxZ6EV_WhB9qQbTIegWh6pWfwYnn61mBIdIpkFoCLVvuAvcTYcVnob07OItjeoptberYS6ibz_8F2RHJuWkLcUYg_sfYbfpA0ufEuKTUHovo&utm_content=254376693&ut m_source=hs_automation. Acesso em: 31 jan. 2024.

ANOTAÇÃO DE CÓRPUS, UM LUGAR PRIVILEGIADO DE OBSERVAÇÃO LINGUÍSTICA: UM ESTUDO DAS APOSIÇÕES DO PORTUGUÊS BRASILEIRO SEGUNDO O MODELO *UNIVERSAL DEPENDENCIES*

Magali Sanches DURAN[90]
Thiago Alexandre Salgueiro PARDO[91]

Resumo: Para a Linguística de Córpus e para o Processamento de Línguas Naturais (PLN), o córpus é uma fonte de conhecimento. A partir dessa constatação, argumenta-se que a anotação de córpus, uma das atividades linguísticas essenciais para o PLN, constitui também uma atividade interessante para outros linguistas, pois permite registrar as análises linguísticas no próprio córpus. A fim de exemplificar como a anotação de córpus pode ser inspiradora para as reflexões linguísticas, discute-se o caso das aposições predicativas à esquerda do sujeito, utilizando a abordagem *Universal Dependencies* para o português.

Palavras-chave: Anotação de Córpus; Aposições Predicativas; PLN; UD; Português.

Os fazeres humanos se modificam em função das tecnologias disponíveis e os estudos linguísticos são um bom exemplo disso. Antes dos anos 90, todo projeto linguístico que não quisesse ser chamado de “linguística de poltrona” tinha que conter um item que se chamava “córpus de estudo”. Era a descrição do conjunto de textos (em papel) sobre o qual o linguista se debruçava para fazer suas análises. Consistia em um esforço de olhar para as realizações da língua, evitando utilizar apenas seu modelo mental. Com a popularização dos computadores, essa acepção da palavra “córpus” incorporou os textos digitais e a linguística de córpus encarregou-se de desenvolver novos métodos para explorá-los e deles extrair conhecimento. Paralelamente, o Processamento de Línguas Naturais (PLN) também se desenvolvia. O PLN adotou os córpus como fonte de conhecimento e passou a utilizar métodos computacionais para “aprender” tarefas envolvendo a língua. Isso só foi possível porque se desenvolveu o que hoje conhecemos como “anotação de córpus”, ou seja, um processo pelo qual atribuem-se etiquetas a partes do córpus de modo a explicitar uma análise humana sobre essas partes.

Há uma dupla responsabilidade para que um córpus anotado propicie um bom aprendizado automático da tarefa. Da parte da linguística, é importante que a anotação seja consistente ao longo de todo o córpus, ou seja, etiquetas iguais sejam atribuídas a fenômenos semelhantes e etiquetas diferentes a fenômenos cuja distinção seja relevante para o projeto. Da parte do PLN, é importante que os métodos utilizados sejam capazes de “capturar” a lógica expressa na anotação, gerando algoritmos que reproduzam automaticamente a anotação humana.

[90] Pesquisadora - Núcleo Interinstitucional de Linguística Computacional (ICMC-USP) - pós-doc C4AI

[91] Professor - ICMC-USP São Carlos-SP (taspardo@usp.icmc.br)

A atividade de anotação de córpus e os métodos de aprendizado de máquina evoluíram muito ao longo das últimas décadas. O aumento da capacidade de processamento dos computadores possibilitou o desenvolvimento de novas abordagens de aprendizado automático, inclusive sem o uso de córpus anotados, chegando-se aos atuais modelos gerativos (como o famoso ChatGPT). Entretanto, a anotação de córpus se mantém relevante para várias tarefas de PLN (por exemplo, o próprio ChatGPT necessitou de córpus anotado para aprender a detectar discursos de ódio) e também para estudos linguísticos, já tendo muitas metodologias testadas (Hovy & Lavid, 2010; Ide & Pustejovsky, 2017).

Neste artigo, apresenta-se um caso de anotação de córpus. O cenário é o de anotação de relações de dependência sintática em um córpus de português brasileiro, o Porttinari-base (Duran et al. 2023). O conjunto de etiquetas escolhido é o do projeto *Universal Dependencies* (UD) (de Marneffe et al., 2021). O fenômeno em foco é o conjunto das aposições que predicam sujeitos, em especial as antepostas.

A proposta da UD, aqui adotada, é fornecer um conjunto de etiquetas aplicável a diversas línguas, de forma a constituir uma base de comparação entre córpus anotados. Na data de escrita deste artigo, a UD tinha 233 córpus anotados em 141 línguas. São previstas 17 etiquetas morfossintáticas e 37 relações sintáticas. O uso dessas etiquetas em português está descrito em dois manuais (Duran, 2021; Duran, 2022) que serviram de base para este artigo. A anotação de dependências sintáticas é feita ligando dois a dois os tokens de uma sentença, de modo que um token seja o *head* da relação e o outro seja o dependente. Cada token pode ser dependente de uma única relação, mas pode ser *head* de várias relações. A anotação das relações de dependência resulta no que chamamos de "árvore sintática de dependências" de uma sentença. Essa árvore (Figura 1) tem como raiz (*root*) o núcleo do predicado da oração principal.

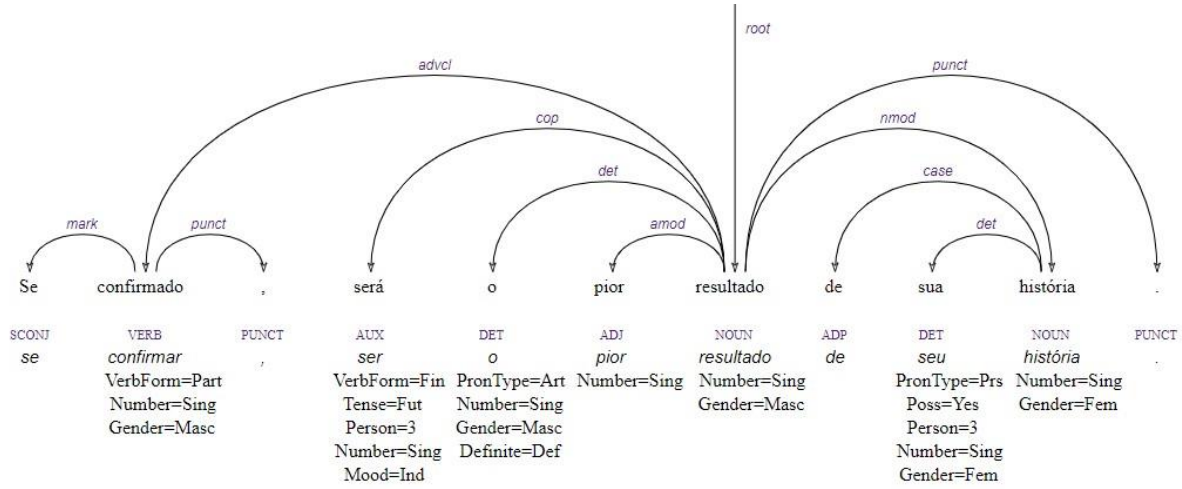


Figura 1 - Exemplo de sentença anotada com etiquetas e diretrizes da abordagem UD

Delineado esse contexto, passamos à discussão das aposições à esquerda do sujeito. Na tradição gramatical do português, as aposições (porções de texto separadas por vírgulas) que predicam o sujeito podem constituir apostos ou orações adjetivas; porém, nos exemplos prototípicos, essas aposições são pospostas aos nominais modificados. Na UD, só são reconhecidas como apostos

relações da esquerda para a direita e entre nominais que sejam intercambiáveis, como nas sentenças 1 e 2 a seguir (aposições em negrito, sujeito sublinhado):

1. O presidente do sindicato dos guardas, **Clovis Pereira**, defende a iniciativa do prefeito.
2. Clovis Pereira, **o presidente do sindicato dos guardas**, defende a iniciativa do prefeito.

Já o exemplo 3 não é considerado aposto na UD:

3. Rodrigo Rocha Loures, **filho de família rica**, achou que valia carregar R$500 mil de Joesley.

A justificativa para isso, na UD, é que, na inversão, o termo à direita, "Rodrigo Rocha Loures", continua sendo sujeito, mostrando que sua função não é intercambiável com a função de "filho de família rica", como fica explícito em (4):

4. **Filho de família rica**, Rodrigo Rocha Loures achou que valia carregar R$500 mil de Joesley.

Poderíamos simplesmente dizer que a restrição de anotação de apostos da UD está equivocada e que temos, sim, apostos à esquerda do sujeito. Contudo, nos deparamos com casos em que o aposto à esquerda está presente e o sujeito está elíptico, como em (5):

5. **Leitor voraz desde garoto,** aos 20 anos começou a vender contos para revistas.

Embora saibamos que "leitor voraz desde garoto" é um predicativo do sujeito elíptico, não haveria um token para ser *head* desta aposição, mesmo que a UD permitisse apostos à esquerda do sujeito. Vamos observar outras aposições à esquerda do sujeito, na forma de adjetivos, como em (6):

6. **Educativas,** brincadeiras clássicas atravessam gerações [...]

Nesse exemplo, o adjetivo "educativas" não tem somente a função de qualificar "brincadeiras", ou seja, não é o mesmo que dizer "brincadeiras clássicas educativas". Poderia se tratar de uma oração adjetiva reduzida de predicativo (sem pronome relativo e sem verbo de cópula), porém orações adjetivas não ocorrem antes de seus antecedentes, o que pode ser observado ao fazermos, para a sentença (6), versões de oração adjetiva desenvolvida posposta ao sujeito (7) e anteposta ao sujeito (8):

7. Brincadeiras clássicas, que são **educativas**, atravessam gerações [...]
8. *Que são **educativas**, brincadeiras clássicas atravessam gerações [...]

A hipótese mais aceitável é a de que a aposição à esquerda do sujeito seja uma oração adverbial, como em (9):

9. [Por serem] **educativas**, brincadeiras clássicas atravessam gerações [...]

Assim, embora as aposições à esquerda do sujeito parecessem, à primeira vista, casos que deveriam ter o sujeito como *head* da relação de dependência, a leitura mais adequada parece ser a de uma oração adverbial.

Essa leitura, por ter o predicado da oração matriz como *head* da relação de dependência, elimina o problema representado pelo sujeito elíptico em (5). Na verdade, essas construções parecem amalgamar dois predicados com um sujeito compartilhado semelhante ao que a tradição gramatical denomina de predicado verbo-nominal. Desdobrados esses dois predicados, teríamos (sujeitos em negrito):

10. **Rodrigo Rocha Loures** é filho de família rica. **Rodrigo Rocha Loures** achou que valia carregar R$500 mil de Joesley.
11. **Ele** é um leitor voraz desde garoto. Aos 20 anos, **ele** começou a vender contos para revistas.
12. **Brincadeiras clássicas** são educativas. **Brincadeiras clássicas** atravessam gerações.

Alguns adjetivos predicativos são mais facilmente identificáveis como orações adverbiais, posto que imprimem uma qualificação circunstancial do sujeito no momento do evento descrito na oração matriz, como no exemplo 13 (aposição em negrito):

13. **Preocupada**, a psicóloga Júlia Prado, 25, pensa em cancelar o pedido.

Essa hipótese ganha força quando testamos o deslocamento da aposição em 14, 15 e 16:

14. A psicóloga Júlia Prado, 25, [por estar] **preocupada**, pensa em cancelar o pedido.
15. A psicóloga Júlia Prado, 25, pensa, [por estar] **preocupada**, em cancelar o pedido.
16. A psicóloga Júlia Prado, 25, pensa em cancelar o pedido, [por estar] **preocupada**.

Em todas essas três versões (14, 15, 16), a possibilidade de uma leitura de "preocupada" como uma oração adverbial reduzida de predicativo[92] (sem conjunção subordinativa e sem verbo de cópula) é factível. Quando a aposição é um sintagma nominal, contudo, como nos exemplos 4 e 5, a versão posposta não é possível, talvez porque um sintagma nominal possa ter outras funções após o verbo e isso pudesse gerar ambiguidade. Contudo, em ambos os casos, a análise seria a mesma, de oração adverbial (**advcl** na UD), como ilustrado nas Figuras 2 e 3:

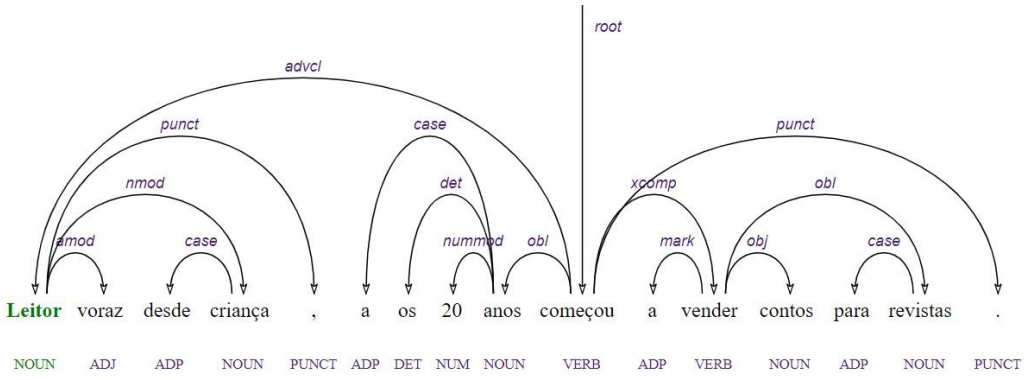


Figura 2 - Aposição (NOUN) à esquerda do sujeito anotada como oração adverbial

[92] Embora as gramáticas só mencionem orações reduzidas de infinitivo, gerúndio e particípio, temos encontrado muitos casos de orações em que o verbo de cópula está elíptico, o que nos levou a empregar o termo "oração reduzida de predicativo".

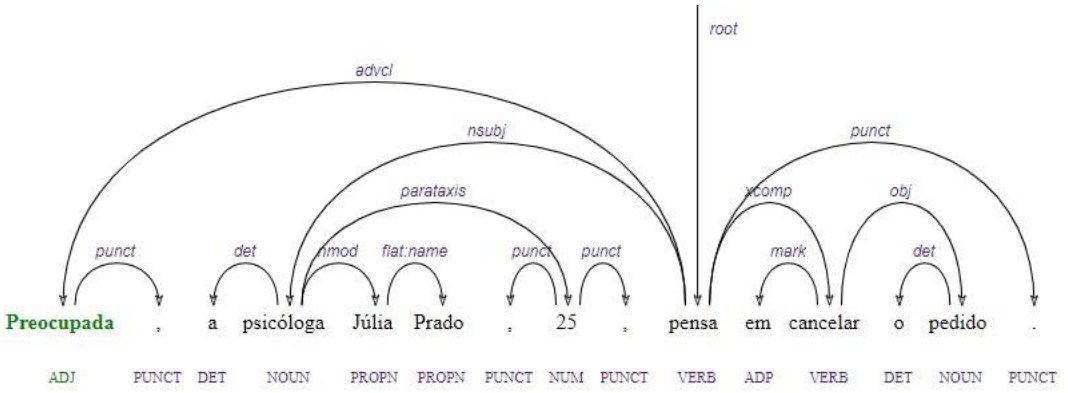


Figura 3 - Aposição (ADJ) à esquerda do sujeito anotada como oração adverbial

A proposta de anotação das aposições à esquerda do sujeito como oração adverbial que modifica outra oração foi implementada recentemente no córpus Porttinari-base. Como esse fenômeno é pouco frequente no córpus e o predicado nominal ora é um adjetivo, ora é um sintagma nominal, há um problema de esparsidade de dados, o que pode prejudicar o aprendizado automático da classificação proposta. O parser treinado na versão anterior desse córpus (Lopes & Pardo, 2024), disponível on-line[93], quando ainda não havíamos adotado um padrão para anotar aposições de sujeito, classifica casos semelhantes ora como oração adverbial (advcl, na UD), ora como adjunto adnominal (amod ou nmod na UD). Em trabalhos futuros, pretendemos anotar mais sentenças que contenham este fenômeno e disponibilizá-las, juntamente com a nova versão do córpus, para retreinamento do parser.

Pelo que pode ser observado, há problemas que emergem durante a tarefa de anotação e que exigem reflexões para que sejam tomadas decisões de anotação. O córpus deixa de ser apenas um lugar para testar hipóteses pré-definidas e passa a ser o próprio *locus* de observação, de levantamento de hipóteses, de reflexão e de armazenamento das decisões tomadas.

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.

Referências

BIBER, DOUGLAS. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In: **The Oxford Handbook of Linguistic Analysis**. Oxford Academic: 2015.

DE MARNEFFE, MARIE-CATHERINE; MANNING, CHRISTOPHER D.; NIVRE, JOAKIM; ZEMAN,

DANIEL. Universal Dependencies. **Computational Linguistics**, 47(2), p.255-308, 2021.

[93] http://portparser.icmc.usp.br:8082/

DURAN, MAGALI SANCHES. (2021). Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). **Relatório Técnico do ICMC n. 434**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

DURAN, MAGALI SANCHES. (2022). Manual de Anotação de Relações de Dependência –Versão Revisada e Estendida. **Relatório Técnico do ICMC n. 440**. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

DURAN, MAGALI SANCHES; LOPES, LOPES; NUNES, MARIA DAS GRAÇAS VOLPE; PARDO,

THIAGO ALEXANDRE SALGUEIRO. The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. *In*: **Proceedings of the 14th Symposium in Information and Human Language Technology** (STIL), p. 115-124, 25-29 setembro, 2023.

HOVY, EDUARD; LAVID, JULIA. Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. **International Journal of Translation**, 22(1), p. 13-36, 2010.

IDE, NANCY; PUSTEJOVSKY, JAMES. **The Handbook of Linguistic Annotation**. Springer: 2017.

LOPES, LUCELENE; PARDO, THIAGO ALEXANDRE SALGUEIRO. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. *In* **Proceedings of the 16th International Conference on Computationa Processing of Portuguese** (PROPOR), Vol. 1, p. 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics (ACL): 2024.

DESAFIOS DA LINGUÍSTICA DE *CORPUS* IMPOSTOS PELA INTELIGÊNCIA ARTIFICIAL: REDISCUTINDO ALGUNS CONCEITOS

Jackson Wilke da Cruz SOUZA[94]

RESUMO: Neste trabalho busco discutir impactos e desafios ocasionados pela Inteligência Artificial à Linguística de *corpus*. Tal discussão parece ser emergente frente a diferentes metodologias supervisionadas e não supervisionadas por humanos quanto à mineração e obtenção de dados textuais. Assim, busco recuperar conceitos sobre autenticidade, processamento e representatividade de informações linguísticas em *corpus*, além de indicações para trabalho com dados de redes sociais.

Palavras-chave: Linguística de *corpus*. Inteligência Artificial. Conceitos. *User Generated Content*.

ABSTRACT: In this paper, I aim to discuss the impacts and challenges posed by Artificial Intelligence to Corpus Linguistics. Such a discussion is timely given the emergence of various human-supervised and unsupervised methodologies for mining and obtaining textual data. Therefore, I intend revisit concepts related to the authenticity, processing, and representativeness of linguistic information in *corpora*, as well as providing insights for working with data from social media.

Keywords: *Corpus* Linguistics. Artificial Intelligence. Concepts. User Generated Content.

INTRODUÇÃO

É notório que as redes sociais têm sido fonte dos processos de produção, circulação e recepção de conteúdos que interessam diferentes segmentos sociais. Tais processos foram potencializados com as últimas ondas da Web (Passarelli; Gomes, 2020), fazendo com que o usuário não apenas consumisse, mas também gerasse conteúdo na internet, trazendo à tona o conceito de *User Generated Content* (UGC).

Santos (2022), ao estudar a proeminência de estudos que observam UGC, aponta que há quatro grandes temas de estudo: (1) conteúdo criado pelo usuário, (2) práticas comunicativas, como o jornalismo, (3) estudos sobre usos linguísticos que ocorrem na interlocução entre usuário e audiência e (4) plataforma Web 2.0. O autor ainda destaca que UGC, na literatura mais recente, tem sido foco nas disciplinas de Comunicação social, mídia e jornalismo, Consumo e negócios, Ciência da informação, Ciências humanas e Ciência política.

Nas áreas de Processamento de Línguas Naturais (PLN) e Linguística de *corpus* (LC), UGC tem sido de interesse especialmente em pesquisas que visam ao processamento de diferentes níveis da língua (como o morfológico, sintático e semântico), além de compreender perfis ideológicos e comportamentais de

[94] Docente da Universidade Federal da Bahia no Instituto de Ciência, Tecnologia e Inovação (ICTI) e no Programa de Pós-Graduação em Língua e Cultura (PPGLinC). E-mail: jackcruzsouza@gmail.com

usuários a partir de suas construções textuais. É importante destacar que, nesse contexto, há diferentes vertentes metodológicas sobre as pesquisas em LC. Há trabalhos que desenvolvem pesquisas com *corpus* assíncrono, em que os textos[95] são compilados e processados fora de um ambiente on-line; já outros, com *corpus* síncrono, em que a compilação e o processamento se dão de maneira on-line e contínua; por fim, outros que utilizam a web como corpus, em que há o material linguístico está sendo produzido e pensado de maneira síncrona ou assíncrona.

O que destaco é que em todas essas abordagens há conceitos propostos pela LC que nos serviram de base até aqui. Porém, por mudanças sócio-históricas, precisamos repensar esses conceitos, especialmente porque as fronteiras entre o digital e o não digital estão cada vez mais borradas (Burnham, 2014). Ainda que discussões sobre IA em PLN não sejam novidade, elas se tornaram populares devido ao destaque midiático que se teve por conta dos *Large Language Models* (LLM). Esses modelos podem ser de língua geral, como o GPT (do inglês, *Generative Pretrained Transformer*), ou de domínio, como o BloombergGPT (Wu *et al.*, 2023). Em ambos os tipos, os modelos são treinados a partir de grande quantidade de dados para que processem aspectos das línguas naturais.

Assim, meu objetivo neste trabalho é discutir como conceitos caros à LC, como autenticidade, naturalidade e extensão, requerem novas discussões frente ao cenário de *Big data* e IA. Além disso, destaco o cuidado que os pesquisadores precisam ter em seus estudos para lidar com (meta)dados linguísticos que envolvem informações advindas de UGC.

REVISITANDO CONCEITOS

Podemos dizer que as áreas de PLN e LC são próximas graças ao fator computacional comum a elas. O entendimento que temos hoje sobre objeto e metodologia da LC apresentou diversos avanços com abordagens do PLN com relação ao processamento de dados linguísticos. Os textos que passaram a integrar os corpora deveriam, então, estar em formatos legíveis por máquina (Sardinha, 2000), possibilitando pré-processamento, processamento e análise em larga escala de dados. Como resultado, espera-se que quanto maior o volume de dados, maior seja o conhecimento da língua e sobre os usos que os falantes fazem dela e se constituem a partir dela.

No início da área, os *corpora* mais extensos tinham em torno de 1 milhão de palavras, como o projeto Brown *corpus* que era composto de 500 textos de 2000 palavras em 15 gêneros. Mais tarde, foram publicados outros *corpora* que tinham uma quantidade de palavras muito mais elevada, como o *corpus* multilíngue *News on the Web*, que soma mais de 5 bilhões de palavras (Tagnin, 2018). Esse crescimento se deu por conta de contribuições de distintas áreas da Linguística, como a Lexicografia, Terminologia e a Tradução (Viana; Tagnin, 2015) a partir do desenvolvimento de *corpora* especializados.

---

[95] Neste trabalho tratarei “texto” como material que integra o *corpus*, mesmo admitindo que há outros materiais que podem integrar os *corpora*, como áudio e imagens.

Aqui cabe, então, discutir um primeiro conceito caro à LC: a *representatividade*. Sinclair (1991) defende que um *corpus* representativo deve ser o maior possível, já que a linguagem é um sistema probabilístico (Halliday, 1991) e o *corpus* é "uma amostra de uma população cuja dimensão não se conhece (a linguagem como um todo)" (Sardinha, 2000, p. 342). Tendo esse pressuposto, no âmbito da língua geral, o *corpus* deve ser extenso, para que a amostra possa ser, de fato, representativa.

Nesse sentido, deparamo-nos com ao menos dois desafios. O primeiro é de ordem metodológica. Sabe-se que as áreas de LC e PLN dispõem de métodos bastante eficazes e robustos para a coleta e processamento de textos, como *Web Scraping* (Zhao, 2022). Nesse método, os dados são extraídos da Web e salvos em sistemas de arquivos ou banco de dados para que possam ser analisados ou recuperados posteriormente. Esse método pode ser feito manualmente ou a partir de técnicas computacionais. Neste último caso, a partir da disponibilização de um endereço eletrônico sem que haja as especificações paramétricas corretas, o sistema computacional pode trazer indiscriminadamente todo o conteúdo do site. Neste último caso, não é trivial dizer que os textos são selecionados para os *corpora* devam ser submetidos a uma curadoria em função do objetivo da pesquisa e sobre o que ele representa.

O segundo desafio diz respeito ao processamento, pois a partir da coleta dos dados, será necessário processar as informações linguísticas que foram coletadas. Caso a coleta dos dados linguísticos seja feita por métodos puramente automáticos e não haja nenhum tratamento para classificar os textos entre produzidos por humanos ou por inteligência artificial, as informações que serão extraídas dos dados poderão ser artificiais. Atualmente, há esforços para classificar textos gerados por IA e por humanos, como o trabalho de Ayapova e Skripnikova (2022), que classifica textos jornalísticos. Entretanto, esse tipo de processamento dos dados ainda parece distante dos estudos em LC, pois deveria estar aliado aos métodos utilizados nas pesquisas, como apontado anteriormente.

Outro conceito da LC que destaco aqui: a *autenticidade* dos conjuntos de textos. A literatura sobre LC (Sinclair, 1991; Sardinha, 2000; Freitas, 2024) aponta para a necessidade de os textos que compõem o *corpus* serem autênticos. Isso significa dizer que o material compilado deva ser produzido por humanos. Reunir textos autênticos é garantir que os resultados descritos e analisados refletem, de fato, a língua e como seus falantes a utilizam. Ao extrair informações linguísticas de seus contextos naturais de uso, a depender do objetivo, deve-se prever a compilação de textos produzidos por usuários da Web. Isso se justifica pelo fato de estarmos lidando com "produtores de conteúdo" e não apenas com "usuários de redes sociais".

Porém, diante da possibilidade de existirem produtos disponíveis que podem ser compilados que não foram produzidos intelectualmente por humanos, a autenticidade tornou a ter relevância. Atualmente, sabe-se sobre a existência de obras literárias completas que foram geradas a partir de IA, como discutido por Dalte (2020). Mais recentemente, algumas discussões giram em torno de questões éticas, como os direitos autorais da obra (Garcia, 2020).

O que levanto aqui não é o fato de se os textos artificiais não devam figurar os *corpora*, mas sim a maneira como eles estão agregados à coletânea de textos. Cabe, então, ao pesquisador tomar dois cuidados: (i) caso opte por métodos automáticos, deve-se aplicar métodos suplementares de identificação e classificação de textos (não) naturais que farão parte do conjunto de dados; (ii) caso opte por incluir textos artificiais em seu conjunto de dados, deverá identificá-los e alertar ao consulente sobre essa característica. Textos artificialmente produzidos podem fazer parte do conjunto de dados linguísticos, desde que o propósito sobre esse tipo de inclusão não prejudique descrições linguísticas ou ainda tome algo como verdadeiro.

