

**STIL 2024**

15th Symposium in Information and Human  
Language Technology

Proceedings of the Conference, Vol. 1

November 18, 2024

## About the workshop

The Proceedings of the XV Brazilian Symposium on Information Technology and Human Language (STIL 2024) present the selected papers presented at the event held from September 18 to 21, 2024, in the city of Belém, PA, in conjunction with the XXXIV Brazilian Conference on Intelligent Systems (BRACIS 2024), XXI National Meeting on Artificial and Computational Intelligence (ENIAC 2024), and XII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2024). In this edition, the proceedings compile the selected papers for the Main Conference and the works accepted in its satellite events:

- XV Brazilian Symposium on Information Technology and Human Language (STIL 2024) received a total of 99 paper submissions, consisting of 78 full papers and 21 short papers. Among the full papers, 24 were accepted, resulting in an acceptance rate of 30%. As for the short papers, 13 were accepted, corresponding to an acceptance rate of 62%. These works were selected through a double-blind peer review process;
- IX Portuguese Description Conference (JDP 2024) received a total of 13 article submissions, of which 10 were accepted, resulting in an acceptance rate of 77%. These works were selected through a double-blind peer review process;
- I Workshop on Data Enrichment in Portuguese (PaDAWan 2024) received a total of 12 article submissions, of which 6 were accepted, resulting in an acceptance rate of 50%. These works were selected through a double-blind peer review process.

## **Acknowledgments**

The Program Committee chairs acknowledge the financial support to this conference provided by the Brazilian Computer Society (SBC). We thank the Program Committees of the XIV Brazilian Symposium in Information and Human Language Technology and Collocated Events for their reviews. Last but not least, we are grateful to the local organization committee led by Carlos Renato Lisboa Francês (UFPA), Evelin Helena Silva Cardoso (UFPA), Jose Jailton Henrique Ferreira Junior (UFPA), Hugo Pereira Kuribayashi (UNIFESSPA) and Jorge Antonio Moraes de Souza (UFRA).

November 2024

Adriana Pagano (UFMG, Brazil)  
Daniela Barreiro Claro (UFBA, Brazil)

## Program chairs

### - STIL

Adriana Pagano (UFMG, Brazil)  
Daniela Barreiro Claro (UFBA, Brazil)

### - JDP

Raquel Meister Ko Freitag (UFS, Brazil)  
Rerisson C. de Araújo (UFBA, Brazil)

### - TILIC

Eloize R. Marques Seno (IFSP, Brazil)  
Marcio Inácio (Univer. Coimbra, Portugal)

### - PaDaWan

Livy Real (CE-PLN/SBC)  
Evandro Fonseca (Blip/PUCRS)  
Paula Cardoso (UFPA)

## Program Committee

Aline Paes - Universidade Federal Fluminense  
Ariani Di Felippo - Universidade Federal de São Carlos  
Arnaldo Candido Junior - Universidade Estadual Paulista  
Carlos Ferreira - Universidade Federal de Ouro Preto  
Cássio Faria da Silva - Rede Gonzaga de Ensino Superior – REGES  
Christopher Shulby - Universidade de São Paulo  
Clarissa Xavier - SiDi  
Cláudia Dias de Barros - Inst. Federal de Edu., Ciência e Tecnologia de São Paulo  
Cláudia Freitas - Universidade de São Paulo  
Diana Santos - Linguateca/Universidade de Oslo  
Diego Furtado Silva - Universidade de São Paulo  
Eduardo Gonçalves - Escola Nacional de Ciências Estatísticas (ENCE/IBGE)  
Eduardo Luz - Universidade Federal de Ouro Preto  
Elisa Marchioro Stumpf - Universidade Federal do Rio Grande do Sul  
Eloize Seno - Instituto Federal de Educação, Ciência e Tecnologia de São Paulo  
Ely Edison Matos - Universidade Federal de Juiz de Fora  
Evandro Ruiz - Universidade de São Paulo  
Evelin Amorim - INESC TEC  
Gabriela Wick-Pedro - Instituto Brasileiro de Informação em Ciência e Tecnologia  
Helen de Andrade Abreu - Universidade Federal de Juiz de Fora  
Helena Caseli - Universidade Federal de São Carlos  
Heliana Mello - Universidade Federal de Minas Gerais  
Hilário Tomaz de Oliveira - Instituto Federal do Espírito Santo  
Jackson Souza - Universidade Federal da Bahia  
Jorge Baptista - Universidade do Algarve  
Juliano Antonio - Universidade Estadual de Maringá  
Lívia Ruback - Universidade Estadual de Campinas  
Leandro H. M. de Oliveira. - Empresa Bras. de Pesquisa Agropecuária (EMBRAPA)  
Livy Real - B2W Digital/GLiC  
Lucelene Lopes - Universidade de São Paulo  
Magali Duran - Universidade de São Paulo  
Marcelo Finger - Universidade de São Paulo  
Maria das Graças V. Nunes - Universidade de São Paulo, São Carlos.

Maria José B. Finatto - Universidade Federal do Rio Grande do Sul  
Marlo Souza - Universidade Federal da Bahia – UFBA  
Maucha Andrade Gamonal - Universidade Federal de Minas Gerais  
Márcio de Souza Dias - Universidade Federal de Catalão  
Nádia Silva - Universidade Federal de Goiás  
Norton Roman - Universidade de São Paulo  
Oto Vale - Universidade Federal de São Carlos  
Pablo Gamallo - University of Santiago de Compostela  
Paula Figueira Cardoso - Universidade Federal do Pará  
Rafael Anchiêta - Instituto Federal de Educação, Ciência e Tecnologia do Piauí – IFPI  
Renata Vieira - Universidade de Évora  
Renato Moraes Silva - Universidade de São Paulo  