Por fim, destaco sobre a disponibilização dos dados do corpus. Ao trabalharmos com corpus, especialmente os de UGC, algumas informações serão necessárias para compreender questões sociais intrínsecas à linguagem, como gênero/sexo, idade e localidade, por exemplo. Essas informações são importantes pois auxiliam no monitoramento de perfis acerca de uma temática, observando como dado grupo de usuários se manifestam linguisticamente sobre ela. Porém, ao tratar os dados, é necessário que estratégias de anonimização (como Supressão, Generalização e/ou Pertubação, como trabalhado em Brito e Machado (2017)) sejam aplicadas, sobretudo por estarmos sob a égide da Lei 13.709, conhecida como a Lei Geral de Proteção dos Dados (LGPD). A LGPD atribui obrigações específicas a quem trata os dados quando a base legal é a pesquisa, entendendo tratamento como compilação, análise, estudo e disponibilização dos dados. Assim, é importante que nossas pesquisas ponderem se o corpus completo pode ser disponibilizado (com ou sem anonimização), se contém nele informações sensíveis (como questões relacionadas a raça, gênero, sexualidade, posições políticas, por exemplo) e por quanto tempo os dados podem ficar sob domínio do pesquisador (Almeida, 2021).

CONSIDERAÇÕES FINAIS

Neste artigo meu propósito foi discutir alguns conceitos da LC sob a luz e os desafios impostos pela Inteligência Artificial aos trabalhos atuais com corpora. É quase impossível voltarmos atrás: estamos presenciando um momento em que passamos a produzir conteúdo sobretudo na Web, e esse conteúdo autêntico disputa espaço, muitas vezes, com um conteúdo artificial. O que propus aqui, então, foi uma breve reflexão sobre como esses conteúdos podem e devem figurar em nossos estudos de descrição e de aplicação sobre a linguagem frente aos métodos e ferramentas computacionais.

Há muitos outros desafios que devem ser debatidos nesse campo e muitos outros conceitos da própria LC que merecem ser discutidos, que culminará em uma nova tipologia e organização e compreensão do objeto corpus. Alguns desses desafios já foram enfrentados em outras disciplinas próximas, como a Linguística Aplicada, quanto à anonimização dos dados, por exemplo; quanto a isso, podemos aprender e utilizar com mais facilidade. Porém, há outros que ainda estão surgindo por conta do avanço da IA; para esses, as soluções ainda estão sendo pensadas conforme os desafios surgem.

Por fim, pondero que as pesquisas em LC não se findaram por estarmos desafiados. Muito pelo contrário: temos novos caminhos a trilhar, o que traz novo

fôlego à área e outras perspectivas de pesquisa. Minha contribuição aqui é chamar a atenção para nossas metodologias de pesquisa e as concepções que estamos utilizando sobre determinados conceitos cunhados há algum tempo, antes da evolução da IA. Importa dizer, então, que o mundo que conhecemos pela literatura clássica em LC mudou por conta do computador; nossos conceitos sobre ela também devem ser repensados.

Agradecimentos: Agradeço ao Programa de Pós-Graduação em Língua e Cultura (PPGLinC) da Universidade Federal da Bahia (UFBA)pelo suporte e apoio.

REFERÊNCIAS

ALMEIDA, F.F. **Guia de proteção de dados pessoais**: pesquisa. CEPI FGV Direito SP, 2021.

AYAPOVA, S. M.; SKRIPNIKOVA, A. I. Ai and human created media texts: experiment results. **Herald of journalism**, v. 64, n. 2, 2022. DOI: https://doi.org/10.26577/hj.2022.v64.i2.08

BRITO, F.T.; MACHADO, J.C. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. **Jornadas de atualização em informática**, [*s.l.*], p. 91-130, 2017.

BURNHAM, T.F. Reconstrução-síntese de contribuições ao XI CINFORM, à guisa de apresentação. In BORGES, J; BARREIRA, M.I.J.S.; CUNHA, F.J.A.P. (Org.). **Mundo digital**: uma sociedade sem fronteiras? 1ed.João Pessoa: Ideia, 2014, v. 1, p. 7-20.

DALTE, P. Inteligência artificial e poesia. **Revista 2i: Estudos de Identidade e Intermedialidade**, v. 2, n. 2, p. 165–177, 2020. DOI: https://doi.org/10.21814/2i.2505

FREITAS, C. Dataset e corpus. *In*: CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (Orgs.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. [s.l.]: BPLN - Brasileiras em PLN, 2023, p. 1–37. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-dadoshttps://brasileiraspln.com/livro-pln/2a-edicao/parte-dados-avaliacao/cap-dataset-corpus/cap-dataset-corpus.pdfavaliacao/cap-dataset-corpus/cap-dataset-corpus.pdf>.

GARCIA, A.C. Ética e inteligência artificial. **Computação Brasil**, n. 43, p. 14-22, 2020.

HALLIDAY, M.A.K. Corpus studies and probabilistic grammar. In AIJMER, K.; ALTENBERG, B. (Orgs). **English corpus linguistics**: Studies in honour of Jan Svartvik, London: Longman, 2014. p. 42-55.

PASSARELLI, B.; GOMES, A.C.F. Transliteracias: A Terceira Onda Informacional nas Humanidades Digitais. **Revista Ibero-Americana de Ciência da Informação**, v. 13, n. 1, p. 253–275, 2020. DOI: https://doi.org/10.26512/rici.v13.n1.2020.29527

SANTOS, M.L.B. The “so-called” UGC: an updated definition of user-generated content in the age of social media. **Online Information Review**, v. 46, n. 1, p. 95– 113, 2021. DOI: https://doi.org/10.1108/oir-06-2020-0258

SARDINHA, T.B. Linguística de *corpus*: histórico e problemática. **DELTA: Documentação de Estudos em Linguística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000. DOI: https://doi.org/10.1590/s0102-44502000000200005 SINCLAIR, John. **Corpus, Concordance, Collocation**. [s.l.]: Oxford University Press, USA, 1991.

TAGNIN, S. E. A Linguística de Corpus vai desbravando novos horizontes. In

FINATTO, MJB; REBECHI, T.; SARMENTO, S; BOCORNY, A. EP (Org). Linguística de corpus: perspectivas. Porto Alegre: Instituto de Letras da UFGRS, p. 11-15, 2018.

VIANA, V.; TAGNIN, S. E. *Corpora* na tradução. São Paulo: HUB Editorial, 2015.

WU, S.; IRSOY, O.; LU, S.; *et al*. **BloombergGPT: A Large Language Model for Finance**. arXiv.org. Disponível em: <https://arxiv.org/abs/2303.17564>.

ZHAO, Bo. Web Scraping. *In*: **Encyclopedia of Big Data**. Cham: Springer International Publishing, 2022, p. 951–953. DOI: http://dx.doi.org/10.1007/978-3-319http://dx.doi.org/10.1007/978-3-319-32010-6_48332010-6_483.

SATIRICORPUS.BR: A *CORPUS* OF SATIRICAL NEWS FOR BRAZILIAN PORTUGUESE

Gabriela WICK-PEDRO[96]
Oto Araújo VALE[97]

ABSTRACT: This paper presents **SatiriCorpus.Br**, a corpus of satirical news in Brazilian Portuguese aimed at investigating linguistic differences between satirical and real news. A subcorpus was created to compare satirical news with their real counterparts. This comparison enhances the understanding of the linguistic features that distinguish satirical humor from factual reporting. The findings offer valuable insights for corpus linguistics and natural language processing (NLP).

Keywords: satirical news; corpus linguistics; satire; linguistic features; natural language processing.

1. Introduction

This work presents the SatiriCorpus, a *corpus* of satirical news for Brazilian Portuguese, and a *subcorpus* composed of satirical news and their respective real news versions, with the purpose of investigating the main differences between these two types of content. Additionally, morphosyntactic aspects are analyzed and described, as well as the differences in verbal occurrences between satirical and real news.

Satirical news have a fictional nature and function as parodies of real events and news, usually using humor, irony, exaggeration, and ridicule to criticize social, political, and cultural issues. However, unlike deceptive content, or popularly known as "fake news," which intentionally disseminates false information to deceive, manipulate, and harm or favor certain agendas, satirical news seek to provoke laughter or amusement in their audience (RUBIN; CHEN; CONROY, 2015; WARDLE; DERAKHSHAN, 2018; TANDOC; LIM; LING, 2018).

Although satirical news are often created to be humorous, there is a risk of some people confusing satirical content with factual information. Moreover, such news can intentionally mislead less attentive readers or those without contextual and cultural knowledge into believing what they are reading (RUBIN *et al.*, 2016). Therefore, it is essential for people to be alert and capable of distinguishing between satirical and non-satirical content to avoid the spread of misinformation.

2. Theoretical Framework

Satire is a literary or artistic genre that uses elements such as humor, irony, exaggeration, and ridicule to criticize social, political, cultural, or individual

[96] Pesquisadora em Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília/DF, bolsista FINATEC. E-mail: gabrielawick@ibict.br.

[97] Docente do Departamento de Letras (DL) na Universidade Federal de São Carlos (UFSCAR), São Carlos/SP.

aspects, aiming to sensitize the public and stimulate reflection on important societal issues (KREUZ; ROBERTS, 1993; SIMPSON, 2003; ATTARDO, 2014). This form of expression can be found in various forms and media, such as literary works, theatrical plays, cartoons, comic strips, television programs, and news (LAMARRE; LANDREVILLE; BEAM, 2009; SINGH, 2012).

Simpson (2003) explores satire as a discursive genre and provides a detailed analysis of the elements that compose this type of humor, examining how satire uses linguistic, stylistic, and rhetorical resources to create political and social criticisms in a satirical manner. For this purpose, he also presents a triad for the configuration of satire based on three positions of the discursive subject: the satirist (the producer), the "satiratee" (the recipient), and the satirized (the target).

According to Simpson (2003), the satirist and the satiratee are two "legitimate" participants, while the satirized is unwelcome in satirical discourse. This target can be an individual, an event, an experience, or even a discourse. The author also highlights that the positions of the subjects in the triad are subject to constant shifts and (re)organizations, which may change during the satirical discourse, depending on changes in the focus of criticism or the humor represented. Thus, it is possible to observe the dynamic and fluid nature of satire, where the positions of the participants are constantly transforming to convey their satirical messages.

3. Methodology

3.1. Construction of SatiriCorpus.Br

SatiriCorpus is a *corpus* of satirical news automatically extracted[98] from the Sensacionalista website[99], a Brazilian electronic news program that satirizes various topics, such as politics and entertainment in Brazil. For its construction, a crawler automatically collected news from the Sensacionalista website, extracting the text of the news body from the page and excluding noise such as tags, HTML, and images.

Following the thematic classification established by the website, the *corpus* was divided into five categories: i) behavior (human behavior; daily life in society) with 56 news articles, 8,014 tokens, 6,947 types, and 267 sentences; ii) entertainment (celebrities; topics from Brazilian and international TV) with 1,406 news articles, 255,972 tokens, 218,531 types, and 10,720 sentences; iii) sports (Brazilian and international sports) with 738 news articles, 120,332 tokens, 103,568 types, and 4,931 sentences; iv) world (international politics) with 1,001 news articles, 178,185 tokens, 158,555 types, and 8,250 sentences; v) country (Brazilian politics) with 1,847 news articles, 316,588 tokens, 271,692 types, and

[98] The corpus extraction period was in January 2019, thus considering the beginning of news postings on the Sensacionalista website from 2016 until the end of 2018.

[99] Available at: https://www.sensacionalista.com.br/. Accessed on: January 22, 2023.

12,347 sentences. In total, the *corpus* contains 5,048 news articles, 879,091 tokens, 759,293 types, and 36,515 sentences.

It is important to note that although there are other portals, such as Piauí Herald[100] and O Bairrista[101], which are also dedicated to satirical journalism, the preference for Sensacionalista, founded in 2009, is justified by its status as the main representative of this type of content in Brazil.

3.2. Construction of the subcorpus

In accordance with the criteria established by Rubin, Chen, and Conroy (2015) for the construction of a fake news corpus, particular emphasis is placed on the alignment between fake and real news, with the objective of verifying positive and negative instances and validating linguistic patterns. Thus, the SatiriCorpus, described in the previous section, was divided into a *subcorpus* composed of 300 news articles, with 150 satirical news articles randomly selected from the "country" category and 150 real news articles related to the satirical news. For the real news, the collection was done manually, first delimiting keywords identified in the satirical news and then manually searching for each real news equivalent to the satirical news.

The *subcorpus* contains 22,993 tokens, 4,843 types, and 1,212 sentences for satirical news, and 107,133 tokens, 11,304 types, and 5,721 sentences for real news. In total, there are 130,096 tokens, 16,147 types, and 6,933 sentences.

Additionally, the morphosyntactic information comes from the parser PALAVRAS (BICK, 2000). Besides syntactic annotations, the tool marks the grammatical class for each word. There are 15 classes in total: adjective, adverb, determinant, compound element, interjection, coordinating conjunction, subordinating conjunction, noun, numeral, personal pronouns, proper nouns, preposition, specifiers, and verbs.

4. Discussion

Based on the data extracted by PALAVRAS (BICK, 2000), the average between grammatical class and the total number of words was calculated. There is a balance of grammatical classes between satirical and factual news. The use of adverbs (5.69% in satirical news and 4.24% in real news), determinants (10.30% in satirical news and 9.55% in real news), and verbs (17.98% in satirical news and 15.06% in real news) occurs proportionally more in satirical news, while prepositions (18.35% in real news and 17.50% in satirical news) and punctuation (15.42% in real news and 13.12% in satirical news) are more relevant in real news.

The analysis of verb tenses was also conducted to find specific characteristics between the news types. However, there is no verb tense with a higher predominance in relation to the news; there is only a higher percentage of infinitive in satirical news (21.35%) compared to real news (16.45%). One

---

[100] Available at: https://piaui.folha.uol.com.br/. Accessed on: January 22, 2023.

[101] Available at: https://obairrista.com/. Accessed on: January 22, 2023

possibility is that real news tend to use more auxiliary verbs compared to satirical news, but the PALAVRAS parser does not have a specific tag for auxiliary verbs. The percentage ratio was calculated between the frequency of each verb tense and the total number of verbs annotated by the parser.

Regarding the analysis of verbal persons, it was expected that real news would have a higher incidence of verbs in the third-person singular and plural because, as Tavares (1997, p. 130–131) indicates, "journalistic text is characterized by the impersonality of the subject," meaning that "the verbal person that refers to the referent (the one being talked about – 'he,' 'they') allows the text to be more objective." The author also points out that the use of the first and second person is not expected in journalistic texts because they make the text more subjective and personal. Thus, based on the data obtained by PALAVRAS, it is noted that satirical news, although not based solely on reality, have a higher incidence of verbs in the first and third-person singular, while real news have more verbs in the first and third-person plural.

5. Conclusions

This study presented the SatiriCorpus, a *corpus* of news for Brazilian Portuguese, and investigated morphosyntactic patterns present in satirical and real news to compare linguistically how they behave.

It is understood that the characteristics extracted by the PALAVRAS parser (BICK, 2000) did not present very significant results, but they can still be seen as indications of a satirical news, such as the higher verbal and adverbial incidence.

As future work, it is hoped to create a parallel *corpus* of real news for the remaining 4,898 satirical news.

References

ATTARDO, Salvatore. Encyclopedia of humor studies. Los Angeles: SAGE Reference, 2014.

BICK, Eckhard. The Parsing System Palavras: Automatic Grammatical Analysis. Aarhus Denmark; Oakville, Conn: Aarhus University Press, 2000.

KREUZ, Roger J.; ROBERTS, Richard M. On satire and parody: The importance of being ironic. *Metaphor and Symbolic Activity*, Routledge, v. 8, n. 2, p. 97–109, 1993.

LAMARRE, Heather; LANDREVILLE, Kristen; BEAM, Michael. The Irony of Satire. *International Journal of Press-politics*, v. 14, p. 212–231, 2009.

LEAL, Sidney Evaldo. Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular. 2021. Tese (Doutorado) — Universidade de São Paulo. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-16072021-115303/. Acesso em: 29 set. 2024.

RUBIN, Victoria et al. Fake news or truth? using satirical cues to detect potentially misleading news. In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, San Diego, California: Association for Computational Linguistics, 2016. p. 7–17.

RUBIN, Victoria L.; CHEN, Yimin; CONROY, Nadia K. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, v. 52, n. 1, p. 1–4, 2015.

SIMPSON, Paul. On the Discourse of Satire. Amsterdam: John Benjamins Publishing Company, 2003.

SINGH, Raj Kishnor. Humour, irony and satire in literature. v. 3, n. 4, p. 63–72, 2012.

TANDOC, Edson C.; LIM, Zheng Wei; LING, Richard. Defining Fake News: A typology of scholarly definitions. *Digital Journalism*, v. 6, n. 2, p. 137–153, 2018.

TAVARES, Maria Alice. O verbo no texto jornalístico: notícia e reportagem. *Working Papers em Linguística*, n. 11, p. 123-142, 1997.

WARDLE, Claire; DERAKHSHAN, Hossein. Thinking about information diso rder:formats of misinformation, disinformation, and mal-information. p. 12, 2018.

FRASEOLOGIA, LINGUÍSTICA DE CORPUS, TRADUÇÃO DE EXPRESSÕES IDIOMÁTICAS E LEXICOGRAFIA: PARCERIAS DE SUCESSO

Isabela MOREIRA DE OLIVEIRA[102]

ABSTRACT: In this study, we analyzed eighty idioms related to food in Brazilian Portuguese to better understand their meaning in context and, based on that, propose translation strategies and equivalents in English. Our goal was to present a translator-oriented, bilingual (Portuguese - English) glossary of idiomatic expressions (IEs). For each entry, we provide a definition, authentic examples, synonyms, and suggest equivalents, with authentic usage examples. We have also investigated their degree of idiomaticity and fixity and attested their use in general language corpora for both languages. What makes this material unique, though, is the fact that it can be consulted electronically by semantic fields / themes, based on the 600+ distinct words we used to describe and categorize each IE from an onomasiological standpoint. We hope this study can contribute to advancing translation of PortugueseEnglish IEs and, perhaps, inspire the creation of new methodologies and lexicographic products aimed at translators.

KEYWORDS: Portuguese-English contrastive studies; semantic fields; idiomatic expressions; Corpus Linguistics; onomasiology; translator-oriented glossary.

Toda vez que compartilho com alguém os detalhes de como foi feita a nossa pesquisa de mestrado, as pessoas se envolvem; no início achava exagerada essa reação entusiasmada, mas com o tempo percebi que a pesquisa envolvia áreas de conhecimento muito relevantes para o dia-a-dia de todos, que nossa pesquisa tinha muita aplicabilidade; é possível perceber isso a partir dos pilares que compõem nosso estudo: a **Fraseologia**, a **Linguística de Corpus**, a **tradução de expressões idiomáticas** e a **Lexicografia** - está certo que em grande parte das vezes as pessoas não sabiam dar nome àquilo sobre o que estávamos conversando, mas nem por isso se mostravam menos cativadas ao perceber o quão tangível era o assunto. Com os pilares da pesquisa firmados, estabelecemos como objetivo elaborar um glossário bilíngue português brasileiro (PB) - inglês americano (EN) de expressões idiomáticas (EIs) com a temática alimentação[103] – que é, reconhecidamente, uma área que carrega marcas culturais muito fortes (TEIXEIRA, 2003), de consulta onomasiológica e online que tivesse tradutoras/es como principais consulentes, mas que poderia também ser usado por demais aprendizes e estudiosos de inglês como L2, assim como poderia ser usado de maneira bidirecional.

Para que o objetivo pudesse ser alcançado, precisamos investigar cada um dos pilares da nossa pesquisa mais a fundo. Para chegarmos na definição

[102] Mestra em Estudos da Tradução pela Universidade de Brasília (UnB), docente temporária do Departamento de Línguas Estrangeiras e Tradução (LET) da UnB e doutoranda do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS). isabela.oliveira@unb.br

[103] O glossário bilíngue, de consulta onomasiológica e online, apresentado no APÊNDICE 1: *Glossário Português-Inglês de Expressões Idiomáticas com a Temática Alimentação*, pode ser acessado a partir do link: https://tinyurl.com/rcetsv6y

do nosso objeto de estudo (i.e. as EIs) investigamos o que compreende a Lexicologia, Terminologia, **Fraseologia** e Paremiologia. De maneira sucinta, segundo Barros (2004, p. 61), a Lexicologia estuda a palavra no nível do sistema linguístico (língua global) e a Terminologia a estuda em nível da(s) norma(s) de universos de discursos especializados (línguas de especialidade).

Saliba (2000) afirma que a Fraseologia compreende unidades lexicais (UL) formadas por duas ou mais palavras gráficas, podendo chegar à extensão de uma oração e chamadas então de unidades fraseológicas (UF) ou fraseologismos; a Paremiologia, segundo Riva (2009), pode ser vista como uma subdivisão dentro dos estudos fraseológicos e assim como a Fraseologia, tem como objeto de estudo UF. Zavaglia (2006) afirma que as UL maiores que a palavra são UF, a autora estabelece também que as UF compreendem diversos tipos de combinações estáveis e que se caracterizam por sua fixidez e idiomaticidade (ZAVAGLIA, 2017). Depois de estabelecer essas convergências e divergências, definimos nosso objeto de estudo - as expressões idiomáticas (EIs) - pela caracterização proposta por Tagnin (2005), que trabalha com UF e sua interface com a Tradução. Segundo a autora, uma expressão idiomática nada mais é do que uma UF que se caracteriza pela convencionalidade (quando uma UF torna-se consolidada pelo uso) e pela idiomaticidade (o significado da UF não pode ser deduzido através da soma de significados de seus componentes).

A essa altura, já tínhamos coletado mais de uma centena de EIs com a temática alimentação com a ajuda da internet, de amigos e familiares; dessa coletânea, selecionamos 80 EIs em PB para compor nosso banco de dados (Figura 1) e começamos a desenvolver nossa ficha de coleta em formato de planilha eletrônica para propor equivalentes em EN, que seriam então apresentados em forma de glossário bilíngue PB - EN de EIs com a temática alimentação, de consulta onomasiológica e online.

Figura 1. Excerto dos campos de coleta no banco de dados.

A1 fx Cód

	A	B	C	D	E	F	G	H	I
1	**Cód**	**EI PB**	**Signif (fonte)**	**Corpus PB**	**Freq PB**	**Ex uso PB (fonte)**	**Ano pub**	**Variante(s)**	**grau fixidez**
2	1	Algo ser mamão com açúcar	Coisa muito fácil. (https://tinyurl.com/c67a6p9e)	NOW	106	Depois de batida na Áustria, Bottas aplaude brita: "Não é para ser **mamão com açúcar**"(...)"Você freia muito tarde na curva 4, você sabe que vai para a brita. Você freia muito forte e muito rápido na 6, você sabe que está na brita, o mesmo com a 7. Eles também colocaram uma zebras sérias nas curvas 9 e 10. Isso é positivo. Você não deveria escapar e voltar assim, **mamão com açúcar**, sabe?", encerrou.	2019		médio
3	2	Algo ser batata	"É ISSO MESMO!" "É batata" é a expressão perfeita quando se quer definir que algo é certeiro e que não tem chance de errar. Ou seja: se "é batata", não precisa nem ter dúvidas! E no mundo culinário, a gente sabe bem como as batatas são versáteis, práticas e deliciosas, transformando pratos comuns em verdadeiras delícias! (https://tinyurl.com/2z7ajfb3)	NOW	3	Essa **é batata**. James Bond é o personagem mais vezes trocado dentro de uma mesma franquia (sem contar Drácula, que caiu em domínio público e aparece em qualquer filme de diferentes franquias, e Jason de Sexta-Feira 13 – que com o uso de maquiagem e sem falar, pode ser vivido por qualquer um). (https://tinyurl.com/3nheurfp)	2015		médio

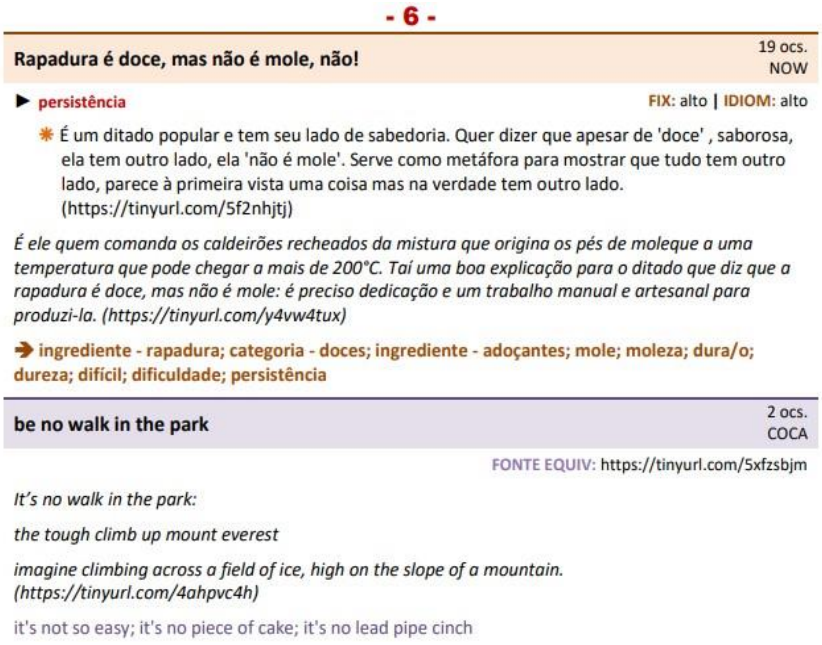
Fonte: Moreira de Oliveira, 2022, p. 58.

Nesse percurso, aplicamos as estratégias de **tradução de EIs** propostas por Baker (1992) e fomos aprimorando nossa ficha de coleta, que por fim era composta por 26 colunas por linha (i.e por EI); em relação aos equivalentes, definimos que apresentaríamos opções de EIs tanto em PB e EN atuais e

convencionais. Acredito que esse objetivo justifica fundamentalmente a nossa abordagem de desenvolver uma ficha de coleta tão extensa, tão detalhada. Para o preenchimento da ficha, tivemos que investigar mais profundamente outro pilar que compõe nossa pesquisa: a **Linguística de Corpus** (LC). Em sua interface com a Tradução, a LC mostrou-se uma grande aliada, já que utilizamos estes três corpora monolíngues como fonte de consulta para atestar o uso das EIs (PB e EN): *Corpus of Contemporary American*

English - COCA e dois subcorpora do *Corpus do Português*, o *Web/ dialects* e o *Now*, todos desenvolvidos pelo mesmo pesquisador, Mark Davies e disponíveis para acesso online e gratuito[104]. Com base na composição da ficha de coleta e no estudo mais aprofundado da **Lexicografia**, definimos a microestrutura do nosso glossário: são 80 verbetes (Figura 2) compostos pelos campos preenchidos nas fichas de coleta, dos quais destacamos grau de fixidez, grau de idiomaticidade e exemplos autênticos de uso nos dois idiomas como campos de extrema relevância para nosso consulente alvo (tradutoras/es); já em relação à sua macroestrutura, os verbetes foram organizados de maneira onomasiológica, i.e. as EIs são consultadas a partir de seus campos semânticos em direção às EIs - esses campos foram definidos no decorrer do preenchimento da ficha de coleta (que resultou em aproximadamente 600 palavras distintas[105]) e acabamos por escolher a consulta ao glossário de forma onomasiológica por priorizar relações de sinonímia por vezes perdidas pela organização semasiológica (que parte das UL em direção ao conceito) e também para que, caso o consultante se esqueça ou desconheça as UL que compõem a EI que está procurando, ainda possa encontrá-la no glossário.

Figura 2: Exemplo de verbete.



- 6 -

Rapadura é doce, mas não é mole, não! 19 ocs. NOW

▶ persistência FIX: alto | IDIOM: alto

✱ É um ditado popular e tem seu lado de sabedoria. Quer dizer que apesar de 'doce' , saborosa, ela tem outro lado, ela 'não é mole'. Serve como metáfora para mostrar que tudo tem outro lado, parece à primeira vista uma coisa mas na verdade tem outro lado. (https://tinyurl.com/5f2nhjtj)

É ele quem comanda os caldeirões recheados da mistura que origina os pés de moleque a uma temperatura que pode chegar a mais de 200°C. Taí uma boa explicação para o ditado que diz que a rapadura é doce, mas não é mole: é preciso dedicação e um trabalho manual e artesanal para produzi-la. (https://tinyurl.com/y4vw4tux)

➔ ingrediente - rapadura; categoria - doces; ingrediente - adoçantes; mole; moleza; dura/o; dureza; difícil; dificuldade; persistência

be no walk in the park 2 ocs. COCA

FONTE EQUIV: https://tinyurl.com/5xfzsbjm

It's no walk in the park:

the tough climb up mount everest

imagine climbing across a field of ice, high on the slope of a mountain. (https://tinyurl.com/4ahpvc4h)

it's not so easy; it's no piece of cake; it's no lead pipe cinch

Fonte: Moreira de Oliveira, 2022, p. 60.

[104] Disponíveis em: https://www.english-corpora.org/coca/ e https://www.corpusdoportugues.org/.
[105] O APÊNDICE 3 *Palavras usadas para caracterizar as temáticas e campos semânticos* pode ser acessado a partir do link: https://tinyurl.com/rcetsv6y

Considerando que o percurso de tradução compreendido pela ficha de coleta que desenvolvemos, aprimoramos e aplicamos na nossa pesquisa foi extenso e meticuloso, conseguimos alcançar bons resultados. O caminho foi exaustivo e demorado, pois foram 80 EIs em PB e seus equivalentes em EN organizados em verbetes e para que chegássemos a cada verbete, 26 colunas foram preenchidas para cada um deles. Outro aspecto que garantiu resultados relevantes e confiáveis foi a consulta aos corpora para atestar o uso das EIs em PB e EN. Nos deparamos com alguns desafios que nos levaram a reflexões que nos guiarão para possíveis melhorias futuramente: usar os corpora, embora extremamente útil para disponibilizar exemplos convencionados de uso, não foi muito eficiente para EIs com grau de idiomaticidade alto, ou baixo grau de fixidez; outro questão foi o limite de consultas estabelecido pelos corpora em suas versões disponíveis online gratuitas, que permitem somente 50 consultas por dia; ao pesquisar colocados que tem ambos significados idiomático e denotativo, os corpora nos apresentavam resultados de ambos e precisamos então filtrá-los; em várias ocasiões, os corpora em PB e em EN nos levaram a páginas inexistentes ou sem relação com o conteúdo alvo; tivemos dificuldade por vezes em atestar o uso de EIs por serem parte da língua falada, e não escrita como os corpora consultados; tivemos que acrescentar à nossa ficha de coleta as colunas “sinônimos” e “variantes” já que nos deparamos diversas vezes com mais de uma opção de equivalente.

A abordagem contrastiva nos permitiu chegar a algumas conclusões que caracterizam as EIs que compõem o glossário: 59% das EIs tem grau de fixidez médio, 37% tem grau alto de fixidez, e 3% grau baixo - EIs com baixo e médio grau de fixidez se misturam mais facilmente no texto, isso pode indicar a dificuldade que tradutores e falantes “ingênuos” (TAGNIN, 2005) tem ao identificá-las em contexto; EIs com grau de fixidez alto apresentam menos variantes quando comparadas àquelas de grau baixo e médio de fixidez, o que também se aplica a flexão de gênero e número; quanto ao grau de idiomaticidade, 49% das EIs tem nível de idiomaticidade alto, 37% médio e somente 13% baixo, esses números podem justificar nossas escolhas por equivalentes ligados ao significado conotativo e não ao denotativo. Ao observarmos as palavras que compõem os campos semânticos, observamos que o campo semântico *ingredientes* aparece em aproximadamente 44% das entradas, o campo semântico dos *pratos,* está presente em cerca de 30% das entradas; os ingredientes que mais apareceram foram *frutas* (10%), *carboidratos* (8%), *tubérculos* (6%) e *carnes* (6%), podemos então perceber que são alimentos que compõem a base da alimentação dos brasileiros; dentre as frutas, são as *frutas tropicais* (10%) que aparecem mais; quanto aos *animais*, que aparecem em 8% das entradas, as *aves* ocorrem em 5% e *porcos* 4%, o que indica que são animais comuns na cultura brasileira, que podem ser criados em nossos quintais; *parte do corpo*, presente em cerca de 19% das entradas nos surpreendeu devida a sua grande ocorrência (não esperávamos uma porcentagem tão alta mesmo que *estômago* e *boca* sejam indicativos da temática alimentação; dentre as palavras que não têm relação com a temática alimentação, *problema* ocorre em 10% das EIs, *superação* em 9% e *dificuldade* ocorre em 6%, o que nos leva ao caráter de ensinamentos ancestrais expressados pelas EIs.

Ao empregar as estratégias propostas por Baker (1992), apresentamos EIs equivalentes em significado, mas não na forma:

54) *ALGUÉM DAR com a língua nos dentes* | *spill the beans*

24) *ALGO ACABAR em pizza* | *come to nought*

52) MANDAR ALGUÉM catar coquinho(s) | *go take a long walk on a short pier* Algumas EIs que tem equivalentes muito similares em relação a sua forma e conteúdo em ambas as línguas:

10)ALGO SER a *cereja* do *bolo* | the *cherry* on the *cake*

11)(não adianta) *chorar* (sobre) o *leite derramado* | *cry* over *spilled milk*

49) Se/Quando a *vida* DAR A ALGUÉM um *limão/limões*, faça *limonada* | when *life* gives you *lemons*, make *lemonade*

Diante disso, concluímos que é possível encontrar EIs que são compartilhadas pelo mundo por conterem ensinamentos ou perspectivas universais sobre determinado assunto e por isso foram adotadas por outras culturas. Por fim, a parceria entre LC e tradução de fraseologismos foi feliz - aliás, construir o glossário nos levou a entender melhor a LC e sua interface com a tradução. Compartilhamos a metodologia aplicada na direção PB > EN[106], que pode ser também usada na direção inversa – esperamos que estudos na área continuem a se expandir.

Agradecimentos: À organização do ELC/ EBRALC 2024, por esse evento tão necessário.

Referências

BAKER, M. *In other words.* Abingdon, Oxon ; New York, NY: Routledge, 1992.

BARROS, Lídia A. *Curso básico de terminologia*. São Paulo: EDUSP, 2004.

RIVA, H. C. Dicionário onomasiológico de expressões idiomáticas usuais na língua portuguesa no Brasil. São José do Rio Preto: 2009, 311 f. *Tese*

(doutorado em Estudos Linguísticos) – Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista.

SALIBA, M. C. Unidades lexicais maiores que a palavra: descrição linguística, considerações psicolinguísticas e implicações pedagógicas. *Dissertação* (mestrado) - Universidade Federal do Paraná. Curitiba: 23/08/2000. Disponível em: https://acervodigital.ufpr.br/handle/1884/24439. Acesso em: 07 ago. 2020.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. São Paulo: Disal, 2005. 223p.

[106] A metodologia completa desta pesquisa está disponível em: https://tinyurl.com/rcetsv6y.

TEIXEIRA, E.D. Em busca de um novo modelo tecno-formal para a construção de dicionários técnicos bilíngues - o exemplo da culinária. *Intercâmbio* (PUCSP), São Paulo, SP, v. XII, p. 243-251, 2003.

ZAVAGLIA, C. Dicionário e Cores. *Alfa*, São Paulo, 50 (2): 25-41, 2006.

ZAVAGLIA, C.; FROMM, G. Fraseologia e Paremiologia: uma entrevista com Claudia Zavaglia. *ReVEL*, v. 15, n. 29, 2017. Disponível em: https:// revel.inf.br. Acesso em: 15 set. 2020.

INPACT - INTERNACIONALIZAÇÃO DA PRODUÇÃO ACADÊMICA COM CORPUS E TECNOLOGIA: A CONSTRUÇÃO DE UMA FERRAMENTA ON-LINE PARA A ESCRITA DE ARTIGOS DE PESQUISA EM INGLÊS NAS HUMANIDADES

Ana Eliza Pereira BOCORNY[107]
Deise Prina DUTRA[108]

RESUMO: O projeto InPACT visa internacionalizar a produção científica brasileira nas Humanidades, desenvolvendo uma ferramenta online de suporte à escrita acadêmica em inglês baseada em linguística de corpus. A ferramenta utiliza elementos fraseológicos extraídos de corpora para auxiliar pesquisadores a atenderem às convenções linguísticas e retóricas das suas disciplinas. Testes preliminares indicam que a ferramenta é eficaz e intuitiva.

Palavras-chave: Linguística de Corpus; Gêneros Acadêmicos; *Lexical Bundles*; *Lexical Frames*.

INTRODUÇÃO

O projeto InPACT tem como objetivo principal contribuir para a internacionalização da produção científica brasileira nas áreas de Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes. Para tanto, o projeto propõe o desenvolvimento de recursos pedagógicos on-line baseados em corpus, que possam auxiliar pesquisadores e estudantes na produção de artigos científicos em inglês. Uma ferramenta on-line de suporte à escrita acadêmica é um dos recursos que está sendo desenvolvido no âmbito do projeto. É importante destacar que diversos trabalhos têm explorado o desenvolvimento de ferramentas para apoiar a produção textual acadêmica, tanto em língua portuguesa quanto em outras línguas.

No Brasil, recursos como o AMADEUS (ALUÍSIO *et al*., 2001), o SciPo (FELTRIN, *et al*., 2003) e o SciPo - Farmácia (SCHUSTER *et al*., 2005) surgiram como esforços pioneiros nessa direção. Internacionalmente, temos exemplos como o AWSUM (MIZUMOTO, et al., 2017) e o Collocaid (FRANKENBERG-GARCIA *et al*., 2019). Neste contexto, a ferramenta que está sendo desenvolvida no âmbito do projeto InPACT emerge como uma resposta às necessidades específicas da produção científica de artigos de pesquisa em inglês por parte da comunidade acadêmica brasileira das áreas-alvo, considerando suas particularidades linguísticas e os padrões discursivos das seções mencionadas.

A presente proposta se alinha a iniciativas passadas e busca avançar na integração da Linguística de Corpus e da tecnologia em prol da escrita acadêmica em inglês, especialmente por buscar elementos fraseológicos recorrentes (*lexical bundles* e *lexical frames*) nos corpora compilados e relacioná-los às funções retóricas que os mesmos realizam nas diferentes

[107] Professora do Magistério Superior, Porto Alegre - RS, Universidade Federal do Rio Grande do Sul (UFRGS). E-mail: ana.bocorny@gmail.com

[108] Professora do Magistério Superior, Belo Horizonte - MG, Universidade Federal de Minas Gerais (UFMG).

seções das 16 disciplinas-alvo[109], no âmbito das Humanidades.

Neste contexto, o presente projeto constituiu-se a partir de pressupostos teóricos de diferentes áreas do conhecimento. Aqui, damos destaque aos (i) estudos sobre gêneros do discurso e (ii) aos princípios da Linguística de Corpus. A concepção de gênero que fundamenta esta pesquisa está alinhada com as perspectivas de Bakhtin (1997), Swales (1990) e Bathia (1997). De Bakhtin (1997), ressalta-se a concepção de gênero do discurso como "tipos relativamente estáveis de enunciados". De Swales (1990, p. 46), destaca-se a ideia de que os gêneros "são veículos de comunicação para atingir um objetivo". Por fim, de Bathia (1997, p.160), sublinha-se o entendimento de análise de gênero como sendo "o estudo do comportamento linguístico situado em contextos acadêmicos ou profissionais".

A Linguística de Corpus parte de uma perspectiva de descrição da língua em uso, seja ela geral ou especializada. A visão da língua como um sistema probabilístico é um dos fundamentos principais da Linguística de Corpus (BERBER SARDINHA, 2004). Assim, os traços linguísticos (lexicais, estruturais, pragmáticos e discursivos) não ocorrem todos com a mesma regularidade (BERBER SARDINHA, 2004). Por esse motivo, a variação dos traços não é aleatória; pelo contrário, existe "um mapeamento regular entre a frequência maior ou menor de um traço e um contexto de ocorrência" (BERBER SARDINHA, 2004, p. 351). Dessa forma, defender que os traços não são aleatórios significa dizer que "a linguagem é padronizada. A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura que se repete significativamente mostra sinais de ser, na verdade, um padrão lexical ou léxico-gramatical." (BERBER SARDINHA, 2004, p. 31).

Por fim, é importante ressaltar que, ao propor a identifição das formas linguísticas que realizam as funções retóricas expressas nas seções de artigos de pesquisa em inglês, este projeto busca preencher um "gap" relatado por Moreno e Swales (2018) e Gray *et al*. (2020). Os referidos autores afirmam que, ainda hoje, poucos estudos são realizados a partir da combinação da perspectiva e princípios dos estudos sobre gêneros do discurso e da Linguística de Corpus para investigar as realizações linguísticas de movimentos retóricos de diferentes gêneros do discurso.

METODOLOGIA

Este estudo conta com uma equipe multidisciplinar que inclui linguistas de corpus, especialistas em *design* e cientistas da informação. A partir desses saberes o desenvolvimento da ferramenta valeu-se de uma combinação de etapas metodológicas que envolveram as três áreas.

A primeira etapa tratou da definição dos objetivos da ferramenta, da identificação dos usuários alvo e de suas necessidades. Como parte dessa etapa

---

[109] **Ciências Humanas (10):** Filosofia (Phil), Sociologia (Soc), Antropologia (Ant), Arqueologia (Arc), Geografia (Geo), Psicologia (Psy), Educação (Edu), Religiões (RelF), Demografia (Dem), **Ciências Sociais Aplicadas (4):** Direito e Ciências Jurídicas (LawLS), Economia (Eco), Comunicação (Com), Ciência Política (PolS), Políticas Públicas (PubP), **Linguística, Letras e Artes (2):** Linguística (Ling), Letras (Lang).

inicial foi criado o *naming* (processo de criação e desenvolvimento de nomes para marcas, produtos, serviços, empresas e outras entidades) e a identidade visual do projeto.

A segunda etapa foi a seleção e organização dos textos utilizados para extração de dados linguísticos usados para informar a construção da ferramenta. Nesta etapa também buscou-se a identificação da estrutura retórica das diferentes seções dos artigos científicos das áreas de Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes.

A terceira etapa foi a definição da arquitetura da ferramenta e a elaboração do design da interface. Nesta etapa, foram realizados testes de usabilidade.

A quarta etapa foi o desenvolvimento de um protótipo da ferramenta e a inserção dos dados linguísticos extraídos na etapa dois. Foram utilizadas tecnologias de programação *web* e bancos de dados linguísticos para criar uma plataforma on-line que fosse intuitiva, que pudesse ser acessada com facilidade e que oferecesse os elementos linguísticos necessários para a produção de artigos de pesquisa em inglês.

RESULTADOS E DISCUSSÃO

Como resultados da primeira etapa da construção da ferramenta tivemos a identificação de três personas que representam os usuários-alvo e suas necessidades. Após o desenvolvimento de estudo, alternativas, seleção e refinamento, o *naming* foi definido como: InPACT - Internacionalização da Produção Acadêmica com Corpus e

Tecnologia. Com o *naming* escolhido, a identidade visual foi elaborada. A Figura 1 mostra o logo definido, a inspiração para o ícone criado a partir do logo, as três aplicações do logo para resultados do projeto e as escolhas de ícones e grafismos para cada disciplina.

Os dados linguísticos extraídos na segunda etapa do estudo foram obtidos a partir do Corpus de Humanidades (CORHUM), um corpus estratificado em 16 disciplinas da área das Humanidades e em quatro seções de artigos de pesquisa (Introdução, Metodologia e Resultados / Discussão e Conclusão - IMR/DC). Com um total de aproximadamente 64 milhões de palavras, o CORHUM conta com 64 subcorpora com aproximadamente 1 milhão de palavras cada. Os subcorpora contêm textos das seções IMR/DC de artigos de pesquisa das disciplinas-alvo Esses artigos foram publicados na plataforma PLOS One, em inglês, de 2013 a 2023. Utilizamos as ferramentas Sketch Engine (KILGARRIFF *et al*., 2004) e AntConc 4.0.10 (ANTHONY, 2022) para extrair conjuntos de *lexical bundles (LBs)* e *lexical frames (LFs)* de cada subcorpus (por exemplo, o conjunto de *LBs* da seção metodologia da disciplina Educação).

Uma vez extraídos, *LBs* e *LFs* foram agrupados por similaridade lexical (unidades com um número de elementos lexicais iguais: *the aim of this paper is to* e *the objective of this paper is to*) e por similaridade retórica (unidades diferentes com funções retóricas iguais: *the aim of this paper is to* e *this paper aims to*). Uma vez agrupados, partiu-se para a análise das funções retóricas de cada conjunto de unidades fraseológicas. A relação entre forma e função se deu a partir de um *framework* representando a estrutura retórica de artigos das

disciplinas-alvo, construído a partir da revisão de 25 estudos prévios (por exemplo, ZHANG; WANNARUK, 2016; YANG; ALLISON, 2003).

Por fim, construiu-se uma base de dados com as informações coletadas que foi incorporada ao protótipo da ferramenta. A metáfora usada para a construção da ferramenta foi a do texto acadêmico como uma parede com tijolos vermelhos e azuis. Os tijolos vermelhos representando os *building blocks of discourse,* formulaicos e convencionais, sob a forma de *LBs* e *LFs* (por exemplo: '*The aim of this paper is to…')* e os tijolos azuis representando o 'conteúdo' relativo ao estudo que o pesquisador está desenvolvendo (por exemplo, '...*extract the most frequent lexical frames from a corpus of abstracts*').

A ferramenta pretende oferecer as opções correspondentes aos tijolos vermelhos relacionando tais opções aos movimentos retóricos das seções onde ocorrem. O texto referente aos tijolos azuis, diz respeito ao conhecimento prévio e aos dados da pesquisa de cada autor de cada artigo de pesquisa. A arquitetura da ferramenta e a elaboração do *design* da interface foram os resultados obtidos na terceira etapa do estudo. A partir do entendimento de que a ferramenta deveria ser capaz auxiliar na produção de um artigo de pesquisa com a estrutura retórica convencionalmente usada nas disciplinas-alvo, utilizando os *LBs* e *LFs* identificados em cada seção, os principais casos de uso da ferramenta foram identificados como: (i) redigir um texto/trecho; (ii) visualizar exemplos; (iii) consultar propósitos da ferramenta; (iv) ver tutoriais.

A partir das constatações descritas foi desenhado o fluxo do usuário simulando o caso de uso 'redigir um texto'. Definido o fluxo do usuário, iniciou-se a geração de alternativas para a ferramenta e o protótipo de uma das opções foi criado. A partir de testes realizados, foram feitas melhorias e aprimoramentos no protótipo inicial da ferramenta.

A ferramenta on-line de suporte à escrita acadêmica em inglês já está em fase de testes. O feedback recebido de pesquisadores das disciplinas-alvo têm sido positivo, indicando que a ferramenta é intuitiva, fácil de usar e eficiente no oferecimento de padrões linguísticos convencionais e específicos das diferentes disciplinas e seções dos artigos de pesquisa da área-alvo.

REFERÊNCIAS

ALUÍSIO, S. M.; BARCELOS, I.; SAMPAIO, J.; OLIVEIRA JR, O. N. How to Learn the Many Unwritten "Rules of the Game" of the Academic Discourse: A Hybrid Approach Based on Critiques and Cases to Support Scientific Writing. **IEEE International Conference on Advanced Learning Technologies**, p. 257-260, 2001.

ANTHONY, L. **AntConc** (Version *4.0.10*) [Computer Software]. Tokyo, Japan: Waseda University, 2021. Disponível em: https://www.laurenceanthony.net/software. Acesso em: 20 jul. 2022.

BAKHTIN, M. **Estética da Criação Verbal**. 2.ed. São Paulo: Martins Fontes, 1997.

BHATIA, V. K. 'Análise de gênero hoje' [Trad. Benedito G. Bezerra]. **Revue Belge de Philologie et d´Historie**, Bruxelles, v. 75, p. 629-652, 1997.

BERBER SARDINHA, T. **Linguística de Corpus**. São Paulo: Manole, 2004.

FELTRIM, V. D. *et al.* A construção de uma ferramenta de auxílio à escrita de resumos acadêmicos em português. In: **Anais do XXIII Congresso da Sociedade Brasileira de Computação**, 2003.

FRANKENBERG-GARCIA, A., REES, G., LEW, R., ROBERTS, J., SHARMA, N. AND BUTCHER, P. ColloCaid: a tool to help academic English writers find the words they need. In: MEUNIER, F.; VAN DE VYVER, J.; BRADLEY, L.; THOUËSNY, S. (Orgs.). **CALL and complexity – short papers from EUROCALL 2019**. Voillans:Research-publishing.net, 2019.

KILGARRIFF, A. *et al*. Itri-04-08 the sketch engine. **Information Technology**, v. 105, n. 116, p. 105-116, 2004.

MIZUMOTO, A; HAMATANI, S; IMAO, Y. Applying the bundle–move connection approach to the development of an online writing support tool for research articles. **Language Learning**, v. 67, n. 4, p. 885-921, 2017.

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.

SCHUSTER, E.; ALUÍSIO, S. M.; FELTRIM, V. D.; PESSOA JR, A.; OLIVEIRA JR, O.N.Enhancing the Writing of Scientific Abstracts: A Two-phased Process Using Software Tools and Human Evaluation. **XXV Congresso da Sociedade Brasileira de Computação**, p. 962-971, 2005.

SWALES, J. **Genre analysis**: English in academic and research settings. Cambridge: Cambridge University Press, 1990.

YANG, R.; ALLISON, D. Research articles in applied linguistics: Moving from results to conclusions. **English for Specific Purposes**, v. 22, p. 365-385, 2003.

ZHANG, B.; WANNARUK, A. Rhetorical Structure of Education Research Article Methods Sections. **PASAA**: Journal of Language Teaching and Learning in Thailand, v.51, p. 155-184, 2016.

ANÁLISE MULTIDIMENSIONAL ADITIVA DOS TEXTOS LEGAIS FEDERAIS BRASILEIROS

Carolina Godoi de Faria MARQUES[110]
Carlos Henrique KAUFFMANN [111]

RESUMO
Neste artigo apresentamos a Análise Multidimensional Aditiva (BIBER, 1988) do LEX-BR-Ius (FERRARI e MARQUES, 2022), um corpus de textos legais brasileiros. Para a sua realização, nosso corpus foi adicionado às dimensões de variação do português brasileiro identificadas por Berber Sardinha, Kauffmann e Acunzo (2014) e comparado com o Corpus Brasileiro de Variação e Registro (CBVR). Os resultados indicam que os textos legais são um registro informacional, letrado e orientado para o futuro.

Palavras chave: Análise multidimensional; LEX-BR-Ius; Legislação federal brasileira; Variação linguística; Linguagem jurídica.

INTRODUÇÃO

As leis, como são popularmente chamados os textos legais, são os pilares da nossa sociedade. Elas estabelecem as normas sob as quais ela se pauta, organizando-a, regulando-a e protegendo-a (SOUZA e SOUZA, 2017). Apesar da sua importância, seu texto é percebido pelo cidadão comum como rebuscado e de difícil compreensão, sendo muitas vezes necessário o auxílio de um profissional do direito para interpretá-lo, o que dificulta o conhecimento e o exercício dos seus direitos e deveres. Essa dificuldade se dá porque a linguagem utilizada nas leis - a linguagem jurídica - é altamente especializada e complexa (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018).

O estudo da linguagem jurídica é um campo de pesquisa em expansão, entretanto são poucos os estudos sobre a linguagem utilizada nas leis e até o momento da realização dessa pesquisa não identificamos corpora compilados para o estudo de textos legais brasileiros nem estudos sobre a sua variação nessa língua (TIERSMA, 1999; PONTRANDOLFO, 2012). Diante dessa lacuna, optamos por investigar a variação linguística na legislação brasileira vigente. Para tanto, partimos da hipótese de que os textos legais são um registro, conforme definição de Biber e Conrad (2009) e realizamos uma Análise Multidimensional (AMD) Aditiva (BIBER, 1988) do LEX-BR-Ius (FERRARI e MARQUES, 2022), um corpus da legislação federal brasileira por nós compilado, adicionando-o às dimensões de variação do português brasileiro (PB) identificadas por Berber Sardinha, Kauffmann e Acunzo (2014).

A ANÁLISE MULTIDIMENSIONAL

[110] Doutoranda, Universidade Federal de Minas Gerais, Belo Horizonte/MG. Bolsista CAPES (n. 939578/2024-00). E-mail: carol.godoi@outlook.com.br

[111] Doutor, pesquisador, Pontifícia Universidade Católica de São Paulo, São Paulo/SP. Bolsista de pós-doutoramento CAPES.

A Análise Multidimensional (BIBER, 1988) é uma abordagem empíricometodológica baseada em corpus para o estudo da variação linguística. Nela adotase a noção de registro, ou seja, uma variedade da língua com traços situacionais, linguísticos e funcionais próprios, utilizada em contextos comunicativos específicos (BIBER e CONRAD, 2009). Biber (1988) propõe que a existência de padrões de coocorrência de traços linguísticos em determinado registro é motivada funcionalmente cuja identificação e a subsequente comparação possibilitaria sua caracterização e descrição (BERBER SARDINHA, 2010). Para tanto, são estabelecidas dimensões de variação a partir da análise estatística de um corpus e da interpretação funcional de seus resultados (Biber 1988).

Segundo Berber Sardinha (2013a) é possível realizar seja a Análise Multidimensional completa, conforme proposta por Biber (1988), quanto a Análise Multidimensional aditiva. As diferenças entre elas se resumem na complexidade dos cálculos estatísticos necessários para a sua realização e no grau de detalhamento da descrição dos registros (BERBER SARDINHA, 2013a; et al, 2019). Ademais, enquanto a AMD completa identifica as dimensões de variação, a aditiva não permite tal identificação, se valendo das dimensões identificadas por uma AMD completa para a sua realização. Considerada por Berber Sardinha et al (2019) como mais simples e flexível, a AMD aditiva fornece um panorama dos registros em análise, obtido a partir da adição do corpus de estudo às dimensões da AMD completa que as identificou e sua comparação com o corpus utilizado para identificá-las.

METODOLOGIA

O LEX-BR-Ius

Para a realização do presente estudo compilamos o LEX-BR-Ius um corpus sincrônico composto por textos legais federais brasileiros em vigência no momento da compilação. Visando garantir uma correta representatividade e não enviesar seu conteúdo e organização interna, optou-se por coletar textos de todos os tipos legais em sua integralidade (SINCLAIR, 2004; BIBER, 1993) no Portal da Legislação, site governamental que disponibiliza as leis atualizadas gratuitamente online, e o balanceamento foi feito segundo a frequência de uso dos textos. Para propiciar buscas e análises linguísticas aprofundadas o corpus foi etiquetado morfossintaticamente com o PALAVRAS (BICK, 2000, 2014) e marcado em Modest XML (HARDIE, 2014) com etiquetas criadas por nós. Quanto à arquitetura do corpus, optamos por manter a divisão do Portal da Legislação separando os textos legais em seis seções: Constituição, Códigos, Estatutos, Emendas à Constituição, Leis complementares e Leis ordinárias., perfazendo um total de 755 normas e 3.300.289 palavras.

A AMD

Para alcançar nossos objetivos, realizamos a AMD Aditiva do nosso corpus. Para tanto, adicionamos o LEX-BR-Ius às dimensões do português brasileiro (PB) identificadas pela AMD do Corpus Brasileiro de Variação e Registro (CBVR) (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014): *(1) Oral versus literate discourse, (2)*

Argumentation, (3) Involved versus informational production, (4) Directive discourse, (5) Future versus past time orientation e *(6) Reported discourse*. Trata-se do estudo mais completo sobre a variação do PB já feito, em que foram analisados 48 registros escritos e orais dessa língua.

O primeiro passo dessa análise foi anotar e contabilizar as ocorrências dos traços linguísticos de cada uma das dimensões analisadas no nosso corpus com o etiquetador PALAVRAS (BICK, 2000 e 2014) e o pós-processador PALAVRAS Tag count (BERBER SARDINHA, 2013b), ambos adotados por Berber Sardinha, Kauffmann e Acunzo (2014) em seu estudo. Em seguida, normalizamos as ocorrências por mil palavras e calculamos seus Z-escores para que as frequências absolutas dos traços linguísticos em analise não enviesassem os dados.

O próximo passo foi calcular a carga fatorial dos textos e a partir dela a carga fatorial do corpus, o que nos permitiu localizá-lo nas dimensões do PB e compará-lo com os registros do CBVR. Para tanto, primeiramente calculamos o escore de dimensão dos textos e, em seguida, a média desses escores. Esta última foi então incorporada à tabela com as médias de dimensão dos registros do CBVR, a nós disponibilizada pelos autores do estudo. Por fim, para constatar a significância estatística dos nossos resultados, realizamos os testes ANOVA e R^2.

RESULTADOS E DISCUSSÃO

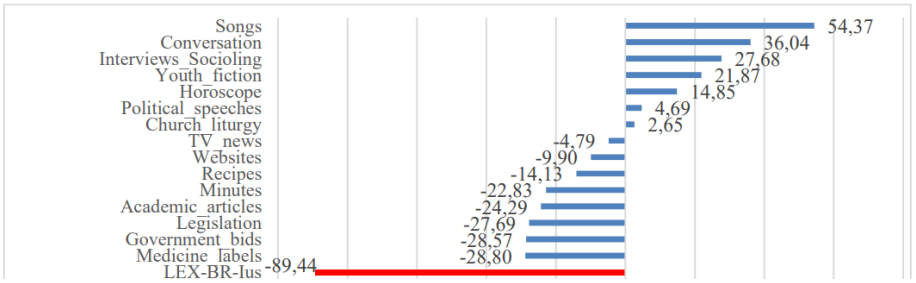
O LEX-BR-Ius pontuou e se distinguiu dos registros do CBVR significativamente em todas as dimensões. Nessa seção, por questões de espaço, apresentaremos apenas as dimensões nas quais nosso corpus se destacou (1, 3 e 5), e reproduziremos nos gráficos apenas parte dos registros abarcados pelo CBVR. Nas demais dimensões (2, 4 e 6) as pontuações obtidas giram em torno de 0. Provavelmente, como os textos legais são impositivos e visam majoritariamente informar e descrever as normas da forma mais clara e detalhada possível, o uso de traços argumentativos, diretivos e do discurso indireto é relegado.

Dimensão 1: Oral vs. literate discourse

Na dimensão 1 os registros são distribuídos segundo o seu grau de oralidade e de letramento conforme o gráfico abaixo. No polo positivo temos os

registros nos quais predominam os traços linguísticos[112] do discurso oral e no negativo aqueles do discurso letrado.

Gráfico 1: Dimensão 1



Fonte: autoras (2024)

O LEX-BR-Ius obteve a maior pontuação negativa dessa dimensão, logo, dentre os registros analisados, é o que carrega mais traços do discurso letrado. Dentre eles destacamos: orações reduzidas de gerúndio (i), passivas sem agente (ii), nominalizações em posição de sujeito (ii), particípios passados e substantivos compostos e abstratos (i, ii), como mostram os exemplos abaixo

(i) I - praticar ato visando fim proibido em lei ou regulamento ou diverso daquele previsto, na regra de competência; (LO8.429_02.06.1992)

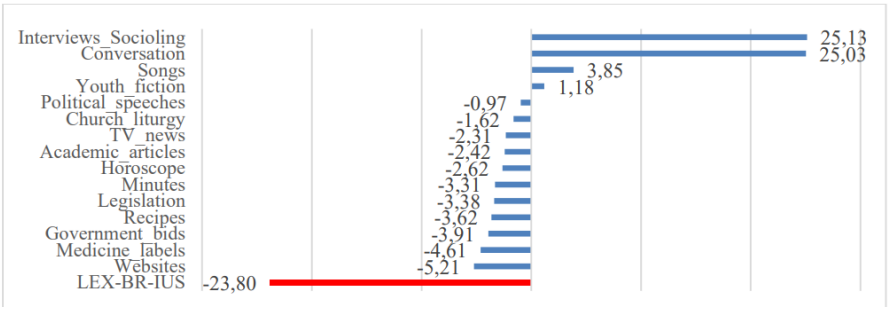
(ii) Art. 15. A participação no CONARE será considerada serviço relevante e não implicará remuneração de qualquer natureza ou espécie. (E9.474_22.12.1997) Os traços linguísticos dessa dimensão contribuem para a complexidade gramatical e densidade informacional dos textos e exercem a função de restringir e detalhar seu conteúdo, compactando um grande volume de informações fornecidas de forma técnica e concisa.

Dimensão 3: Involved vs. informal production

Nessa dimensão temos, no polo positivo, os registros marcados pela interação e, no polo negativo, os registros informacionais, sendo o único traço desse polo o *typetoken ratio*, que mede a densidade lexical dos textos.

[112] Identificados pela AMD realizada por Berber Sardinha, Kauffmann e Acunzo (2014).

Gráfico 2: Dimensão 3



Fonte: autoras (2024)

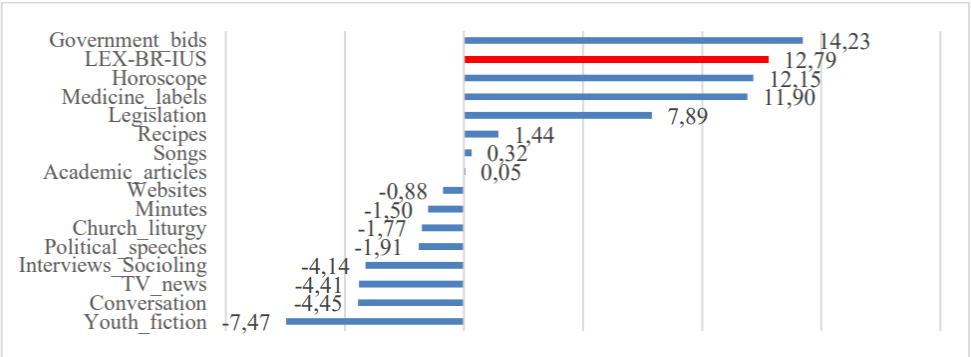
O LEX-BR-Ius obteve a maior pontuação negativa, indicada pela diversidade de vocabulário utilizado nos textos legais, que envolve termos técnicos, termos em latim, e também substantivos em geral que especificam ao máximo o sujeito do texto.

(i) Art. 2º São águas públicas de uso comum: a) os mares territoriais, nos mesmos incluídos os golfos, bahias, enseadas e portos; e) as nascentes quando forem de tal modo consideráveis que, por si só, constituam o "caput fluminis";
(C24.643_10.07.1934)

Dimensão 5: Future vs. past oriented

A dimensão 5, por sua vez, abarca a variação na orientação temporal predominante no discurso que vai do passado (polo negativo) ao futuro (polo positivo).

Gráfico 3: Dimensão 5



Fonte: autoras (2024)

Nessa dimensão nosso corpus apresentou uma das maiores pontuações positivas, sendo caracterizado por um discurso orientado para o futuro. Esse é

marcado por verbos no futuro do subjuntivo e do presente (i), modais haver (ii) e poder (ii), orações subordinadas (i), conjunções coordenadas e advérbios de probabilidade.

(i) Art. 39. O pedido passivo de cooperação jurídica internacional será recusado se configurar manifesta ofensa à ordem pública. (C13.105_16.03.2015)

(ii) Art. 10. O juiz não pode decidir, em grau algum de jurisdição, com base em fundamento a respeito do qual não se tenha dado às partes oportunidade de se manifestar, ainda que se trate de matéria sobre a qual deva decidir de ofício. (C13.105_16.03.2015)

Esses traços exercem a função de descrever e especificar como se dará a aplicação das normas e quais serão suas consequências a partir da sua promulgação.

CONCLUSÃO

A AMD aditiva nos permitiu traçar um perfil linguístico dos textos legais federais brasileiros a partir do seu mapeamento em todas as dimensões de variação do PB identificadas por Berber Sardinha, Kauffmann e Acunzo (2014). Nossos resultados indicam que os textos legais são caracterizados por um discurso letrado, de caráter informacional e orientado para o futuro, sendo que a argumentação e a diretividade exercem um papel secundário nos textos e o discurso direto é favorecido àquele indireto. Ademais, nosso corpus computou cargas fatoriais únicas e variação estatisticamente significativa em todas as dimensões analisadas, confirmando nossa hipótese de que os textos legais seriam um registro.

Agradecimentos

Essa pesquisa foi parcialmente financiada pela CAPES (bolsa nº 88887.626989/202100) e FAPEMIG (bolsas PROBIC e PIC-JR-FAPEMIG) as quais agradecemos.

Referências

BERBER SARDINHA, T. A abordagem metodológica da Análise Multidimensional. *Gragoatá*. Niterói, n. 29, p. 107-125, 2. sem. 2010. Disponível em: https://periodicos.uff.br/gragoata/article/view/33077. Acesso em: 19 jan. 2022.

BERBER SARDINHA, T. Variação entre registros da Internet. In: SHEPHERD, T. G.; SALIÉS, T. G. (Eds.). *Linguística da Internet*. São Paulo: Contexto, 2013a, p. 55–85.

BERBER SARDINHA, T. *Pós-processador PT Tag Count*. 2013b.

BERBER SARDINHA, T.; PINTO, M. V.; MAYER, C.; ZUPPARDI, M. C.;

KAUFFMANN, C. H. Adding Registers to a Previous Multi-Dimensional Analysis. In:

BERBER SARDINHA, T.; VEIRANO PINTO, M. (eds.). *Multi-Dimensional Analysis*: Research Methods and Current Issues. London: Bloomsbury, 2019. p. 165-186.

BERBER SARDINHA, T.; KAUFFMANN, C.; ACUNZO, C. M. Dimensions of register variation in Brazilian Portuguese. In: VEIRANO PINTO, M. (Eds.). *Multi-dimensional analysis: 25 years on a tribute to Douglas Biber*. John Benjamins Publishing Company, 2014.

BIBER, D. Representativeness in Corpus Design. *Literary and Linguistic Computing,* v. 8, n. 4, Oxford: Oxford University Press, p. 243-257, 1993.

BIBER, D. *Variations across speech and writing***.** Cambridge: CUP, 1988.

BIBER, D; CONRAD, S**.** *Register, genre, and style*. Cambridge: CUP, 2009.

BICK, E. PALAVRAS, a constraint grammar-based parsing system for Portuguese. In: T. SARDINHA, Berber, e FERREIRA, T. São Bento (Eds.), *Working with Portuguese corpora*. London: Bloomsbury, p 279–302, 2014.

BICK, E. *The Parsing System "Palavras":* Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dr. Phil. thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press, 2000.

CARAPINHA, C. A linguagem jurídica. Contributos para uma caracterização dos Códigos Legais*. REDIS: Revista de Estudos do Discurso*, n. 7, 2018. Disponível em:

https://ojs.letras.up.pt/index.php/re/article/view/6200. Acesso em: 10 set. 2021.

GOŹDŹ-ROSZKOWSKI, Stanisław. Legal Language. In: CHAPELLE, Carol A. (Org.*). The Encyclopedia of Applied Linguistics.* John Wiley e Sons, 2012, p. 3281-3287.

HARDIE, A. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal*, v. 38, n. 1, 73-103, 2014. Disponível em: https://doi.org/10.2478/icame2014-0004. Acesso em: 20 jul. 2021.

FERRARI, L. A.; MARQUES, C. G. F. O LEX-BR-Ius: arquitetura e decisões na compilação de um corpus representativo das leis federais brasileiras. *ANTARES*, v.14, n.34, 2022. Available at: http://www.ucs.br/etc/revistas/index.php/antares/article/view/11150/5328. Access:19 dez. 2022.

PONTRANDOLFO, Gianluca. Legal Corpora: an overview. *Rivista Internazionale di Tecnica della Traduzione*, Trieste, v. 14, p. 121-136, 2012. Disponível em: https://www.openstarts.units.it/bitstream/10077/9783/1/12Pontrandolfo.pdf. Acesso em: 6 set. 2020.

SINCLAIR, John. Corpus and Text. In: WYNNE, M (eds.). *Developing Linguistic Corpora:* a Guide to Good Practice. Oxford: Oxbow Books, 2005, p. 1-16. Disponível em: 6 set. 2020.

SOUZA, C. F. de; SOUZA, P. de T. F. de. Direito e democracia - o significado das leis e do legislativo na teoria da democracia. *Revista Do Direito*, (51), 145-156, 2017. Disponível em: https://doi.org/10.17058/rdunisc.v1i51.7784. Acesso em: 6 set. 2020.

TIERSMA, P. *Legal Language*. Chicago: The University of Chicago Press, 1999.

A CRIAÇÃO DO MACHADO DE ASSIS CATÁLOGO & CORPUS (MACC)

Ursula Puello SYDIO[113]

RESUMO O MACC, ou Machado de Assis Catálogo & Corpus, é um recurso digital que permite que pesquisadores acessem um abrangente catálogo e um corpus bilíngue da obra machadiana traduzida para o inglês. Enquanto o seu Catálogo reúne 39 títulos, incluindo romances, contos e antologias do autor, o seu Corpus conta com 6 romances (11 traduções) e 90 contos (373 traduções).

Palavras-chave Linguística de corpus; Machado de Assis; Estudos da tradução; catálogo; Literatura brasileira traduzida.

INTRODUÇÃO

Machado de Assis é o autor brasileiro mais pesquisado no exterior de acordo com Costa (2016). No entanto, ainda não existia de um catálogo completo e atualizado das traduções para o inglês, tampouco um corpus eletrônico que facilitasse a consulta a essas traduções. Com essa lacuna em mente, foi desenvolvido o website MACC (SYDIO, 2023a), o Machado de Assis Catálogo & Corpus, que está disponível no endereço https://macc.fflch.usp.br/pt-br/.

O Catálogo do MACC lista 39 publicações traduzidas no período de 1921 a 2023, incluindo romances, contos e antologias da obra machadiana. Já o Corpus traz 6 romances (11 traduções) e 90 contos (373 traduções).

Este trabalho tem como objetivo apresentar os métodos utilizados para a criação do catálogo e do corpus, além de explicar como os pesquisadores podem usufruir desse recurso digital.

FUNDAMENTAÇÃO TEÓRICA

O Catálogo surgiu antes do Corpus, afinal, antes de iniciar a compilação do corpus, era necessário estabelecer quantas e quais obras do Machado de Assis foram traduzidas. Embora a obra do autor em português esteja amplamente catalogada, a mesma situação não se aplica às traduções, que são mais dinâmicas e sujeitas a frequentes retraduções.

A obra do Machado de Assis chegou em terras anglófonas em 1921, com a tradução de três de seus contos na coletânea *Brazilian Tales* (ASSIS, 1921). No entanto, os seus romances só chegaram às prateleiras das livrarias dos Estados Unidos e Reino Unido na década de 1950, desde então, os seus títulos continuaram a ser traduzidos e retraduzidos. Por exemplo, o autor recebeu duas

[113] Doutoranda em Letras Estrangeiras e Tradução (LETRA), Universidade de São Paulo, São Paulo - SP, bolsista CAPES, contato através do e-mail ursulapuellosydio@gmail.com.

retraduções de *Memórias Póstumas de Brás Cubas* (1881) em 2020 e outra retradução de *Dom Casmurro* (1889) em 2023. As recentes retraduções demonstram a importância de um catálogo atualizado, ainda que tenhamos a consciência da impossibilidade de atingir à completude (Pym, 2014). Apesar de sua incompletude, Pym (2014) também enfatiza que os catálogos são fundamentais para a criação de corpora.

O próximo passo foi a construção do Corpus. Como define Tagnin (2015, p.1), os "corpora são bancos de textos de linguagem autêntica, criteriosamente construídos, destinados à pesquisa e legíveis por computador". Além disso, os corpora eletrônicos são o objeto de estudo da Linguística de corpus, que é "uma abordagem empírica para o estudo da língua [...] especialmente útil no estudo da Tradução" (Tagnin, 2015, p.1), pois as ferramentas computacionais permitem a análise de corpora mais volumosos.

A decisão de disponibilizar o Catálogo e o Corpus em um recurso digital voltado para pesquisadores foi tomada com base em projetos similares que precederam o MACC. Entre os websites com catálogos de tradução ou corpora eletrônicos que serviram de inspiração para o MACC, estão o CorTrad do projeto CoMET (TAGNIN; TEIXEIRA; SANTOS, 2009), o COMPARA, da Linguateca (FRANKENBERG-GARCIA; SANTOS, 2002) e o website Poesia Traduzida no Brasil (ASEFF, 2018).

METODOLOGIA

A pesquisa iniciou pela criação do catálogo, ou seja, o primeiro passo foi o levantamento das traduções de obras machadianas para língua inglesa, uma vez que os títulos em português já estavam bem catalogados. Os títulos foram levantados a partir de artigos, teses, catálogos de tradução online e, principalmente, de busca por obras do autor em grandes sites de comércio de livros. Cada título identificado foi registrado em um banco de dados que reunia informações como título em inglês, título em português, gênero literário, tradutor(a), ano de publicação, editora e país. Essa base de dados permitiu traçar um panorama histórico das traduções de Machado de Assis para o inglês, desde 1921 até os dias atuais. Em seguida, o banco de dados foi dividido por gênero literário. Uma tabela era dedicada aos romances, como estes geralmente são publicados fora de antologias, foi fácil estabelecer a obra correspondente em português. Por outro lado, na tabela dedicada aos contos, o processo foi um pouco mais longo, uma vez que a maioria dos contos está reunida em coletâneas e foi necessário investigar conto a conto para concluir a fase de catalogação.

Com as obras catalogadas, a próxima etapa foi a compilação do corpus e ela foi dividida nas seguintes fases:

a) Conversão dos textos dos livros eletrônicos em arquivos .txt;
b) Pré-processamento, limpeza e organização dos textos;
c) 1ª parte do alinhamento: foi criado um programa computacional em Python para extrair cada .txt e converter um único arquivo .csv por obra, com o texto em português na primeira coluna e as traduções nas colunas seguintes;

d) 2ª parte do alinhamento: verificação e correção manual (obra por obra) do alinhamento por parágrafos dos arquivos .csv;
e) Criação de um banco de dados em PostgreSQL, que seria usado para buscas no website do MACC.

Por fim, com o corpus eletrônico, literário (contos e romances machadianos), bilíngue (português e inglês) e paralelo (alinhado por parágrafos) pronto, foi possível dar início a última etapa da pesquisa: a construção de website gratuito e fácil de navegar que contaria com as ferramentas para buscas tanto no Corpus quanto no Catálogo, em palavras, a criação do MACC.

DISCUSSÃO DOS DADOS

O MACC traz os dados mais relevantes levantados durante a pesquisa de forma estruturada, pesquisável e acessível para a comunidade acadêmica em um website gratuito, basta realizar um cadastro prévio.

A seção do Catálogo pode ser consultada de três formas. Na primeira, temos “Linha do tempo” que lista os 39 títulos catalogados em inglês em ordem cronológica. Na segunda, temos a “Busca por obras em língua portuguesa”, que permite o visitante pesquisar a partir do título em português ou do gênero literário. Na terceira, temos a “Busca por obras em língua inglesa”, que permite o visitante filtrar a sua busca a partir de informações como ano de publicação, título em inglês, título em português, gênero literário e país.

A seção de “Busca no Corpus” traz um corpus machadiano que totaliza 2.105.695 palavras, divididas em um subcorpus em português, composto por 6 romances e 90 contos, e outro em inglês, composto por 11 traduções dos romances e 373 traduções dos contos. Através da ferramenta de busca desenvolvida especialmente para o MACC, o visitante pode pesquisar a partir do subcorpus em inglês ou em português.

Em suma, o MACC é fruto de uma pesquisa de mestrado (SYDIO, 2023b), reunindo um catálogo atualizado de traduções da obra machadiana para inglês e o maior corpus paralelo bilíngue da obra machadiana disponível para consulta. A sua principal contribuição para os Estudos da Tradução, os Estudos Machadianos e para a Linguística de Corpus é a disponibilização desses para a comunidade científica.

Agradecimentos

Agradeço à Profa. Dra. Luciana Carvalho Fonseca, por sua orientação durante o mestrado, à Profa. Dra. Stella Esther Ortweiler Tagnin, por sua contribuição durante a qualificação e por me orientar agora no doutorado, e à Profa. Dra. Marlova Gonsales Aseff e à Profa. Dra. Elisa Duarte Teixeira pelas contribuições enquanto a banca examinadora do mestrado.

REFERÊNCIAS

ASEFF, Marlova. Catálogo da poesia traduzida no Brasil (1960-2009). 1. Ed. Brasília, 2018. ISBN: 978-85-540456-0-9. Disponível em: http://poesiatraduzida.com.br/ Acessado em: 10 jul. 2024.

ASSIS, Machado de. **Brazilian Tales.** Trad. Isaac Goldberg. Boston: The Four Seas Company, 1921.

ASSIS, Machado de. **Obra Completa de Machado de Assis**. Rio de Janeiro: Nova Aguilar, 1994. Disponível em: http://machado.mec.gov.br/ Acessado em: 10 jul. 2024.

COSTA, C. B. **DOM CASMURRO EM INGLÊS: TRADUÇÃO E RECEPÇÃO DE UM CLÁSSICO BRASILEIRO**. [s.l.] Universidade Federal de Santa Catarina, 2016.

FRANKENBERG-GARCIA, A.; SANTOS, D. COMPARA, um corpus paralelo de português e de inglês na Web. **Cadernos de Tradução IX**.1 Florianópolis, 2002, pp. 61-79.

PYM, A. **Method in Translation History.** Routledge, 2014.

SYDIO, Ursula Puello. MACC: Machado de Assis Catálogo & Corpus. [S. l.], 1 jan. 2023a.

SYDIO, Ursula Puello. **Machado de Assis Catálogo & Corpus (MACC):** A construção de um catálogo e um corpus paralelo das traduções da obra machadiana para língua inglesa. 126 f. Dissertação (Mestrado) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2023b.

TAGNIN, S. E. O. A Linguística de Corpus na e para a Tradução. Em: **Corpora na Tradução**. São Paulo: HUB, 2015.

TAGNIN, S. E. O.; TEIXEIRA, E. D.; SANTOS, D. CorTrad: a multiversion translation corpus for the Portuguese-English pair. **Arena Romanistica**, v. 4, p. 314-323, 2009.

ANOTAÇÃO SEMÂNTICA MULTIMODAL A PARTIR DO CORPUS AUDITION: UMA CONTRIBUIÇÃO DA SEMÂNTICA DE FRAMES PARA A PESQUISA EM TRADUÇÃO AUDIOVISUAL ACESSÍVEL

Maucha Andrade GAMONAL[114]
Adriana Silvina PAGANO[2]
Tiago Timponi TORRENT[115]

ABSTRACT: This work presents the multimodal semantic annotation conducted through the Audition corpus. The corpus consists of a series of short films from different genres and includes semantic annotation of the original audio transcription, subtitles and closed captions, overlay text, and audio description. The annotation methodology follows the approach of Belcavello et al. (2024), and the theoretical framework supporting the research is Frame Semantics (Fillmore, 1982).

Palavras-chave: Semântica de Frames; Tradução Audiovisual Acessível; audiodescrição; anotação semântica; corpus multimodal.

Introdução

O estudo e o desenvolvimento de metodologias e tecnologias que relacionem conteúdo audiovisual a práticas de acessibilidade é um dever da comunidade científica. Em conformidade com o paradigma dos direitos humanos, a inclusão é um direito social, como ratificado em 2008 pela Convenção sobre os Direitos das Pessoas com Deficiência (Brasil, 2008). Nesse sentido, aliar a Semântica de *Frames* pode ser útil para a Tradução Audiovisual Acessível, especialmente para a prática da audiodescrição, uma vez que possibilita avaliar os efeitos das escolhas lexicais em uma audiodescrição a partir de uma abordagem teórico-metodológica de *frames* semânticos.

Neste trabalho, apresentamos a prática de anotação semântica multimodal no corpus Audition, um corpus compilado com diversos objetos modais, incluindo material audiovisual em formato curta-metragem, transcrição das audiodescrições de cada curta, além de legendas e outras informações visuais. Selecionamos anotações de diversos curtas, dentre eles aquelas produzidas para o curta-metragem Eu não Quero Voltar Sozinho tanto da transcrição da audiodescrição quanto das imagens do curta em que há tais inserções.

---

[114] Residente de pós-doutoramento no Programa de Pós-Graduação em Linguística na Universidade Federal de Minas Gerais, Minas Gerais, atualmente é bolsista Capes. mgamonal@ufmg.br. [2] Professora Titular de Linguística Aplicada da Universidade Federal de Minas Gerais, Minas Gerais, bolsista de produtividade em Pesquisa IC do CNPq.

[115] Professor Associado do Departamento de Letras e do Programa de Pós-Graduação em Linguística da Universidade Federal de Juiz de Fora, Minas Gerais, bolsista de Produtividade em Pesquisa 2, CNPq.

Além da audiodescrição oficial produzida pelo grupo Tramad, cuja anotação foi realizada por Dornelas (2023), nós incluímos a anotação de outra versão de audiodescrição produzida para o mesmo material audiovisual (Vieira, 2015). A metodologia do trabalho segue os procedimentos de anotação de *frames* semânticos da FrameNet Brasil (Torrent et al., 2022) voltadas à prática multimodal (Belcavello et al., 2024).

Os dados de anotação deste trabalho compõem o dataset semântico de objetos multimodais da FrameNet Brasil. Tais dados serão úteis para a aplicação tecnologia linguística em Processamento Automático de Linguagem Natural, como algoritmos de inteligência artificial para rotulação semântica automática.

Fundamentação teórica *Frames* semânticos

O conceito de *frame* na Semântica advém de Fillmore (1985) por meio da longa trajetória de estudos que se consolidou com a proposição da teoria Semântica de *Frames*. Sua proposta se distancia dos estudos formais e se aproxima dos estudos empíricos, uma vez que, como ele mesmo destaca, o interesse está em investigar as continuidades entre a linguagem e a experiência (Fillmore, 1982, p.112). Para ele, as escolhas linguísticas de uma comunidade fala revelam categorias de experiência codificadas por seus membros.

O autor utiliza a analogia da gramática e um conjunto de ferramentas. Assim como as ferramentas são identificadas por suas especificidades de forma e composição, assim é a fonologia e a morfologia de uma língua. E, semelhantemente à gramática, as ferramentas servem a diversos propósitos tendo em vista a grande quantidade de situações para as quais são úteis. Desse modo, Fillmore convida a pensar o texto não como um registro de "pequenos significados" em busca de um "significado maior", mas como um registro de ferramentas utilizadas para uma determinada atividade (Fillmore, 1982, p.113). Ao *frame* cabe a função de representar tal sistema de conceitos.

> frame é um sistema de conceitos relacionados de modo que, para entender qualquer um deles, é necessário entender toda a estrutura de conceitos na qual se enquadram, quando um dos elementos é introduzido em um textoou em uma conversa, todos os outros serão disponibilizados automaticamente[116]. (Fillmore, 1982, p.111)

As unidades lexicais *construção*.n, *montar*.v, *reformar*.v, *erguer*.v, *reforma*.n se reúnem no *frame* `Construir`[117], de acordo com o repositório de *frames* da FrameNet Brasil. Em sua definição, há ações de montagem ou de construção, em que o AGENTE [118] une um COMPONENTE para formar a ENTIDADE_CRIADA (FrameNet Brasil, 2024). Sabe-se que o *frame* é definido a partir de seus elementos, os chamados Elementos de *Frame* (FE), e esses

[116] A frame is any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available.
Tradução dos autores.

[117] Seguindo convenções estabelecidas, *frames* são registrados em fonte Courier.

[118] Seguindo convenções estabelecidas, elementos de *frames* são registrados em letras maiúsculas.

elementos, por suas vezes, são agrupados por meio da afinidade semântica atribuída ao conjunto de unidades lexicais (LU) que instanciam o *frame*.

1) Até que [alguém AGENTE] construiu CONSTRUIR [um barzinho ENTIDADE_CRIADA], e, pouco a pouco, se tornou um vilarejo completo. [#181606]
2) Em outras partes da cidade, quando queriam reformar [um bairro ENTIDADE_CRIADA], eles demoliam tudo e construíam prédios bem feios no lugar. [#181367]

As sentenças (1) e (2) integram o conjunto de anotações de corpus da FrameNet Brasil. Diz-se que tanto *construir*.v como *reformar*.v são LUs que evocam o *frame* `Construir` nos exemplos ilustrados. Por mais que cada uma tenha suas especificidades semânticas, elas dispõem de afinidades que as conectam a uma categoria da experiência. E isso é visto por intermédio de seus FEs.

Audiodescrição

A AD é uma forma de tradução audiovisual acessível da informação visual em linguagem auditiva verbal (Pagano et al., 2016). Conforme Fryer (2016) destaca, consiste em um comentário verbal que fornece informações visuais para aqueles que não as percebem por conta própria.

A Tradução Audiovisual Acessível (TAV), segundo Fryer (2016, p.2), refere-se à tradução de todos os produtos audiovisuais, incluindo filmes, documentários, programas de televisão e conteúdo online. Como a autora menciona, ao contrário das legendas, dublagens ou *voice-over*, a audiodescrição não se desenvolve a partir de um texto verbal preexistente. Tal realidade, para alguns autores, caracteriza a AD como uma mediação intersemiótica, intermodal ou cross-modal (Jiménez Hurtado, 2007, apud Fryer, 2016).

No Brasil, conforme Franco e Araújo (2022), a efetividade do acesso à audiodescrição e a outras práticas de acessibilidade estão diretamente ligadas ao cenário político do país. Elas destacaram a agenda política do ano vigente da publicação da obra, em que investimentos em projetos culturais foram extintos, além do próprio Ministério da Cultura. Mesmo assim, as autoras enfatizam a resiliência de todos os agentes de cultura e acessibilidade em garantir inclusão por meio da TAV.

Metodologia

Corpus Audition

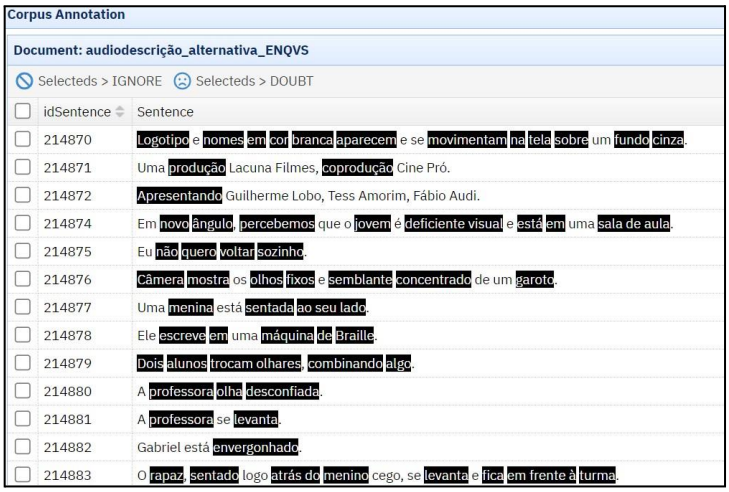
O curta-metragem Eu não quero voltar sozinho é uma obra audiovisual produzida em 2010 pela Lacuna Filmes, cuja audiodescrição foi elaborada pelo grupo Tradução, Mídia e Audiodescrição (TRAMAD). O curta é uma das obras audiovisuais que compõem o Audition, corpus compilado para as tarefas de anotação multimodal da FrameNet Brasil (Silva et al., 2023).

No Audition, os curtas-metragens são de diferentes gêneros com documentários, animações e *live actions*. Em comum, todos possuem a opção de audiodescrição. Para este trabalho, nosso material de análise inclui a audiodescrição oficial do curta e a anotação multimodal de outra versão produzida por Vieira (2015).

Frames semânticos em anotação multimodal

Os procedimentos de anotação seguem conforme a metodologia da FrameNet Brasil (Torrent el al., 2022; Belcavelo et al., 2024). Na Figura 1, as sentenças transcritas da audiodescrição são mostradas no software de anotação webtool. As marcações destacam as Unidades Lexicais que evocam algum *frame* no banco de dados disponível.

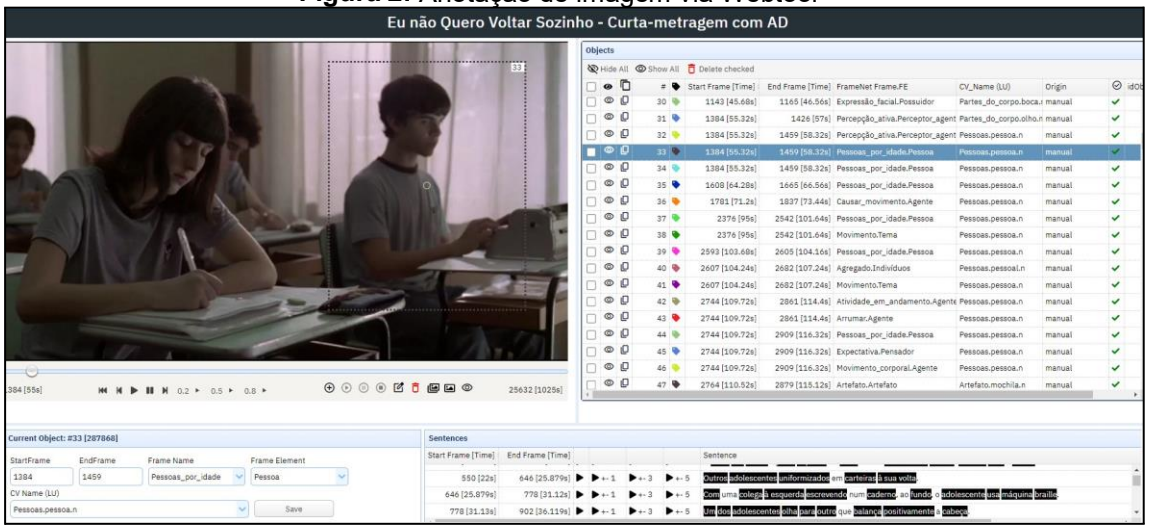
Figura 1: Sentenças transcritas para anotação linguística via Webtool



Fonte: Captura de tela da Webtool (FrameNet Brasil, 2024)

A Figura 2 exibe a tela de anotação de imagem também via Webtool. Por meio dela, o anotador identifica os objetos visuais e os relaciona aos *frames* do banco por meio dos elementos que atuam no *frame*, os chamados Elementos de *Frame*.

Figura 2: Anotação de imagem via Webtool



Fonte: Captura de tela da Webtool (FrameNet Brasil, 2024)

Discussão dos dados

Os resultados da anotação multimodal das duas transcrições das audiodescrições geraram um agrupamento de *frames* semânticos, Elementos de *Frame* e Unidades Lexicais. A partir do produto das anotações, verifica-se a construção de sentido nesses dois materiais audiodescritos, identificando similaridades e particularidades acerca de cada opção tradutória.

A percepção do audiodescritor acerca da obra audiovisual mostra a centralidade da perspectiva adotada na construção de uma AD. A Semântica de *Frames* e a FrameNet Brasil são apresentadas neste trabalho como aparatos teórico-metodológicos tanto para análise quanto para criação de conteúdo audiovisual acessível com o interesse de ser útil para a criação de recursos que possibilitem inclusão e acessibilidade no âmbito audiovisual.

Agradecimentos

Este trabalho recebeu suporte do CNPq por meio dos processos 151361/2023-1 and 313103/2021-6, e da CAPES por meios dos processos 88887.936139/2024-00 and 8887.683333/2022-00. Os autores expressam agradecimento a todos os estudantes e demais pesquisadores que atuaram na criação do corpus Audition.

Referências bibliográficas

BRASIL. Decreto Legislativo nº 186, de 9 de julho de 2008. Aprova o texto da Convenção sobre os Direitos das Pessoas com Deficiência e seu Protocolo Facultativo, assinados em Nova York, em 30 de março de 2007. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/decreto/d186.htm. Acesso em: 15 jul. 2024.

BELCAVELLO, F.; VIRIDIANO, M.; MATOS, E.; TORRENT, T. T. Charon: A FrameNet Annotation Tool for Multimodal Corpora *In*: Proceedings of The 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022. Marseille, France: ELRA, 2022, p. 91-96.

DORNELAS, L. A audiodescrição sob a perspectiva da semântica de frames: análise dos frames evocados pelo texto da audiodescrição e pelas imagens dinâmicas num curta-metragem. 2023. Dissertação de mestrado - Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

FILLMORE, C. J. Frame Semantics. *In* The Linguistic Society of Korea (org.). Linguistics in the Morning Calm. Seoul: Hanshin, 1982, p. 111-137.

_____. Frames and the semantics of understanding. *In*.: Quaderni di Semantica. Vol. VI, nº 2, Dezembro de 1985.

FRANCO, E. P. C. ; ARAÚJO, V. L. S. . Audio Description in Brazil. In: Taylor, Christopher; Perego, Elisa. (Org.). The Routledge Handbook of Audio Description. 1ed.Londres: Routledge, 2022, v. 1, p. 596-612.

FRYER, L. (ed.). (2016). An Introduction to Audio Description a Practical Guide. London: Routledge

FRAMENET BRASIL. Software de anotação Webtool. Disponível em: https://webtool.framenetbr.ufjf.br/. Acesso em:15 de jul. 2024.

PAGANO, A. S.; TEIXEIRA, A. L. R.; MAYER, F. A. Accessible Audiovisual

Translation. In: JI, Meng; LAVIOSA, Sara (ed.). *The Oxford Handbook of Translation and Social Practices*. Oxford: Oxford University Press, 2020. cap. 4, p. 66-82.

TORRENT, T. T.; MATOS, E. E.; BELCAVELLO, F.; VIRIDIANO, M.; COSTA, A. D.; GAMONAL, M. A.; MARIM, M. C. Representing context in FrameNet: *a multidimensional, multimodal approach*. Frontiers in Psychology - Language Sciences, 13, article 838441.

SILVA, A. C.; RABELO, I.; OLIVEIRA, I. M.; SOUZA, M.; GAMONAL, M.; ROZA, R. Coleta, composição e etapas de pré-processamento de corpus: procedimentos para a anotação multimodal da FrameNet Brasil. *In*: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 14. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 362-366. VIEIRA, G. M. Audiodescrição do curta Eu não Quero Voltar Sozinho. *In*: Trabalho acadêmico para Núcleo Temático sobre Cinema e Representações Sociais.Universidade Federal do Vale do São Francisco (UNIVASP), 2015.

O PROCESSAMENTO DA LINGUAGEM NATURAL NO ÂMBITO DA PROMOÇÃO DA ACESSIBILIDADE TEXTUAL E TERMINOLÓGICA

Heloísa Orsi Koch DELGADO[119]
Bruna Rodrigues da SILVA[120]

RESUMO: Este recorte de pesquisa apresenta análise de texto sobre Transtorno de Humor Bipolar, sob viés do Processamento da Linguagem Natural. O objetivo foi verificar se o trecho seria compreendido pelo leitor médio brasileiro. Como o texto se mostrou complexo, foram adotados preceitos de tradução intralinguística, com vistas a promover sua acessibilidade textual e terminológica.

Palavras-chave: Acessibilidade Textual e Terminológica; linguagem simples; Transtorno do Humor Bipolar; NILC-Metrix; escolaridade limitada.

INTRODUÇÃO

Este trabalho apresenta recorte de projeto de pesquisa realizado no estágio pós-doutoral da primeira autora e inserido no âmbito do Grupo de Pesquisa Acessibilidade Textual e Terminológica (GEATT) da Universidade Federal do Rio

Grande do Sul (UFRGS). A pesquisa se insere nas áreas de Terminologia, Tradução Intralinguística e Acessibilidade Textual e Terminológica (ATT), tendo como enfoque metodológico o Processamento da Linguagem Natural (PLN).

A partir desse direcionamento teórico-prático, analisamos os elementos textuais e terminológicos — de possível complexidade — de textos sobre o Transtorno do Humor Bipolar (THB)[121], escritos em português do Brasil. Esses trechos fazem parte do dicionário on-line sobre esse transtorno (DicTrans), voltado para estudantes de Medicina e profissionais da saúde. A análise de dificuldade textual teve como referência o perfil de leitor adulto com baixo nível de instrução formal (Ensino Fundamental Completo) e pouca experiência de leitura.

Tal investigação, de natureza qualiquantitativa, baseou-se nos índices de complexidade da linguagem, obtidos por meio de ferramentas linguístico-computacionais. Dessa forma, verificamos se o *corpus* utilizado está adequado ao perfil de leitor desejado ou se mudanças textuais precisariam ser realizadas para elaborar um texto mais acessível. Avaliamos quais trechos deveriam ser reformulados e quais índices seriam apontados pelas ferramentas. Salientamos que, tanto os traços globais, quanto os particulares dos textos foram analisados em um todo de significação e de comunicação, possibilitando que nosso *corpus*

[119] Linguista, tradutora, professora universitária. Universidade La Salle, Canoas, Rio Grande do Sul. heloisa.orsi.koch.delgado@gmail.com.br.

[120] Doutoranda pelo PPG-Letras/UFRGS, professora da rede pública de ensino, Porto Alegre, Rio Grande do Sul.

[121] Pode ser também chamado de Transtorno Afetivo Bipolar (TAB).

servisse para ilustrar dados comparativos e ilustrativos sobre os níveis de complexidade textual nele encontrados.

A escolha pelo THB justifica-se pela condição grave que costuma ser acometida por episódios maníacos e depressivos, afetando 140 milhões de pessoas no mundo, conforme resultados de 2019 da OMS. No Brasil, a Associação Brasileira de Transtorno Bipolar (ABTB) estima que cerca de 8% da população adulta brasileira sofre do quadro psicopatológico (AMARO, 2020). Dados do Ministério da Saúde mostram que foram registrados 4.839.833 procedimentos no Sistema de Informação Ambulatorial (SIA/SUS) entre os meses de março e maio de 2019 (FLORES, s.d.).

FUNDAMENTAÇÃO TEÓRICA

Os textos escritos são os campos naturais de termos de áreas específicas do saber. Krieger & Finatto (2004, p.106) já afirmavam que

> a relevância do texto está diretamente vinculada ao princípio comunicacional que postulam. Isso corresponde a considerar o texto como *habitat* natural das terminologias dotados de elementos linguísticos, pragmáticos, comunicativos e discursivos, e como objeto de comunicação entre destinador e destinatário.

Nesse sentido, Ciapuscio (2003) observa que o uso de determinadas terminologias varia de acordo com o nível para o qual os textos serão destinados: no caso de textos escritos por ou para especialistas, o conceito de um termo é pleno, enquanto para o público geral e leigo, apenas os traços que são relevantes para a caracterização dos termos permanecem. As considerações dessas autoras guiaram a reformulação dos trechos aqui apresentados, buscando a acessibilidade textual e terminológica.

Quanto à reescrita dos trechos originais, fizemos uso das estratégias da Tradução Intralinguística - cuja função é, primordialmente, explicar os signos verbais de uma língua utilizando outros signos da mesma língua - visto que almeja subsidiar uma crescente demanda pela compreensão de discursos técnicos e científicos (ZETHSEN, 2009; JAKOBSON, 1959). Ao descrever a tradução intralinguística e apontar micro estratégias para serem utilizadas, Zethsen (2009, p.16) destaca que, “entre os achados mais importantes, há uma forte tendência à simplificação”.

Com relação ao enfoque aplicado, utilizamos os princípios do PLN, marcado pela observação de índices em textos individuais, de forma que programas computacionais, tais como o NILC-Metrix[122], analisem um texto e disponham de uma quantidade determinada de informações. Portanto, a possibilidade de um tratamento individualizado de cada texto é válida e necessária para pesquisas que tenham como objetivo examinar textos, um a um, para compará-los, de maneira multifatorial, com outros (FINATTO, 2018). Pode-se definir o PLN, então, como uma vertente da inteligência artificial, que incorpora inúmeras técnicas para interpretação da linguagem com base em métodos estatísticos e de aprendizado de um determinado número de regras de

[122] Disponível em http://fw.nilc.icmc.usp.br:23380/nilcmetrix. Acesso em set. 2024.

funcionamento de uma língua por meio de análises de *corpora* de exemplos típicos do mundo real (MANNING & SCHÜTZE, 1999).

METODOLOGIA

Em linhas gerais, os passos metodológicos compreenderam a observação de índices numéricos do *corpus* de estudo, submetido ao NILC-Metrix, de acordo com os preceitos do PLN.

As etapas realizadas foram: *i)* escolha dos excertos textuais disponíveis no DicTrans e envio à ferramenta; *ii)* análise dos índices obtidos sob uma perspectiva quantitativa; *iii)* averiguação dos trechos de potencial complexidade, considerando o perfil do leitor (olhar qualitativo); *iv)* aplicação da tradução intralinguística; *v)* envio da nova versão à ferramenta; e *vi)* análise dos novos resultados. Também foi realizada a repetição dos passos anteriores caso o texto ainda não se mostrasse adequado ao propósito de pesquisa. Ao final, foi realizada a comparação dos índices obtidos entre os textos originais e simplificados.

As métricas de interesse, além da frequência de palavras, são: **simplicidade textual** (reflete a quantidade de palavras em cada sentença do texto, bem como usos mais simples e estruturas sintáticas mais familiares, que apresentem menor desafio para a compreensão) e **simplicidade lexical** (verifica se o texto contém palavras que possuam significado complexo ou evoquem imagens mentais fáceis de processar e de entender). A ferramenta utilizada agrupa métricas desenvolvidas em mais de uma década no NILC, iniciadas com o Coh-Metrix-Port. Trata-se de um sistema computacional que contém por volta de 200 métricas propostas em estudos de discurso, psicolinguística, linguística cognitiva e computacional, que tem o objetivo de analisar a complexidade textual para o português (WICK-PEDRO E SANTOS, 2021). As métricas analisadas aqui são:

Tabela 1 – Métricas do NILC-Metrix

Nome da métrica	Interpretação	Descrição
Índice de Leiturabilidade Flesch **(1)**	Quanto maior o resultado da métrica, menor a complexidade textual.	Busca uma correlação entre tamanhos médios de palavras e sentenças.
Média dos valores das frequências das palavras de conteúdo do texto via Corpus Brasileiro **(2)**	Quanto maior a frequência das palavras, menor a complexidade do texto.	Média dos valores das frequências das palavras de conteúdo do texto, variando entre 1 e 7.
Proporção de sentenças longas em relação a todas as sentenças do texto **(3)**	As longas são mais complexas do que as sentenças curtas e médias; as muito longas, mais complexas do que as longas.	Proporção de sentenças longas em relação a todas as sentenças do texto.
Proporção de sentenças curtas em relação a todas as sentenças	Quanto maior a proporção de sentenças curtas, menos	Proporção de sentenças curtas em relação a todas as sentenças do

do texto **(4)**	complexo é o texto.	texto.

Fonte: NILC-Metrix (2024)

RESULTADOS E CONSIDERAÇÕES

Os resultados apontaram para a complexidade dos textos originais. Sendo assim, de fato, foi preciso fazer uso da tradução intralinguística para que os textos se tornassem potencialmente compreensíveis ao público.

Quadro 1 – Exemplo de trecho sobre episódio maníaco

Episódio maníaco
Reconhece-se um episódio maníaco quando as seguintes evidências comportamentais se manifestam. Período distinto de humor anormal e persistentemente elevado, expansivo ou irritável e aumento anormal e contínuo da energia ou da atividade dirigida a objetivos ou com duração mínima de uma semana e presente na maior parte do dia, quase todos os dias (ou qualquer duração se a hospitalização se fizer necessária). Durante o período de perturbação do humor e aumento da energia ou atividade, três (ou mais) dos seguintes sintomas (quatro dias se o humor é apenas irritável) estiverem presentes em grau significativo e representem uma mudança notável do comportamento habitual: autoestima inflada ou grandiosidade; redução da necessidade de sono; mais loquaz que o habitual; fuga de ideias ou experiência subjetiva de que os pensamentos estão acelerados; distratibilidade, ou seja, a atenção é desviada muito facilmente por estímulos externos insignificantes ou irrelevantes; aumento da atividade social, profissional, escolar ou sexual e agitação psicomotora; envolvimento excessivo em atividades com elevado potencial para consequências dolorosas (surtos desenfreados de compras, indiscrições sexuais ou investimentos financeiros insensatos).

Fonte: Dictrans (2012)

O índice Flesch (1) corresponde ao índice -23.1875, que extrapola o índice mais alto de complexidade, indicando um texto consideravelmente difícil para pessoas com grau de alfabetização limitado e condizente com a de um especialista da área. A métrica (2) indica um valor de 4.48262, apontando para uma maior frequência de palavras de conteúdo (substantivos, verbos, adjetivos e advérbios), ou seja, teoricamente, menor a complexidade do texto. Para a métrica (3), o seguinte cálculo foi feito: o texto registra 3 sentenças de 11, 51 e 112 palavras (1 frase curta e 2 consideravelmente longas). A proporção é 2/3, resultando no índice de 0,66 e indicando que o texto é complexo (as diretrizes da acessibilidade textual postulam que as frases devem conter, no máximo, 25 palavras). A última métrica (4) aponta para o valor de 0,5 feita com base no cálculo da métrica (3), isto é, 1/3. Aqui, evidenciamos um índice um pouco maior do que 0,33 (resultado esperado), o qual pode significar que a diferença (0,17) esteja associada ao excedente de palavras nas frases muito longas (principalmente a de 112 palavras) em comparação à curta. Vejamos, abaixo, a tradução intralinguística com base no texto original.

Quadro 2 – Tradução intralinguística sobre o episódio maníaco

Episódio maníaco
O episódio maníaco acontece quando a pessoa tem o humor diferente do normal. Fica muito alegre e feliz ou irritada e com muita energia por muitas horas durante o dia e por, pelo menos, uma semana. Quando a pessoa fica assim, ela pode se achar melhor do que os outros, dormir menos, falar muito e mudar de um assunto para o outro rapidamente e estar desatenta. Pode gastar demais e conversar sobre sexo sem sentir-se envergonhada. A pessoa pode ficar fora de si e ter que parar de trabalhar, sair com amigos e namorar. Pode ter que ir para o hospital para ficar mais segura e não fazer mal a si e nem aos outros.

Fonte: As autoras.

O conteúdo do texto original consiste em levar o conhecimento sobre o episódio maníaco a pessoas inseridas no mundo médico ou da saúde em geral. Assim, é aceitável que termos da área estejam presentes sem que seja necessário aplicar subsídios da tradução intralinguística para deixar o texto mais fácil de entender.

Porém, quando se trata de textos para uma fatia significativa da população – que possui um baixo nível de instrução e letramento -, estratégias linguísticas que favoreçam à compreensão precisam ser levadas em consideração. Assim, a primeira ação tomada foi a de enxugar o texto. Em seguida, optamos pela paráfrase, ou seja, renúncia a componentes formais ou funcionais do texto que possam levar à falta de acesso ao conhecimento. Essa estratégia resultou em um número significativamente menor de frases e palavras, em que se evidenciam ponderações sobre o todo do texto. Nota-se, portanto, que facilitar um texto extrapola os limites da simples troca de palavras, sugere-se "uma dada escrita por uma alternativa entre várias possíveis" (FINATTO; TCACENCO, 2021).

O índice Flesch (1) foi alterado para 68.45462 (o 100 indica que o texto é muito fácil), revelando um grau de leiturabilidade fácil. A métrica (2) variou para 5.005 (maior do que no texto original), apontando para uma maior frequência de palavras de conteúdo, indicando um texto menos complexo. Para o cálculo da métrica (3), observamos que o texto registra 6 sentenças de 13, 23, 30, 10, 18 e 21 palavras, então, a proporção é 4/6 (quatro frases muito longas[123] para 6 frases no total), resultando no índice de 0,66 e indicando que o texto é complexo. No entanto, dois pontos merecem consideração: i) a tolerância de duas palavras explicadas na nota 9 e ii) as diretrizes da acessibilidade que sugerem o limite de 25 palavras aceitável e não de apenas 15 para frases longas e mais de 15 para muito longas. A última métrica (4) mudou para 0,28571, isto é, 1/6. Aqui, evidenciamos um índice menor do que 0,5, indicando a existência de mais frases longas do que curtas. Cabe, da mesma forma, considerar o exposto na métrica (3) e salientar que o uso de outras ferramentas poderia evidenciar resultados diferenciados (por conta de diferentes descrições e interpretações das métricas), oferecendo contrapontos relevantes. Dessa forma, seria possível afirmar que, para fins deste recorte de estudo, a reformulação simplificada apontou para pontos de reflexão dos resultados, mostrando que deve haver uma ponderação

[123] Como o texto não apresenta frases longas e a classificação entre as sentenças curtas, médias, longas e muito longas possui uma tolerância de apenas duas palavras, pode haver um desvio de classificação na contagem das frases em função dessa tolerância.

sobre os dados quantitativos. Estes devem contar com a reflexão qualitativa baseada no olhar e na experiência de tradutores intralinguísticos, direcionados para a simplificação textual, como forma de encontrar pontos de equilíbrio essas duas abordagens. Por fim, reforçamos a relevância da testagem entre leitores reais, para validar estratégias e identificar problemas ainda impensados.

REFERÊNCIAS

AMARO, D. **Entre a euforia e a depressão: 8% da população é bipolar**. Edição do Brasil. Disponível em http://edicaodobrasil.com.br/2020/03/20/entre-euforia-ehttp://edicaodobrasil.com.br/2020/03/20/entre-euforia-e-depressao-8-da-populacao-brasileira-e-bipolar/depressao-8-da-populacao-brasileira-e-bipolar/, 2020.

CIAPUSCIO, G. **Textos especializados y terminología**. Barcelona: IULA, 2003.

DICTRANS, Dicionário sobre o Transtorno do Humor Bipolar (DELGADO, H.O.K.; VERNETTI, C. L. ; SANTOS, C. A. dos). Porto Alegre, 2019. Disponível em:

https://www.dictrans.org/conheca.php. Acesso em: junho de 2023.

FINATTO, M. J. B. Corpus-amostra português do século XVIII: textos antigos de Medicina em atividades de ensino e pesquisa. DOMÍNIOS DE LINGU@GEM , v. 12, p. 435, 2018.

FINATTO, M. J. B.; TCACENCO, L. M. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. **Tradterm,** 37 (1), p. 30-63, 2021. Disponível em:https://www.revistas.usp.br/tradterm/article/view/168327. Acessado em: 02 jun. 2021. DOI https://doi.org/10.11606/issn.2317-9511.v37p30-63

FLORES, J. **Entre crises e likes**. Blog da UOL. Disponível em:

https://www.uol.com.br/vivabem/reportagens-especiais/transtorno-afetivo-bipolar-oque-e-por-que-ele-pode-piorar-com-a-pandemia. Acesso em: jan. de 2023.

JAKOBSON, R. **On linguistic aspects of Translation**. Massachusetts: Harvard University Press, 1959.

KRIEGER, M. G., FINATTO, M. J. B. **Introdução à Terminologia**: teoria & prática. 1a ed. São Paulo: Contexto, 2004.

WICK-PEDRO, G.; SANTOS, R. L. S. Complexidade textual em notícias satíricas: uma análise para o português do Brasil. *In*: **SIMPÓSIO BRASILEIRO**

DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL). Porto
Alegre: Sociedade Brasileira de Computação, 2021. p. 409-415.

MANNING, C.D., SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Massachussets: MIT Press, 1999.

ZETHSEN, K. K. (2009). **Intralingual Translation**: An Attempt at Description. Meta, 54 (4), 795–812. https://doi.org/10.7202/038904ar

CORPUSCRIPT: AN AUTOMATED TEXT-CLEANING TOOL FOR CORPUS LINGUISTICS

Jhonatan Henrique LOPES Alves[124]
Ana Eliza Pereira BOCORNY[125]
Deise Prina DUTRA[126]
Carolina Godoi de Faria MARQUES [127]
Gustavo Leal TEIXEIRA[128]
Danilo Duarte COSTA[129]

Introduction

The process of corpus compilation remains a significant challenge in the field of corpus linguistics. This paper introduces CorpuScript, an innovative text-cleaning software aimed at aiding researchers in the process of corpus preparation. By combining software engineering with corpus linguistics methods, this tool can significantly improve the workflow for corpora compilation, specifically in the task of corpus cleaning.

The necessity for CorpuScript emerged from recurring challenges experienced by our research team, particularly during our current corpus research project, in which a considerable large number of texts needed to be cleaned before being used for data analysis.

Considering the pressing need for an automated solution that could improve the text-cleaning process in our research project, CorpuScript was carefully developed to help us accelerate the corpus compilation, while meeting the requirements outlined in our corpus design.

THEORETICAL FRAMEWORK

The importance of clean, well-prepared corpora in linguistic research is wellestablished in the literature. Biber, Conrad, and Reppen (1998) emphasize that corpusbased investigations rely on empirical analysis of large, principled collections of natural texts.

Notably, a standard procedure in corpus building is the conversion of the selected texts into plain text (ASCII or UTF-8), since this type of file format can be run in most corpus analysis tools, as mentioned in Ädel (2020) and Reppen (2022).

---

[124] Undergraduate Student, Universidade Federal de Minas Gerais, Belo Horizonte, MG. Scholarship: FAPEMIG (APQ-01173-22)

[125] Professor, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[126] Professor, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

[127] Doctorate Student, Universidade Federal de Minas Gerais, Belo Horizonte, MG. Scholarhip: CAPES

(n. 88887.939578/2024-00)

[128] Professor, Universidade Federal de Minas Gerais, Montes Claros, MG

[129] Professor, Universidade Federal do Vale do Jequitinhonha e Mucuri, Diamantina, MG

However, Gries and Newman (2013, p. 263), point out that files will still “almost invariably require some editing for them to be used most effectively”. To ensure they can be read by computer software that are used for corpus analysis, Coxhead (2020, p. 470) stresses that text files “should be as clean as possible”.

In this context, corpus cleaning refers to the task of removing extraneous material from text data, such as headers, footers, special characters, line breaks, and other non-linguistic elements that do not contribute to the actual linguistic content (Weisser, 2016). Cleaning a corpus is a fundamental step in the corpus compilation process, since those unwanted items “may adversely affect the accuracy of the analytical procedures we intend to carry out, as well as impinging on the corpus’s representativeness” (McEnery and Brooks, 2022, p. 43).

Although cleaning many texts manually is rather time-consuming, it remains a frequently used method. For instance, when instructing English language students to use corpora for improving their learning, Poole (2018) suggests that they clean their texts by using the ‘find and replace tool’ in a text processor. Similarly, a non-automatic text cleaning approach was adopted in a study by Charles (2015) with students of English for Academic Purposes (EAP).

Automating the text cleaning process is certainly a highly welcomed advancement. To this end, according to Anthony (2020), high-level, functional programming languages are well-suited for the creation of brief, straightforward programs aimed at expediting the cleaning and processing of corpora. Languages such as Pearl, Python, and R have been extensively employed in corpus linguistics applications, encompassing tasks involved in corpus cleaning.

METHODS

CorpuScript was developed using Python, incorporating a suite of robust libraries to manage various aspects of text processing. The primary libraries utilized include:

1) Regular Expressions (“re” module): Fundamental for executing complex pattern matching and substitution operations essential for text cleaning.

2) SpaCy: A comprehensive natural language processing library employed for tasks such as tokenization, lemmatization, part-of-speech tagging, and stop word removal, thereby enhancing the linguistic accuracy of the text processing pipeline.

3) BeautifulSoup (bs4): Utilized for parsing and stripping HTML content from textual data, ensuring that only plain text is processed.

4) PySide6: Leveraged to develop the graphical user interface (GUI), facilitating user interaction and accessibility for researchers without programming expertise.

Additional Python Standard Libraries: Modules such as os, sys, unicodedata, logging, json, urllib.request, time, random, multiprocessing, and threading were integrated to handle file operations, system interactions, logging mechanisms, JSON data processing, network requests, time management, concurrency, and synchronization.

The core text cleaning functionality of CorpuScript is structured through a modular preprocessing pipeline, comprising the following key steps to standardize and prepare textual data for analysis:

HTML Stripping: Implemented via BeautifulSoup, this step removes any embedded HTML tags within the text, ensuring that only unformatted text is retained for subsequent processing.

Character Filtering: This module removes specified characters or sequences from the text based on user-defined parameters, allowing for the exclusion of unwanted symbols or tokens that may interfere with text analysis.

Diacritic Removal: Utilizing the unicodedata module, this process eliminates diacritical marks from characters, normalizing the text to its basic alphabetic form and enhancing consistency across different text inputs.

Script Filtering: Specific modules are employed to remove characters from nonLatin scripts, such as Greek and Cyrillic, maintaining a uniform character set within the corpus and eliminating potential noise from multilingual data.

Unicode Normalization: Applied using unicodedata.normalize with the NFKC (Normalization Form KC) standard, this step ensures that characters are represented in a consistent and compatible form, reducing discrepancies caused by varied Unicode encodings.

Whitespace Normalization: This process involves adjusting whitespace by removing unnecessary spaces preceding punctuation marks, standardizing spacing around punctuation, brackets, and braces, and collapsing multiple consecutive whitespace characters into a single space, thereby enhancing the readability and uniformity of the text.

Line Break Removal: By replacing newline characters with spaces, this module transforms multiline text into a continuous flow, which is beneficial for certain types of text analysis.

Bibliographical Reference Removal: Through the use of regular expressions, this step detects and removes bibliographical references embedded within the text, such as in-text citations, to focus the analysis on the main content.

Lowercasing: Converting all text to lowercase ensures uniformity, facilitating case-insensitive processing and comparison in subsequent analysis stages.

Lemmatization: Utilizing SpaCy’s lemmatization capabilities, this module reduces words to their base or dictionary forms, which aids in consolidating different morphological variants of a word, thus improving the semantic consistency of the corpus.

Tokenization: This process involves splitting the text into sentences or words using SpaCy's tokenization tools, enabling more granular analysis and manipulation of the textual data.

Stop Word Removal: SpaCy's predefined stop word list is employed to filter out common, non-informative words, thereby focusing the analysis on more meaningful and content-rich terms.

Unicode Category Filtering: This module removes characters belonging to specific Unicode categories, such as superscript and subscript characters, further refining the text and eliminating potential formatting artifacts.

Regular Expression Substitutions: Advanced pattern matching and replacements are conducted using user-defined regular expressions, allowing for customizable and flexible text cleaning operations tailored to specific dataset requirements.

The preprocessing pipeline is designed to be highly modular and configurable, enabling users to selectively apply cleaning steps based on their specific research needs. Each preprocessing module is implemented as a distinct component, facilitating ease of maintenance, scalability, and the ability to extend or modify the pipeline as needed. This modular architecture ensures that CorpuScript can accommodate a wide range of text processing tasks, from simple cleaning operations to more complex linguistic transformations, thereby supporting comprehensive corpus preparation for subsequent linguistic analysis.

Furthermore, CorpuScript's GUI, developed with PySide6, provides an intuitive interface for configuring processing parameters, selecting files or directories for processing, and monitoring progress through real-time feedback mechanisms. Concurrent processing capabilities, managed via Python's multiprocessing and threading modules, enable efficient handling of large datasets by leveraging multiple CPU cores. Logging functionalities ensure that all processing activities are meticulously recorded, facilitating debugging and audit trails.

RESULTS AND DISCUSSION

The implementation of CorpuScript has the potential to profoundly impact corpus compilation and research. The most striking advantage is the considerable reduction in time spent on pre-processing time.

A prime example of this efficiency gain was observed in our large-scale corpus research project. Initially, we estimated a six-month period for corpus preparation alone. However, with the introduction of CorpuScript midway through the project, we were able to complete the preparation phase in a matter of days, demonstrating the software's significant impact on research productivity.

The software's ability to maintain consistency across large volumes of text has also improved the quality of prepared corpora. By minimizing human error and ensuring uniform application of cleaning rules, CorpuScript can contribute to the reliability and validity of corpus-based studies.

CONCLUSION

CorpuScript represents a significant advancement in the field of corpus linguistics. By automating and streamlining the text cleaning process, it addresses long-standing challenges in corpus preparation, considerably reducing processing time, while minimizing human error and ensuring consistency across large corpora.

The software's impact extends beyond time-saving, enabling researchers to work with larger corpora and conduct more comprehensive analyses, thereby contributing to the advancement of corpus linguistics research.

As we continue to refine and expand the capabilities of CorpuScript, we invite collaboration and feedback from both the linguistic and software engineering communities. The goal is to further enhance its functionality and broaden its applicability to other scientific domains, ultimately contributing to more efficient, accurate, and comprehensive linguistic research. While the current version of CorpuScript has already demonstrated significant value, several points for future enhancement have been identified.

These future developments aim to further enhance the software's functionality, adaptability, and integration with existing research workflows, solidifying its role as an essential tool in corpus linguistics research.

AKNOWLEDGEMENTS

The authors wish to thank the following organizations for supporting and funding the research reported here: Federal University of Minas Gerais and Grant #APQ01173-22, Minas Gerais, Research Foundation (FAPEMIG).

REFERENCES

ÄDEL, Annelie. Corpus compilation. In: PAQUOT, Magali; GRIES, Stefan Th (Eds.). **A practical handbook of corpus linguistics**. Springer Nature, 2020.

CHARLES, Maggie. Same task, different corpus. In: BOULTON, Alex; LEŃKOSZYMAŃSKA, Agnieszka. **Multiple affordances of language corpora for datadriven learning**. John Benjamins 2015, p. 131-154, 2015.

GRIES, Stefan; NEWMAN, John. Creating and using corpora. In: PODESVA, Robert J.; SHARMA, Devyani (Ed.). **Research methods in linguistics**. Cambridge University Press, 2014.

MCENERY, Tony; BROOKES, Gavin. Building a written corpus: what are the basics?. In: O'KEEFFE, Anne; MCCARTHY, Michael (Ed.).**The Routledge handbook of corpus linguistics**. 2nd Edition. Routledge, 2022. p. 35-47.

POOLE, Robert. **A guide to using corpora for English language learners**. Edinburgh University Press, 2018.

REPPEN, Randi. Building a corpus: what are key considerations?. In: O'KEEFFE, Anne; MCCARTHY, Michael (Ed.).**The Routledge handbook of corpus linguistics**. 2nd Edition. Routledge, 2022. p. 13-20.

HOW TO USE SHAPE AND STEM CORPORA TO HELP RESEARCH-PAPER WRITING IN ENGLISH FOR ACADEMIC PURPOSES CLASSES[130]

Paula Tavares PINTO[131]
Luciano Franco da SILVA[132]
Talita SERPA[133]
Diva Cardoso de CAMARGO[134]

RESUMO: Corpora do tipo "faça-você-mesmo" são bancos de dados linguísticos poderosos que podem ser usados para apoiar a redação acadêmica e a tradução nas áreas de Humanidades, Ciências e Matemática (VANTAROLA, 2002; MAIA, 2002; FRANKENBERG-GARCIA, 2019; CARVALHO et al., 2021). Este trabalho discutirá as possibilidades de compilar rapidamente dois corpora especializados nas áreas de SHAPE e STEM com a ferramenta AntCorGen (ANHONY, 2019). Ambos os corpora serão explorados com o Sketch Engine (KILGARIFF, 2004) para mostrar como os pesquisadores podem usá-los para redigir seus próprios artigos acadêmicos. Os leitores aprenderão maneiras de encontrar adjetivos e verbos frequentes e relevantes, bem como blocos lexicais usados para dar ênfase à escrita acadêmica. Além disso, eles encontrarão formas específicas de explorar os corpora de SHAPE e STEM para identificar estruturas acadêmicas recorrentes para cada seção de um artigo acadêmico, ou seja, introdução, metodologia, discussão e conclusões.

Palavras-chave: linguística de corpus; redação de artigos científicos; corpora DIY; disciplinas SHAPE e STEM.

ABSTRACT: Do-it-yourself corpora are powerful language databases that can be used to support academic writing and translation in the areas of Humanities, Science and Math (VANTAROLA, 2002; MAIA, 2002; FRANKENBERG-GARCIA, 2019; CARVALHO et al., 2021). This paper will discuss the possibilities of quickly compiling two specialized corpus in the areas of SHAPE and STEM with the tool AntCorGen (ANHONY, 2019). Both corpora will be explored with Sketch Engine (KILGARIFF, 2004) to show how researchers can use them to write their own research papers. Readers will learn about ways to find frequent and relevant adjectives and verbs, as well as lexical bundles that are used to bring emphasis to their academic writing. Also, they will find specific ways to explore SHAPE and STEM corpora to find recurrent academic structure for each research paper section, that is to say, the introduction, methodology, discussion and conclusions.

Keywords: corpus linguistics; research paper writing; DIY corpora; SHAPE and STEM disciplines.

---

[130] Based on Pinto et al. (2024), available at <https://lume.ufrgs.br/bitstream/handle/10183/272634/001197497.pdf?sequence=1> Access on Oct. 12th, 2024.

[131] 2 Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo(CNPq).paula.pinto@unesp.br.

[132] Instituto Federal do Paraná (IFPR), Goioerê, Paraná.

[133] Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo(CAPES).

[134] Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP), São José do Rio Preto, São Paulo(CNPq).

Introduction

Writing research papers in English may be a challenge for newcomer authors at the beginning of their academic careers. For those who are non-native speakers of English and did not have the chance to use academic English with frequency it may be even harder. Most of the time these researchers are used to reading scientific papers, but do not have much experience in writing them, which may take years of experience and hard work.

Some of the scholars who have studied academic writing in depth are Swales and Feak (2004, 2009), Hyland (2004, 2014), Lee and Swales (2006), Cortes (2007), Flowerdew (2010). Even though these authors have widely described the features of academic writing, there are some characteristics that may still not be as salient for novice researchers such as the use of academic collocations and lexical bundles. Some authors use word combinations that do not sound natural to their scientific community and this may impair their article acceptance. Some of the scholars who have pointed out the academic issues found in research papers of non native speakers of English are Charles (2012), Howarth (2013), Chang and Swales (2014), Karpenko-Seccombe, (2020), Pinto et al. (2021) and Pinto et al. (2024).

In this context, Corpus Linguistics has played an important role in providing a range of writing tools to help researchers from different fields to find language patterns in academic discourse that are recognized by their peers. This happens because authors will rely on large collections of academic texts, hereafter, corpora, which can show them how their research community generally writes and the specific terminology and frequent patterns that can be rapidly identified and retrieved for writing purposes. This methodological approach can be used in different areas such as Math, Humanities and Biological areas. In order to do that, authors can use pre-compiled specialised corpora or compile their own collection of research papers published in high impact journals and use them as a Do-it-Yourself corpus (VANTAROLA, 2002; MAIA, 2002; FRANKENBERG-GARCIA et al., 2019; PINTO et al., 2024). By using corpora, the writer will be able to observe the useful information according to his specific needs and will develop an autonomous process of learning that will lead him to mastering the academic English based on his interpretation of his peers' writing.

This paper will discuss how specialised corpora can be explored by researchers who want to compile their own language database to help them write different sections of their own research papers. We will illustrate our proposal by taking examples from SHAPE disciplines, which involve Social Sciences Humanities, Arts for People and Economy, as well as STEM disciplines, which involve Science, Technology, Engineering, and Mathematics.

AntCorGen for the compilation of SHAPE and STEM areas

AntCorGen (ANTHONY, 2019) is a tool used to quickly compile specialised corpora with research papers from the PLOS one platform. A tutorial video of this tool was recorded by its creator in a short video[135]. Below we will talk about the

[135] AntCorGen tutorial <<https://www.youtube.com/watch?v=WrsIzE9to4o> access on June 30th, 2023. [7] PLOS available at <https://plos.org/about/> access on June 27th, 2023.

compilation of SHAPE Plos and STEM Plos and their exploration for academic writing.

SHAPE disciplines stand for Social Sciences Humanities, Arts for People and Economy. All these disciplines and subareas can be found at PLOS, which is a nonprofit, open access multi-disciplinary publisher[7]. All areas of SHAPE can be easily accessed in AntCorGen and the researcher can choose the parts of research papers he wants to analyse. Since we wanted to have mostly written material we selected the articles' abstracts, introduction, materials & methods, results & discussion and conclusions.

We called this corpus SHAPE Plos and, since it was compiled for describing the process in this chapter, we set the maximum of 100 articles , but it is possible to have a much larger study corpus if we wanted to. After this compilation we had a study corpus of 445,291 words to be explored.

STEM disciplines are related to both Biology and Hard Sciences. Although the figure below seems to have only Biology and Life Sciences, the actual list of disciplines selected was longer and we could include areas such as Math and Computer Sciences as well. In the same way, we selected 100 articles for STEM Plos corpus.

After this compilation, we had a specialised corpus of STEM disciplines with a total number of 297,255 words to be observed and compared to the results from SHAPE Plos.

Analyses with Sketch Engine

We uploaded both corpora, SHAPE and STEM to Sketch Engine (KILGARIFF, 2014) so we would be able to observe the frequent adjectives and verbs in each broad area and see the similarities and differences between them. We could also generate concordance lines with search words, terms and phrases that can be used by researchers to explore and observe how international researchers in their area have been writing different sections of their research papers.

Building your Research paper with SHAPE Plos and STEM Plos corpus

If a researcher wants to have examples of research papers in SHAPE and STEM disciplines, he can search for common expressions in the corpus. In our case, we have divided both subcorpora into research sections that are usually found in research articles. Based on Karpenko-Seccombe (2020), we are going to discuss how researchers can use their own specialised corpora for writing their research papers.

Writing the Introduction Section

According to Swales and Feak (2009, 2011), a research paper introduction typically contains three main steps or *moves*: a) establishing the area of research, where the author will show the importance of a field and introduce previous research in his area; b) establishing a gap in the knowledge or problem to be solved and c) presenting his paper, where he will identify his objectives, introduce expected outcomes and describe the structure of his work.

In order to explore introductions in SHAPE Plos and STEM Plos corpora, we searched for concordance lines with the query phrase " this paper" and we selected some of the lines to be used as examples here:

1. **This paper** attempts to fill the gap of existing research concerning the link between public pension and fertility. [SHAPE Plos]

2. In **this paper** ,we perform a comprehensive survey of the worldwide linguistic landscape as emerging from mining the Twitter microblogging platform. [SHAPE Plos]

3. In **this paper** , we are interested in measuring linguistic regularities both at the level of word structure and at the level of word order. [SHAPE Plos]

4. **This paper** explores the ways abortion attitudes intersect with causal beliefs about gender categories, within the unique social context of a national referendum held to legalise abortion in the Republic of Ireland. [SHAPE Plos]

5. In **this paper**, we introduce a novel mobile application called "Medikamentenplan" ("Medication Plan"), which was developed to support medication compliance and vital sign documentation. [STEM Plos]

6. In **this paper**, we propose a concise, improved and effective privacy framework for wearable device manufacturers, as well as application developers, capable of providing greater privacy and security to the wearable device owners. [STEM Plos]

7. **This paper** innovatively proposes countermeasures to improve the innovation of e-commerce practitioners in rural areas. [STEM Plos]

8. The objective of **this paper** is to outline our approach of establishing and implementing this IT infrastructure. [STEM Plos]

We can see that authors from SHAPE and STEM use similar strategies to introduce their research papers. In 1, 4 and 7, authors used the structure *This paper + [adverb] + verb (infinitive)*. In examples 2, 3, 5 and 6, authors opted to use *In this paper + we + verb (infinitive)*. Finally, in example 8, the author preferred to introduce his paper by using the structure *The objective of this paper is + to + verb (infinitive)*.

We can see a pattern in the previous examples that can be used in a more confident way by researchers of SHAPE and STEM.

Final Considerations

In this paper we presented an overview on how to compile specialised corpora in SHAPE and STEM with the AntCorGen tool and how researchers can use those corpora to access the academic language used by their peers. By doing so, researchers will confirm or refute ways of presenting their studies according to each research paper section, as well as the best way of describing their methodological approach, and call attention to their studies contribution. We hope this chapter may inspire research teams to start building their own language database that can be used by future members and can be constantly updated.

Acknowledgments
The authors would like to thank the support by CNPq (Process Number #307287/2021-1); FAPESP (Process Number # 2022/05908-0); CAPES.

References

ANTHONY, L.. AntCorGen (Version 1.1.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software, 2010.

FRANKENBERG-GARCIA, A.; BOCORNY, A. E. P.; TAVARES-PINTO, P.; SARMENTO,
S. Supporting the internationalization of Brazilian research. *Workshops delivered at the Federal University of Rio Grande do Sul and at São Paulo State University*, Porto Alegre and São José do Rio Preto, April-June 2019. 2019.

CARVALHO, C. T. de; LARANJA, L. A. N.; PINTO, P. T. DIY Corpora: o que são e para quem são? *Tradterm*, v. 37, n. 1, p. 64-87, 2021. Available at: https://doi.org/10.11606/issn.2317-9511.v37p64-87. Access on Oct. 12th,. 2024.

CHANG,, Y. Y., & SWALES, J. M. Informal elements in English academic writing: threats or opportunities for advanced non-native speakers?. In *Writing: Texts, processes and practices* (pp. 145-167). Routledge, 2014

CHARLES,, M. 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, *31*(2), 93-102, 2012.

CORTES, V. Exploring genre and corpora in the English for academic writing class. *The ORTESOL Journal*, *25*, 8-14, 2007.

FLOWERDEW, L. Using corpora for writing instruction. *The Routledge handbook of corpus linguistics*, 444-457, 2020

HOWARTH, P. A. *Phraseology in English academic writing*. Max Niemeyer Verlag, 2013

HYLAND, K. *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press 2004.

HYLAND, K. Disciplinary discourses: Writer stance in research articles. In: _____. *Disciplinary discourses*: Social interactions in academic writing. 2. ed. Londres: Routledge, 2014. p. 99-121.

KARPENCO-SECCOMBE, T. *Academic writing with corpora: A resource book for data-driven learning*. Routledge, 2020.

KILGARIFF, A., BAISA, V., BUSTA, J., JAKUBÍČEK, M., KOVÁR, V., MICHELFEIT, J., ... & SUCHOMEL, V. The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36, 2014.

LEE, D., & SWALES, J. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for specific purposes*, *25*(1), 56-75, 2006.

MAIA, B. Do-it-yourself, disposable, specialised mini corpora–where next? Reflections on teaching translation and terminology through corpora. Cadernos de tradução, 1(9), 221-235, 2002.

PINTO, P.T.; SILVA, Luciano Franco da ; SERPA, TALITA ; CAMARGO, DIVA CARDOSO DE . Do-It-Yourself Corpora to Support SHAPE and STEM Research Paper Writing. In: SARMENTO, S., REBECHI, R., MATTE M. L.. (Org.). *English for Academic purposes: reflections, description e pedagogy*. 01ed.Porto Alegre: Zouk, 2024, v. 01, p. 97-126.

PINTO, P. T.; CAMARGO, D. C. de; SERPA, T.; SILVA, L. F. da. Analysing the behaviour of academic collocations in a corpus of research-papers: a data-driven study/Analisando o comportamento de colocações acadêmicas em um corpus de artigos científicos: um estudo dirigido por dados. *Revista de Estudos da Linguagem*, v. 29, n. 2, p. 1229-1252, 2021.

SKETCH ENGINE <https://auth.sketchengine.eu/#login> Access on March 13th, 2024.

SWALES & FEAK, C. B. *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). Ann Arbor, MI: University of Michigan Press, 2004.

SWALES & FEAK, C. B. *Abstracts and the writing of abstracts* (Vol. 2). University of Michigan Press ELT, 2009.

TAVARES-PINTO, P.; REES, G.; FRANKENBERG-GARCIA, A. Identifying collocation issues in English L2 research article writing. In: CHARLES, Maggie; FRANKENBERG-GARCIA, Ana (org.). *Corpora in ESP/EAP Writing Instruction*: Preparation, Exploitation, Analysis. 1. ed. Londres: Routledge, 2021. p. 01-20.

"VOCÊ ESTÁ TENDO PRAZER COM SEU TORTURADOR?" A CONDIÇÃO FEMININA NOS RELATOS DE TORTURA À COMISSÃO NACIONAL DA VERDADE

Eu sou leve, sabe, eu tô viva, estamos vivos, vamos ficar vivos. Por que olhar pra trás? Não vive quem fica arrastando cordéis de caixões

Regina Duarte

Giovana de Castro MARCHESE [136]
Luciana Carvalho FONSECA [137]

RESUMO: Esta pesquisa faz uma análise testemunhos de mulheres que foram presas políticas durante o regime ditadura empresarial-militar no Brasil à Comissão Nacional da Verdade, com o fim de infestigar como o gênero é performado em seus discursos ao narrarem a violência sexual sofrida nas sessões de tortura. Este estudo lança mão de Foucault (1970, 1975), Butler (2018, 2004) e Segato (2022) para as análises, e da ferramenta Sketch Engine para investigação do corpus.

Palavras-chave: estupro; ditadura militar; patriarcado; memória; Sketch Engine.

Introdução

Em tempo de revisionismo histórico, trabalhos em prol da memória pública são um imperativo ético no Brasil para que possamos recuperar a força das lutas sociais que trazem à luz a violência do regime ditatorial de 1964 a 1985. A epígrafe deste resumo é um trecho retirado de uma entrevista concedida por Regina Duarte, secretária de Cultura do governo Bolsonaro em 2020, à CNN

Brasil no dia 7 de maio de 2020[3]. Quando questionada pelo jornalista Daniel Adjuto sobre tortura durante a ditadura militar no Brasil, a secretária ri e afirma que "na humanidade, não para de morrer" e que "sempre houve tortura[138]". Ao minimizar as mortes e sevícias causadas pelo Estado brasileiro, Regina Duarte contribui para a naturalização da violência no país, fazendo eco com os discursos do então presidente Jair Bolsonaro.

---

[136] Doutoranda no Programa de Estudos Linguísticos e Literários em Inglês do Departamento de Letras Modernas – DLM, Faculdade de Filosofia, Letras e Ciências Humanas – FFLCH, Universidade de São Paulo – USP, São Paulo, SP. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Email: giovana.cmr@usp.br

[137] Professora Doutora nos Programas de Pós-Graduação Estudos Linguísticos e Literários em Inglês (ELLI) e Letras Estrangeiras e Tradução (LETRA), Departamento de Letras Modernas – DLM, Faculdade de Filosofia, Letras e Ciências Humanas – FFLCH, Universidade de São Paulo – USP, São Paulo - SP

[138] A entrevista pode ser assistida na íntegra no canal da CNN Brasil no YouTube: https://www.youtube.com/watch?v=v9gLHrP7RNw Acesso em: 07 de outubro de 2024

A normalização do regime ditatorial e, consequentemente, da tortura e a desqualificação das memórias críticas à ditadura e dos atores a ela relacionados nos impossibilitam de lidar com nossos erros históricos, o que contribui para a perpetuação das estruturas de violência e de opressão. Pauta (não tão) velada do governo Bolsonaro (2019-2022), que tinha no torturador Carlos Brilhante Ustra um herói. Um exemplo dessa perpetuação da violência é o aumento de 10,8% dos casos de feminicídio no primeiro semestre de 2021 em relação a 2019, primeiro ano do governo bolsonarista.

Dessa forma, buscando desempenhar o meu papel de cidadã brasileira e pesquisadora em consonância com a perspectiva delineada por Pedretti (2021, p. 54), que postula a responsabilidade das instituições públicas e da sociedade civil, incluindo acadêmicos e movimentos sociais, no esforço de reorganizar as demandas por *memória, verdade e justiça* após quatro anos de desarticulação dessas iniciativas, apresento esta pesquisa que examina os relatos de tortura durante a ditadura militar submetidos à Comissão Nacional da Verdade (doravante CNV, 2012 – 2014). O enfoque deste estudo está nas narrativas femininas que delineiam a violência de gênero perpetrada contra mulheres, a fim de investigar, com base em uma análise crítica do discurso, como gênero é performado nesses discursos.

Fundamentação teórico-metodológica

O objeto de estudo desta pesquisa é um corpus monolíngue em português, *offline*[139], intitulado CNV_Mulheres, composto pela transcrição dos depoimentos de presas políticas à CNV entre 2012 e 2014[140]. O corpus é composto por 103 depoimentos.

A ferramenta *Sketch Engine*, software de análise linguística desenvolvido pela Lexical Computing CZ, foi usada para, através das palavras-chave, levantar o tema central do corpus. Durante a investigação, me chamou atenção o grande número de palavras relacionadas à violência sexual, como *estuprar, violentar, abuso, (chamar de) puta*. Foi feito um levantamento das palavras-chave dentro desse campo lexical e, acessando as linhas de concordância correspondentes a elas, foi dado início à análise dos relatos.

Assim, com base em Foucault (1970, 1975), Butler (2018, 2004) e Segato (2022) esta pesquisa investiga como gênero é performado nos discursos sobre tortura sexual de ex-presas políticas em seus testemunhos à CNV. Discussões preliminares lançam luz sobre discursos de incerteza do estupro, negação do estupro, dessubjetivação política da mulher militante e menstruação como instrumento de tortura.

---

[139] Para os vários tipos de corpora, ver TAGNIN, Stella E. O**.** A Linguística de Corpus na e para a Tradução. In: TAGNIN, Stella E.O.; VIANA, Vander (Org.). **Corpora na Tradução**. São Paulo: Hub Editorial, 2015. p. 19-56.

[140] As transcrições dos depoimentos podem ser encontradas no site da própria CNV: http://cnv.memoriasreveladas.gov.br/todos-volume-1.html Acesso em: 08 de setembro de 2023.

Discussão de dados: negação do estupro

No corpus, muitas das depoentes que negam estupro nas sessões de tortura fazem uso da dupla, ou até mesmo tripla, negação. Por exemplo, Karen Leslie Raborg Sage Keilt ao ser questionada por Mezarobba, da CNV, se ela havia sido estuprada na cela onde aconteciam as torturas, responde: "*Não, não*". Já, Dagmar Pereira da Silva faz uso da tripla negação ao ser questionada por um interlocutor não identificado da CNV se houve abuso sexual durante as torturas, respondendo: "*Não, não, não*". Ana Maria Ramos Estêvão, por sua vez, explica que foi ameaçada de estupro, mas reitera que "*não* chegaram a cumprir, *não*." Ainda, Leslie Denise Beloque, menciona que havia diferenças na intensidade das torturas a depender da equipe responsável, "mas *nunc*a, por exemplo, *nenhuma* tentativa de assédio sexual, insinuações ou ameaças de estupro" (grifos da autoria).

A dupla negação poderia estar relacionada tanto à recusa de ingresso no tema abuso sexual quanto à ênfase da negação do abuso sexual em si, com o propósito de evidenciar que a violência sexual de fato não ocorreu. Nos dois casos, a escolha das depoentes pode estar relacionada à culpabilização das vítimas de estupro na sociedade, que as estigmatiza e as coloca como merecedoras dessa violência devido à sua recusa à performance de gênero esperada pelo patriarcado (BUTLER, 2018). Essa formação discursiva da mulher violentada como merecedora da violência aparece muito claramente na fala de Lúcia Maria Sálvia Coelho quando explica por que acredita não ter sido violentada em suas sessões de tortura:

Essa parte sexual não me fizeram, porque eu estava em tamanho pânico. Mas eu acho que eu devia, no começo, estar com uma cara muito realmente do tipo que me criaram, de professora séria.

Se, de acordo com Orlandi (1999) "há sempre no dizer um não-dizer necessário", não é difícil descortinar o não dito no discurso da depoente. Aqui, Lúcia Maria performa gênero em seu discurso de acordo com o esperado pela ordem patriarcal: a mulher que tem cara de "professora séria" não é passível de violência sexual. Mas, se a mulher com cara de "professora séria" não sofre violência sexual, quem são as outras que sofrem? Segundo Maria Dalva Leite de Castro de Bonet, as putas. A militante explica que os militares buscavam convencê-la de que algumas mulheres, por serem putas, não ligavam para a violência sexual que sofriam. Sua fala começa sendo interrompida pela voz do próprio torturador, na forma de discurso direto, enfatizando que a opinião é de uma terceira pessoa, não a dela:

Tem mulher que chega aqui nem se liga pra isso" pra tentar formular na tua cabeça que você é especial. E elas são as putas e você...na dor acredita em qualquer coisa.

Nessa formação discursiva ecoa a crença de que a mulher que não se desvia das regras de comportamento social impostas a ela pelo patriarcado tem menos chance de ser vítima de violência sexual. Ou seja, aquelas mulheres que

> foram violentadas ou estupradas teriam dado alguma razão para que o crime ocorresse. Dessa forma, associada à negação da violência sexual, temos a justificativa da violência sexual. Em muitos casos, a mulher está consciente do crime sexual que sofreu durante o encarceramento, mas, após a liberdade, se silencia, pois acredita tê-lo merecido. Esse ato de silenciar-se não é em si uma escolha, mas uma resposta aos mecanismos de culpa e vergonha que foram inculcados nessas mulheres por meio de discursos patriarcais que estabelecem quem sofreu o estupro e por que o sofreu. A justificativa da violência sexual, nesse sentido, serve como ferramenta de manutenção da ordem social e da dominação (FOUCAULT, 1975), consolidando ainda mais o poder patriarcal sobre corpos e subjetividades.

Ainda nessa direção, era comum que, durante a tortura sexual, os militares buscassem induzir a vítima a acreditar que estava tendo prazer com eles, e que, se não colaborasse, seus companheiros ficariam sabendo do ocorrido. É o que relata Ana de Miranda Batista à CNV:

[...] a tortura era a seguinte, também, além de todas as outras: "Você sabe onde você está?" Voz bem cava, "Você sabe onde você está?", "Você está tendo prazer com o seu torturador?" E começava a bolinar o teu corpo todo. "E você sabe que o que seus companheiros vão dizer, que você gozou com um torturador?", "Você não vai poder sair da prisão, você vai ter que ficar do nosso lado porque se não nós vamos contar para os seus companheiros o que você fez aqui".

De acordo com TEGA (2019), essa estratégia contribui para a desorganização da mulher torturada e prejudica o trabalho de resolução do trauma. A pergunta “Você está tendo prazer com o seu torturador?” pode levar à mulher torturada a questionar se houve ou não consentimento na violência sofrida. Isso ocorre de tal maneira que muitas mulheres ainda acreditam que houve consentimento na violência sexual a qual foram submetidas, como explica Miriam Lewin (2013) a respeito de suas companheiras presas na Escola Superior de Mecânica da Armada (ESMA), em Buenos Aires. Ainda, ao ameaçar contar aos companheiros de Ana que ela teria “gozado” com seu torturador, o torturador faz de seu corpo um espaço não apenas de submissão, mas também de silenciamento, uma vez que o medo da estigmatização por parte da vítima é condição recorrente em crimes de estupro, como discutido acima. É importante comentar que a estratégia discursiva de Ana de trazer para seu relato a voz do torturador por meio da citação direta descortina, como afirma Seligmann-Silva (2008), a memória como um misto de verbalidade e imagens. A imagem do torturador do passado interrompendo o discurso de Ana no presente em primeira pessoa indica a atemporalidade da situação traumática. Isso se dá porque, segundo Levi (1990), o trauma é a memória de um passado que não passa.

Considerações finais

Voltemos ao convite de Regina Duarte em nossa epígrafe: “vamos ficar vivos”. Sem memória, verdade ou justiça, vivos sim, mas em quais condições? Quando nos é negado o direito de olhar para os erros do passado, ou quando esses erros são minimizados, a possibilidade de transformação é anulada.

Os relatos que estão sendo analisados mostram que as sessões de tortura são um microcosmo do patriarcado. Os donos da vida (SEGATO, 2022) irão disciplinar os corpos femininos, buscando torná-los submissos e “dóceis” (FOUCALT, 1975) para que ocupem o lugar social relegado às mulheres dentro do patriarcado: o lugar do espaço privado, procriando, voltadas para o cuidado familiar e subalternizadas. Para Kehl (2010), tornar públicas as lutas que foram esquecidas é primordial “na elaboração de traumas sociais” (p. 128), afinal, aquilo que não somos capazes de elaborar, tendemos a repetir. A violência institucionalizada no Brasil contra as mulheres é um exemplo disso.

Referências bibliográficas

BUTLER, Judith. **Problemas de gênero**: feminismo e subversão da identidade. Rio de Janeiro: Civilização Brasileira, 2018. 303 p.

ESTEVEZ, Alejandra (org.). **Lembrar é agir**: memória, verdade e direitos humanos. São Paulo: Letra e Voz, 2021. p. 53-68.

FOCAULT, Michel. **A ordem do discurso**: aula inaugural no college de france, pronunciada em 2 de dezembro de 1970. 24. ed. São Paulo: Edições Loyola, 1970. 74 p. Edição 2014, tradução de Laura Fraga de Almeida Sampaio.

FOCAULT, Michel. **Vigiar e punir**: nascimento da prisão. 16. ed. Petrópolis: Vozes, 1975. Edição de 2014. Tradução de Raquel Ramalhete.

KEHL, Maria Rita. Tortura e sintoma social. In: TELES, Edson; SAFATLE, Vladimir (org.). **O que resta da ditadura**. São Paulo: Boitempo, 2010. p. 123 - 132.

LEVI, Primo. **Os afogados e os sobreviventes**: os delitos, os castigos, as penas. São Paulo: Paz & Terra, 1990. 168 p.

TEGA, Danielle. **Tempos de dizer, tempos de escutar**: testemunhos de mulheres no Brasil e na Argentina. São Paulo: Intermeios, 2019. 271 p.

PEDRETTI, Lucas. Entre políticas de memória e camadas de esquecimento. In: VIÑAR, Maren e Marcello. **Exílio e Tortura**. São Paulo: Escuta, 1992. 154 p.
SEGATO, Rita. **Cenas de um pensamento incômodo**: gênero, cárcere e cultura em uma visada decolonial. Rio de Janeiro: Bazar do Tempo, 2022. 256 p. Tradução de Ayelén Medail.

SELIGMANN-SILVA, Márcio (org.). **O Espaço Biográfico**: catástrofe e representação. [S.l]: Editora Escuta, 2008. 264 p.

ANÁLISE MULTIDIMENSIONAL LEXICAL EM CORPORA DE RESENHAS E VIDEORRESENHAS *ONLINE*: UMA ABORDAGEM DA LINGUÍSTICA DE CORPUS COMO ÁREA AUTÔNOMA DE PESQUISA CIENTÍFICA

Mauricio José Ferreira LOPES[141]

Resumo: Este estudo analisa variações léxico-discursivas em resenhas escritas e videorresenhas de influenciadores literários no Instagram e YouTube. A Linguística de Corpus (LC) e a Análise Multidimensional (AMD) foram usadas para identificar padrões linguísticos (Biber, 1988; Berber Sardinha, 2000). A Análise do Discurso (AD) de Pêcheux auxiliou na interpretação das práticas discursivas ideológicas e sociais (Pêcheux, 2010). A combinação de LC com Inteligência Artificial (IA) ampliou as possibilidades de análise, especialmente em relação à formação de comunidades digitais (Silva, 2019).

Palavras-chave: Linguística de corpus; análise multidimensional; práticas discursivas; redes sociais; resenhas literárias.

Introdução

O estudo investiga as variações léxico-discursivas em corpora de resenhas escritas e videorresenhas literárias, produzidas por cinco influenciadores digitais literários (IDLs) cujos perfis e canais são hospedados nas redes sociais Instagram e YouTube. A pesquisa posiciona a Linguística de Corpus (LC) como uma área de investigação científica autônoma, utilizando a Análise Multidimensional (AMD) para identificar padrões linguísticos em registros distintos, com base nos métodos de Biber (1988) e Berber Sardinha (2000). Para além da análise quantitativa oferecida pela AMD, a Análise do Discurso (AD) de Pêcheux é incorporada ao estudo, a fim de interpretar as práticas discursivas, observando-se suas formações ideológicas e sociais, conforme abordado por Pêcheux (2010). A intersecção entre LC e AD permite uma análise integrada dos registros, revelando como as práticas discursivas refletem contextos sociais e como as dimensões discursivas emergentes mostram formações ideológicas subjacentes aos discursos dos influenciadores. Além disso, o uso de técnicas de Inteligência Artificial (IA) possibilita uma análise mais sofisticada de grandes volumes de dados linguísticos, oferecendo novas oportunidades para investigar os discursos produzidos em plataformas digitais, como observado por Silva (2019).

O estudo examina como influenciadores literários configuram suas práticas discursivas de acordo com o público-alvo e as características das diferentes plataformas. Resenhas publicadas no Instagram tendem a apresentar uma abordagem mais introspectiva e analítica, enquanto as videorresenhas no YouTube enfatizam a comunicação direta e a interação com o público. A combinação entre LC e IA permite uma análise mais precisa de gêneros e subgêneros literários, oferecendo insights valiosos sobre as dinâmicas discursivas em plataformas digitais. A pesquisa destaca o papel fundamental da LC como ciência autônoma e interdisciplinar, que fornece uma compreensão

[141] Professor de Língua Estrangeira na rede pública municipal de São Paulo. Mestre e doutorando em Linguística Aplicada e Estudos da Linguagem pela PUC-SP, bolsista CAPES.

crítica das práticas discursivas contemporâneas e suas implicações sociais e ideológicas. O estudo, assim, contribui para a consolidação da LC como uma área científica que dialoga com outras disciplinas, em função de sua natureza transdisciplinar, ampliando o escopo da análise linguística em contextos digitais e colaborando para a formação de comunidades discursivas online e a disseminação do conhecimento literário.

Questões de Pesquisa

Como se caracterizam os discursos subjacentes às dimensões discursivas que emergem da análise fatorial dos corpora de resenhas e videorresenhas literárias?

De que forma a Linguística de Corpus, integrada com a Inteligência Artificial, pode ser reconhecida como uma área autônoma de pesquisa científica, além de uma mera abordagem metodológica?

Quais são as tendências e preferências literárias dos IDLs no Instagram e no YouTube, e como essas escolhas refletem as interações culturais e sociais das comunidades de leitores?

Objetivos

Analisar e interpretar as dimensões discursivas emergentes das variáveis fatoriais nos corpora CLIRI e CLIVY.

Argumentar a favor da LC como área autônoma de pesquisa científica, explorando sua integração com técnicas de IA para análise de grandes volumes de dados linguísticos.

Fundamentação Teórica

A LC, desde o seu início com o desenvolvimento do Brown Corpus nos anos 1960, evoluiu de uma metodologia de análise de linguagem baseada em corpora para uma área de pesquisa com princípios teórico-epistemológicos robustos. Biber (1988) introduziu a Análise Multidimensional (AMD), que permite identificar dimensões de variação linguística em corpora, utilizando análise fatorial para mapear padrões coocorrentes de características léxico-discursivas. Esta abordagem é fundamental para compreender como registros distintos, como resenhas e videorresenhas, se diferenciam em função de suas características discursivas.

A LC fornece uma base empírica para identificar padrões de uso linguístico, enquanto a AD corrobora as dimensões discursivas subjacentes a esses padrões à luz das práticas sociais e formações ideológicas. Ambas reconhecem a dimensão social e ideológica da linguagem, destacando como as práticas discursivas refletem e moldam as estruturas sociais. A integração dessas duas abordagens permite uma análise mais rica e multifacetada dos textos e discursos, proporcionando uma compreensão mais profunda e abrangente de como a linguagem funciona como um fenômeno social e ideológico nas redes sociais (BIBER, 1995; PÊCHEUX, 2010).

Ademais, a integração com a AD, particularmente a de linha francesa, enriquece as interpretações ao fornecer um quadro teórico para entender os significados subjacentes aos padrões identificados. Essa combinação permite

uma análise mais holística, profunda e abrangente dos dados, considerando tanto a quantificação linguística quanto a interpretação dos contextos sociais e culturais em que esses discursos estão inseridos.

Metodologia

Design dos Corpora

Foram criados dois corpora:

- CLIRI (Corpus de Resenhas do Instagram): Composto por 517 resenhas de IDLs, abrangendo um total de 144.358 palavras.
- CLIVY (Corpus de Videorresenhas do YouTube): Composto por 504 videorresenhas, transcritas automaticamente, totalizando 1.252.978 palavras.

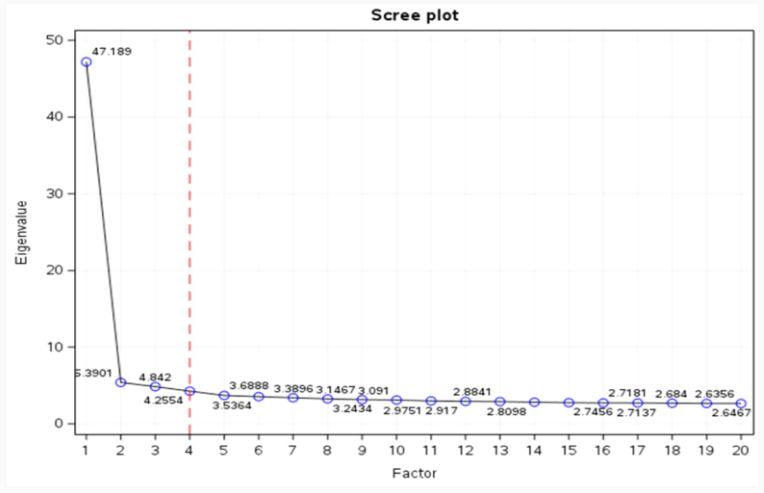
Procedimentos Metodológicos

Coleta de Dados: As resenhas foram coletadas de perfis e canais de IDLs no Instagram e YouTube, focando em influenciadores com alto engajamento e produção crítica.

Processamento dos Dados: As videorresenhas foram transcritas usando software de reconhecimento de fala, seguido de uma normalização linguística para remover ruídos textuais.

A análise fatorial identificou quatro fatores principais com base no critério de índice de variância (*eigenvalues*) acima de 1. A linha vermelha pontilhada no gráfico a seguir marca a posição do ponto de corte, indicando que os fatores além do quarto têm pouca relevância para a explicação da variabilidade total dos dados.

Gráfico 1: *Eigenvalues* ou índice de variância associado ao respectivo fator extraído da análise processada pelo Fonte SAS *University Edition*.



Os fatores com *eigenvalues* mais altos (à esquerda do ponto de inflexão) são os mais significativos, explicando a maior parte da variância

nos dados. Neste gráfico, os primeiros quatro fatores apresentam *eigenvalues* substancialmente maiores, indicando que são os mais relevantes para descrever as variações léxico-discursivas nos corpora analisados. A linha vermelha pontilhada sugere que os primeiros quatro fatores devem ser retidos para uma interpretação mais eficaz, pois representam a maior quantidade de variação explicada nos dados.

Dimensões Discursivas: Aplicação da AMD lexical para identificar padrões lexicais e discursivos. Quatro dimensões principais foram identificadas:

Dimensão 1: Comunicação e Interatividade vs. Introspecção e Análise. Dimensão 2: Motivação e Engajamento Cultural vs. Reflexão e Realismo Existencial. Dimensão 3: Análise Crítica e Psicológica vs. Contextualização Histórica e Descritiva.

Dimensão 4: Análise Pontual da Narrativa vs. Foco na Comercialização do Livro.

Interpretação dos Fatores: As dimensões foram interpretadas discursivamente, capturando as propriedades comunicativas primordiais detectadas nos registros pela AMD lexical.

Discussão dos Dados

Os resultados indicam que há variações significativas nas abordagens de resenhas e videorresenhas literárias:

Dimensão 1: Videorresenhas focam na comunicação direta e criação de uma comunidade, enquanto as resenhas do Instagram são mais introspectivas e analíticas.

Dimensão 2: Videorresenhas abordam aspectos culturais e de engajamento, enquanto as resenhas escritas exploram reflexões existenciais e sociais.

Dimensão 3: Resenhas no Instagram enfatizam a análise crítica e psicológica, enquanto as videorresenhas priorizam a contextualização histórica e descritiva das obras.

Dimensão 4: Resenhas escritas tendem a ser mais detalhadas na análise narrativa, enquanto videorresenhas focam em aspectos comerciais e logísticos da leitura.

Considerações Finais

O estudo revelou que as resenhas no Instagram tendem a adotar uma abordagem analítica e reflexiva, explorando aspectos individuais e críticos das obras literárias. Em contrapartida, as videorresenhas no YouTube priorizam a interação e o engajamento, com foco na narrativa e na comunicação com a audiência. Essa diferenciação reforça a ideia de que a profundidade intelectual e a acessibilidade não são mutuamente excludentes no domínio discursivo das redes sociais, mas podem coexistir de maneira complementar. A análise demonstrou que as redes sociais, apesar de sua natureza fragmentada e dinâmica, têm o potencial de disseminar conhecimento literário e promover debates críticos.

O estudo também sustenta a importância de reconhecer a LC como uma área científica autônoma. A abordagem da LC permite explorar criticamente a linguagem em seu contexto de uso, superando a visão limitada de que a pesquisa linguística se restringe a aspectos técnicos ou metodológicos. Assim, ao posicionar a LC como campo independente, os pesquisadores têm a oportunidade de desenvolver estudos transdisciplinares, contribuindo para uma compreensão mais profunda da linguagem e de suas manifestações na sociedade contemporânea.

Ao reafirmar a LC como área autônoma, este estudo científico destaca a importância de uma abordagem crítica e reflexiva na condução dos estudos linguísticos, evidenciando a capacidade da LC de dialogar com diversas disciplinas e enriquecer o entendimento sobre as dinâmicas discursivas nas redes sociais. Em suma, o estudo demonstrou que a LC, além de fornecer insights valiosos sobre práticas discursivas específicas, possui um papel fundamental na construção de uma teoria crítica da linguagem, consolidando-se como uma área científica capaz de produzir conhecimento inovador e relevante para o estudo da linguagem em diferentes contextos.

Essa abordagem multidisciplinar contribui para uma compreensão mais aprofundada dos discursos literários em redes sociais, mostrando como influenciadores digitais configuram suas práticas discursivas para diferentes audiências e situações comunicacionais e interativas no domínio discursivo digital.

Referências

ARAÚJO, J.; SOUSA, M. M. N.; CAVALCANTI, J. M. Comunidade discursiva e redes sociais: os resenhadores do Skoob. *Revista Intercâmbio*, v. 45, p. 28-51, 2020. Disponível em: https://revistas.pucsp.br/index.php/intercambio/article/view/50439. Acesso em: 12 jul. 2024.

BAKER, P.; McENERY, T. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Palgrave Macmillan, 2015.

BERBER SARDINHA, T. Linguística de Corpus: Histórico e Problemática. *Revista Delta*, v. 16, n. 2, p. 45-67, 2000.

BERBER SARDINHA, T.; VEIRANO PINTO, M. *Multi-Dimensional Analysis: Research Method and Current Issues*. Bloomsbury Academic, 2019.

BIBER, D. *Variation Across Speech and Writing*. Cambridge University Press, 1988.

BIBER, D.; CONRAD, S. *Register Genre and Style*. Cambridge University Press, 2009. PÊCHEUX, M. *Semântica e Discurso*. Editora Unicamp, São Paulo, 2010.

SILVA, E. Análise do Discurso e Linguística de Corpus. *Revista de Estudos Linguísticos*, v. 37, n. 2, p. 112-134, 2019.

PEDAGOGIA DA TRADUÇÃO E OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL (ODS)

Emiliana FERNANDES BONALUMI [142]
Diva CARDOSO DE CAMARGO[143]

RESUMO: Esta comunicação a respeito da pedagogia da tradução (DÍAZ FOUCES, 1999, 2001; LAVIOSA, 2008, 2010, 2020; YYY, 2016a, 2016b; SERPA et al, 2021; THOW, 2022) teve por intuito analisar termos multi-palavras em um corpus compostos de notícias originais em línguas inglesa e portuguesa sobre os ODS, em especial, item (11) Cidades e Comunidades Sustentáveis. Também fizemos uso da aprendizagem movida por dados (JOHNS, 1991, 2002; GRANGER, 1998, 2002; BERBER SARDINHA, 2004, 2006, 2010; BOULTON, 2010; LAVIOSA, 2008, 2010, PINTO, 2021; PINTO & YYY, 2021; PINTO et al. 2022). A fim de gerar os termos analisados, utilizamos a ferramenta on-line Sketch Engine. Após examinarmos as listas dos primeiros doze termos em línguas inglesa e portuguesa, foi compilado um glossário composto das similaridades e diferenças encontradas respectivamente em quatro termos de línguas inglesa e portuguesa. Outrossim, compilamos um glossário menor constituído de cinco termos, por meio de um texto previamente selecionado em língua portuguesa a respeito do item (11) dos ODS, bem como discutimos alguns dos traços de normalização de Scott (1998), fazendo uso da versão em língua inglesa elaborada pelos discentes.

Palavras-chave: Pedagogia da Tradução, Objetivos de Desenvolvimento Sustentável (ODS), Termos Multi-palavras, Glossário, Traços de Normalização.

INTRODUÇÃO

A pedagogia da tradução aborda o emprego de teorias e práticas de tradução em seu ensino (LEONARDI, 2010; YYY, 2016a, 2016b; SERPA et al, 2021). Sendo assim, julgamos importante utilizá-la na disciplina de Prática de Tradução em Língua Inglesa III em uma das unidades da Universidade Estadual do Estado de São Paulo, uma vez que acreditamos que para efetuar uma tradução, é necessário o conhecimento de sua teoria. Já, a opção de utilizar os Objetivos de Desenvolvimento Sustentável - ODS das Nações Unidas neste trabalho se deve ao fato de que se trata de um tema de saliência desde sua adoção em setembro de 2015, e vem sendo discutido por pesquisadores em âmbito nacional e internacional.

Por seu turno, a aprendizagem movida por dados tem sido desenvolvida em sala de aula desde sua criação por Johns em 1986. É uma abordagem que

[142] Docente do Curso Letras-Inglês da UFR (Universidade Federal de Rondonópolis, MT).

[143] Docente do Departamento de Letras Modernas da UNESP, São José do Rio Preto.

utiliza textos autênticos extraídos dos corpora para diversas finalidades. Recorremo-nos à aprendizagem movida por dados a fim de elaborar as listas de frequência dos termos multi-palavras, fazendo uso da ferramenta on-line Sketch Engine (https://www.sketchengine.eu/).

Com esse propósito, compilamos um corpus jornalístico composto respectivamente de oito textos originais em língua inglesa e oito textos originais em língua portuguesa, a respeito do item (11) Cidades e Comunidades Sustentáveis dos ODS, para a investigação em corpora, cujo objetivo deste trabalho é observar a versão de cinco termos multi-palavras, bem como discutir alguns dos traços de normalização de Scott (1998), por meio da versão de um texto previamente selecionado.

FUNDAMENTAÇÃO TEÓRICA

No que tange à pedagogia da tradução, podemos mencionar que as primeiras investigações neste campo foram a de Díaz Fouces (1999, 2001), nas quais “buscam criar metodologias de ensino que observem as competências da tradução”. Por meio desses métodos, “discentes devem estar aptos a codificar e sistematizar informações presentes nos textos” (DÍAZ FOUCES, 1999, 2001 apud SERPA et al (2021). Por seu turno, Laviosa (2008) comenta que “os corpora pequenos e especializados são feitos e usados não apenas como recursos de busca de equivalentes na tradução, mas também como repositórios de dados a fim de aperfeiçoar a compreensão dos discentes a respeito das regularidades da tradução” (LAVIOSA, 2008 apud YYY 2016b, p. 159). No trabalho intitulado “A Corpus-based Proposal for Teaching a Translational Habitus: Initial dialogues with Bourdieu’s sociological approaches”, Serpa et al (2021) apresentam atividades didáticas utilizando a aprendizagem movida por dados e a pedagogia da tradução. Em 2022, Thow publica o estudo “Translation Pedagogy in the Comparative Literature Classroom: Close Reading and the Hermeneutic Model of Translation” nos mostrando como podemos utilizar a pedagogia da tradução em uma sala de aula de literatura comparada, nos indicando meios para realizá-la.

Acerca dos ODS das Nações Unidas, sabemos da importância desse tema atualmente e por este motivo, além das investigações de Pinto (2021) e Pinto et al (2023) e do Seminário de Estudos Linguísticos da UNESP, realizado em setembro de 2023, com o tema “A Linguística face aos Objetivos de Desenvolvimento Sustentável da Agenda 2030”, decidimos por também trabalhá-lo em sala de aula, acreditando que seja de extrema relevância apresentar ao graduando uma variedade de temas que possam vir a enriquecer sua formação. Já, a aprendizagem movida por dados foi criada por Tim Johns e é, de acordo com Berber Sardinha,

> uma das propostas mais sólidas para a utilização de material de corpus na sala de aula. [...] A ênfase é desenvolver no aluno a habilidade de descoberta (discovery learning), e o papel do professor é propiciar meios para que os alunos adquiram estratégias de descoberta. O computador entra como elemento central da aprendizagem, no papel de informante, e não de substituto do professor (BERBER SARDINHA, 2004, p. 290- 291 – grifo nosso).

A fim de elaborar as listas de frequência dos termos multi-palavras, utilizamos a ferramenta on-line Sketch Engine, na qual se fizermos o upload de textos em línguas inglesa e portuguesa, ela nos fornecerá os termos multi-palavras. além de nos trazer linhas de concordância em seu contexto.

METODOLOGIA DE INVESTIGAÇÃO

Apresentamos, a seguir, a composição dos corpora, bem como os procedimentos e as formas de análise adotadas para o nosso estudo.

MATERIAL EMPREGADO NA COMPILAÇÃO DOS CORPORA

Em razão à limitação de espaço, o material empregado na compilação dos corpora será apresentado durante o evento.

PROCEDIMENTOS DE ANÁLISE

PASSOS PARA A ANÁLISE COM BASE NAS CONCORDÂNCIAS

Empregamos a ferramenta on-line Sketch Engine para os TOs em línguas inglesa e portuguesa. Após termos examinado as listas dos primeiros doze termos em línguas inglesa e portuguesa, foi compilado um glossário de línguas inglesa e portuguesa composto das similaridades e diferenças encontradas em quatro termos. Também, elaboramos um glossário menor constituído de um texto em língua portuguesa acerca do item (11) dos ODS e, por meio da versão em língua inglesa do referido texto realizada pelos discentes da disciplina, discutimos alguns dos traços de normalização de Scott (1998).

DISCUSSÃO E ANÁLISE DOS RESULTADOS

A seguir, apresentamos os dados extraídos do texto “Cidades portuguesas a caminho da sustentabilidade”, bem como de suas versões para a língua inglesa efetuadas por sete discentes da disciplina, além da discussão a respeito de alguns traços de normalização de Scott (1998), a saber: tamanho da sentença, reordenação de elementos, pontuação, padrão de repetição simples, mudança em palavras menos comuns, e adição no texto traduzido, .

Como podemos notar por meio do glossário menor constituído de cinco termos extraídos do referido texto, os termos multi-palavras smart city, capital verde e cidades inteligentes apresentaram apenas uma versão para a língua inglesa, respectivamente o próprio termo smart city, green capital e smart cities.

Referindo-se ao termo multi-palavra ecossistema da sociedade, verificamos que houve variações em suas versões para a língua inglesa realizada pelos alunos: Society’s ecosystem (1); ecosystem of Society (4); ecosystems of the Society (1); e societal ecosystem (1).

No tocante ao termo linha orientadora no país, percebemos que as versões para a língua inglesa também foram variadas pelos discentes: guideline in the country (5); guiding line in the country (1); e guiding direction in the country (1).

Devido à limitação de espaço, não será possível apresentar os dados com detalhes nesta seção, porém, durante a apresentação, estes serão efetivamente discutidos e analisados.

CONSIDERAÇÕES INICIAIS

Esperamos que por meio deste trabalho seja possível observar as semelhanças e diferenças nos TOs em línguas portuguesa e inglesa a respeito do léxico contido no item (11) Cidades e Comunidades Sustentáveis dos ODS, por meio da compilação de dois glossários de línguas inglesa e portuguesa.

Acreditamos que a pedagogia da tradução é uma abordagem que merece destaque, uma vez que trata da teoria e da prática da tradução, sendo muito relevante para os discentes, uma vez que por meio da teoria, podemos ampliar nosso conhecimento e transmiti-lo no exercício da versão para a língua inglesa.

REFERÊNCIAS

BAKER, M. Corpus-based translation studies: the challenges that lie ahead. In: SOMERS, H. (ed.). Terminology, LSP and translation studies in language engineering, in honour of Juan C. Sager. Amsterdam/Philadelphia: John Benjamins, 1996, p.175-186.

BERBER SARDINHA, T. Lingüística de Corpus. São Paulo: Manole, 2004.

BURNETT, S. A corpus-based study of translational English. Manchester: UMIST, 1999. Dissertação de mestrado.

YYY. Uso de Corpora para uma Pedagogia da Tradução. Revista Língua & Letras. V. 17:36, p. 188-205, 2016a.

YYY. Language of Translation and Interculturality for a Corpus-based Translation Pedagogy. In: FONTANILLE, J. (Org.). Traduire: signes, textes, pratiques. Liège: Presses Universitaires de Liège, p. 155-173, 2016b.

DÍAZ FOUCES, Ó. Didáctica de la traducción (português – español), Vigo: Servicio de Publicacións da Universidade de Vigo, 1999.

DÍAZ FOUCES, Ó. Sociologia de la traducción, Quaderns: Revista de traducció v. 6, p. 63-77, 2001.

JOHNS, T. MicroConcord: a language learner's research too I. System, Oxford, Pergamon, v.14, n.2, jun/ 1986, p.l5l-62.

LAVIOSA, S. Discovery and Justification Procedures in the Corpus-Based Translation Classroom. Translation Challenges: From Training to Profession, Hammamet, Tunisia, 28-29 November 2008.

LAVIOSA, S. A transcultural conceptual framework for corpus-based translation pedagogy, 2010 In: Proceedings of Using corpora in contrastive and translation studies - UCCTS. Ormskirk, 2010. v. 01.

LAVIOSA, S. Translation and Language Education: Pedagogic Approaches Explored. New York: Routledge/Taylor & Francis, 2014.

LAVIOSA, S. The Instrumental and Hermeneutic Models of Translation in Higher Education. In: Engel,
N., Köngeter, S. (eds) Übersetzung. Springer VS, Wiesbaden, 2020. https://doi.org/10.1007/978-3-658-20321-4_3

SERPA, T.; PINTO, P.T.; YYY. A Corpus-based Proposal for Teaching a Translational Habitus: Initial dialogues with Bourdieu's sociological approaches. Trans. Revista de Traductología. V. 25, p. 507-525, 2021.

SCOTT, M. N. Normalisation and Reader's Expectation: A Study of Literary Translation with Reference to Lispector's A Hora da Estrela. Tese (Doutorado em Filosofia). Liverpool: Universidade de Liverpool, 1998.

THOW, E. Translation Pedagogy in the Comparative Literature Classroom: Close Reading and the
Hermeneutic Model of Translation. L2 Journal, V. 14 (2), p. 91-106, 2022. DOI: 10.5070/L214252048

O C-ORAL-ESQ, corpus brasileiro de fala espontânea de pessoas com esquizofrenia

Bruno Nevis Rati de Melo ROCHA
Tommaso RASO

Introdução

O trabalho visa apresentar (a) o C-ORAL-ESQ (RASO et al., 2023), Corpus Oral de Esquizofrênicos, (b) algumas medidas descritivas da atual fase de compilação do corpus e (c) perspectivas futuras do corpus. O C-ORAL-ESQ, em estágio avançado de compilação, contará com 43 registros de interações entre psiquiatras e pacientes com esquizofrenia durante consultas médicas. O corpus será transcrito e segmentado segundo os critérios estabelecidos para o corpus C-ORAL-BRASIL (RASO; MELLO, 2012) e alinhado no Elan (WITTENBURG et al., 2006). Grande parte das gravações será etiquetada informacionalmente segundo os preceitos da Language into Act Theory (CRESTI, 2000). Por fim, o C-ORAL-ESQ contará com uma seção multimodal composta de 18 gravações em áudio e vídeo que seguirá os parâmetros estabelecidos para o C-ORAL-BGEST, apresentados por Barros (2021), permitindo estudos de expressões faciais e gestos.

A Language into Act Theory

A L-AcT é uma teoria que tem como objetivo principal descrever a estrutura informacional da fala espontânea. Em particular, a teoria se interessa por explicar a maneira pela qual o falante codifica prosodicamente as diferentes funções que as unidades tonais do enunciado podem assumir, partindo do entendimento de que todo enunciado é um ato de fala (AUSTIN, 1962). Os enunciados são entendidos como a menor unidade superior ao nível da palavra que possui autonomia prosódica e interpretabilidade pragmática (ou seja, é percebido como prosodicamente concluído e veicula uma ilocução). Um enunciado pode ser formado por uma ou mais unidades tonais, as quais desempenham funções diversas. A função básica, expressa por ao menos uma unidade em todo e qualquer enunciado, é justamente a de realizar uma ilocução. Assim, em enunciados compostos por uma única unidade tonal, essa é necessariamente sua unidade ilocucionária. Além da função ilocucionária, existem outros tipos de funções que uma unidade pode desempenhar, como aquela de estabelecer um domínio cognitivo para a ilocução (cf. CRESTI, 2000 e RASO; MONEGLIA, 2014 para uma descrição detalhada de todas as funções).

Metodologia

O setting das gravações

As gravações do C-ORAL-ESQ são realizadas em instituições hospitalares públicas de Belo Horizonte. De 2019 a 2023, foram feitas gravações em áudio de consultas ocorridas no ambulatório do Instituto Raul Soares (IRS/FHEMIG). A partir de 2023, começaram a ser feitas gravações em áudio e vídeo seja no IRS, seja no Hospital das Clínicas da UFMG (HC/EBSERH).

As gravações são conduzidas por uma equipe de dois pesquisadores que se dirigem ao hospital com os equipamentos (dois conjuntos de microfones sem fio omnidirecionais Sennheiser EK100/SK100, um gravador digital de alta resolução Tascam DR-100 e duas câmeras GoPro Hero7). A equipe se encontra com o residente responsável pela consulta a ser gravada, o qual conversou antecipadamente com o paciente que atenderá naquele dia sobre a possibilidade de ser gravado. Em seguida, a equipe coloca os microfones no residente e no paciente e posiciona as câmeras de vídeo. As duas câmeras são usadas para gravar imagens do paciente, sendo uma posicionada frontalmente a ele e outra posicionada de forma levemente lateral, para permitir observar os gestos com maior profundidade. Antes de sair do consultório, a equipe lê os TCLEs do médico, do paciente e de eventuais acompanhantes e recolhe as assinaturas necessárias. Fora do consultório, a equipe conecta os microfones sem fio ao gravador e inicia a gravação. Durante toda a consulta, a equipe permanece fora do consultório e não interage direta ou indiretamente com os participantes.

Esse desenho metodológico foi feito para potencializar a naturalidade da situação, com o objetivo de garantir a validade ecológica dos dados.

As etapas de tratamento dos dados

Todas as gravações d o C-ORAL-ESQ passam por uma série de etapas obrigatórias de tratamento dos dados: (a) transcrição e segmentação prosódica segundo os critérios adotados pelo C-ORAL-BRASIL (RASO; MELLO, 2012), (b) primeira e segunda revisões da transcrição e da segmentação, (c) alinhamento texto-som-espectrograma no *software* Elan e (d) revisão final da transcrição. As consultas que possuem gravações em vídeo possuem procedimentos adicionais, sendo eles (a) a anonimização da face do paciente e (b) a inserção dos vídeos no alinhamento do Elan. A anonimização é feita por meio de um procedimento computadorizado que captura pontos do rosto importantes para a veiculação das expressões faciais e os reproduz em um rosto criado por inteligência artificial. Essa versão anonimizada será disponibilizada com o corpus e poderá ser usada em apresentações e publicações científicas sem permitir que se descubra a identidade do participante

Por fim, um conjunto de gravações do C-ORAL-ESQ será selecionado para passar pelo procedimento de etiquetagem informacional segundo a L-AcT.

O corpus

Estado atual

O C-ORAL-ESQ irá registrar 43 consultas entre pacientes com esquizofrenia e seus psiquiatras, com uma média de 1.180 palavras produzidas pelo paciente em cada gravação (além das palavras dos médicos e de eventuais acompanhantes dos pacientes).

A Tabela 1 mostra o número de palavras produzidas por cada grupo de participantes (pacientes, profissionais da saúde, acompanhantes e outros). Nela, lê-se que as gravações têm, em média, 2.360 palavras, sendo que a menor possui 749 palavras e a mais extensa, 4.868, totalizando 101.606 palavras. Os

pacientes produzem em média 1.180 palavras por gravação, mas com uma variação muito grande (DP 810), indo de 236 palavras a 2.717.

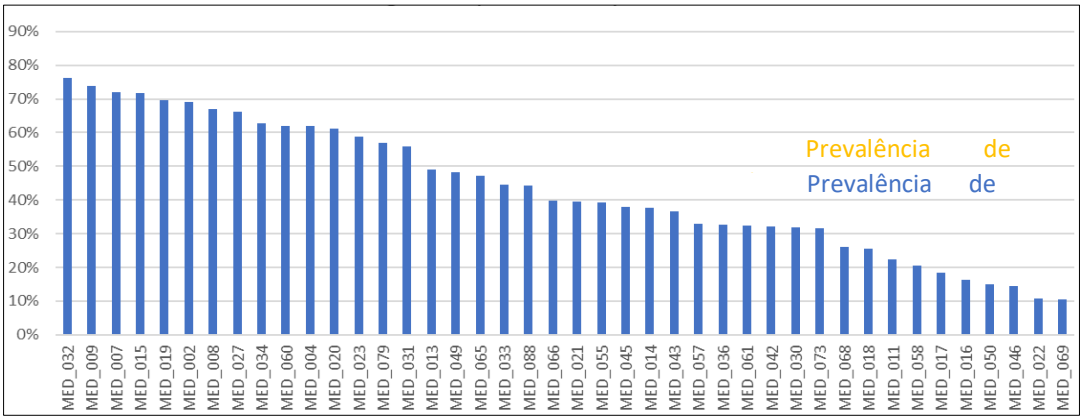
Tabela 1 – Medidas descritivas do corpus

		Palavras	Média	DP	Mediana	Mínimo	Máximo
Pacientes		48.086	1.118	810	836	236	2717
	Médicos	46.522	1.082	475	974	260	2157
Acompanhantes		6.642	332	300	265	10	801
	Outros	3561	57	67	17	2	172
	Total	101.606	2.360	1.031	2.237	749	4.868

Fonte: os autores.

O alto desvio padrão no número de palavras produzidas por todos os grupos de participantes, em especial o de pacientes, é o reflexo de características primordiais das consultas: a variação de tipos e de número de assuntos a serem tratados no dia, a disponibilidade do paciente para interagir com o psiquiatra, a estratégia adotada pelo psiquiatra para lidar com o paciente no dia, a presença ou ausência de acompanhantes e o grau de participação deles nas consultas etc. Com efeito, o alto desvio padrão no número de palavras dos participantes é tanto uma característica esperada quanto desejada em um corpus que busca documentar a fala espontânea produzida em interações médico-paciente não roteirizadas.

Figura 1 – Percentual de palavras de pacientes com relação às palavras de outros participantes

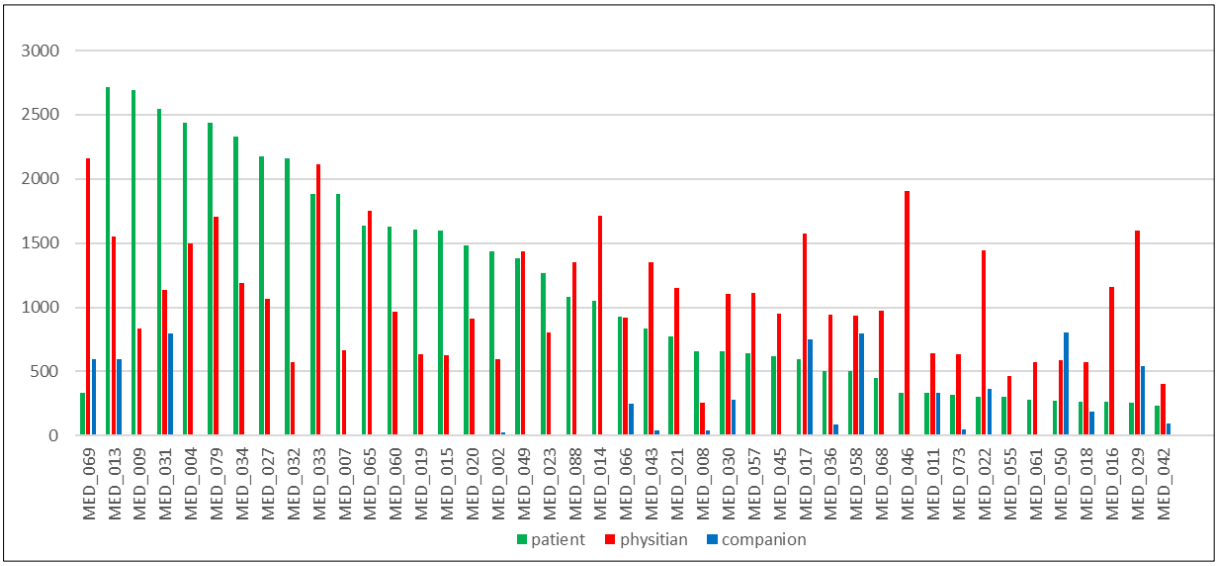


Fonte: Elaboração própria.

A figura 1 mostra o percentual de palavras de pacientes com relação às palavras de outros participantes (psiquiatras, psicólogos, acompanhantes e demais pessoas) em cada gravação. As 15 primeiras colunas, cujo limite superior vai além da linha horizontal tracejada, identificam gravações em que a fala do paciente é predominante na consulta, chegando a ocupar 79% de uma gravação no caso mais extremo. As demais são casos em que o paciente produz um número de palavras inferior ao dos demais participantes, chegando ao caso de uma consulta em que o paciente produz aproximadamente 11% do total de palavras. É interessante notar que existe uma grande variação com relação a esse parâmetro, podendo ser observada uma tendência aproximadamente linear na distribuição dos valores percentuais de palavras dos pacientes com relação às palavras dos outros participantes.

A figura 2, por sua vez, indica o número de palavras por participante em cada gravação (destacando palavras de pacientes em verde, de médicos em vermelho e de acompanhantes em azul). Vale notar que, nas gravações com o menor número de palavras de pacientes (últimas colunas à direita), é maior a quantidade de consultas em que os acompanhantes produzem uma quantidade expressiva de palavras.

Figura 1 – Número de palavras de pacientes (verde), médicos (vermelho) e acompanhantes (azul)



Fonte: Elaboração própria.

Perspectivas futuras

Paralelamente à compilação do C-ORAL-ESQ, está sendo elaborado um corpus de controle específico para esse corpus. O corpus de controle possuirá a mesma quantidade de gravações do C-ORAL-ESQ e será balanceado em gênero, idade e escolaridade. Suas gravações serão multimodais (áudio e vídeo) e registrarão consultas médicas de pacientes com doenças crônicas e sem histórico pessoal e familiar de transtornos mentais. As gravações do corpus de controle passarão pelos mesmos procedimentos metodológicos aplicados ao C-ORAL-ESQ.

Conclusões

O corpus C-ORAL-ESQ fornecerá um material de grande ineditismo não somente no panorama nacional, mas também internacional. Grande parte das pesquisas sobre a fala de pessoas com esquizofrenia é feita a partir do exame de fala eliciada. O C-ORAL-ESQ, por outro lado, apresenta um grande conjunto de dados de fala espontânea obtidos em interações reais, não roteirizadas. O corpus já tem sido usado em uma série de pesquisas preliminares que tem permitido observar com mais atenção a estrutura informacional da fala dos pacientes (COSTA JR., 2022) e pode ser usado tanto para pesquisas de cunho linguístico como médico.

Referências bibliográficas

AUSTIN, J. L. *How to do things with words*. Oxford University Press, Oxford 1962.

BARROS, C. A. *A relação entre unidades gestuais e quebras prosódicas: o caso da unidade informacional Parentético*. Dissertação (Mestrado em Estudos Linguísticos), Universidade Federal de Minas Gerais, 2021.

CAVALCANTE, F. A., *The topic unit in spontaneous american English: a corpus-based study*. Dissertação (Mestrado em Estudos Linguísticos), Universidade Federal de Minas Gerais, 2016.

COSTA JR. J. C., *Padrão informacional de stanzas de pacientes com esquizofrenia*. Tese (Doutorado em Linguística), Universidade Federal de Minas Gerais, 2022.

CRESTI, E., *Corpus di Italiano parlato*, Accademia della Crusca, Firenze 2000.

MONEGLIA, M.; RASO, T. Notes on language into act theory. In: RASO, T.; MELLO, H. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins, 2014, 468-495.

RASO, T.; MELLO, H. *C-ORAL-BRASIL I*. Corpus de referência do português falado informal. Belo Horizonte: UFMG, 2012.

WITTENBURG, P.; BRUGMAN, H.; RUSSEL, A.; KLASSMANN, A.; SLOETJES, H. ELAN: a Professional Framework for Multimodality Research. Em: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation, 2006.

Agradecimentos

Nossa sincera gratidão à presença, contribuição e empenho de todes.

Foi o esforço colaborativo que permitiu a realização deste evento.

Comissão Organizadora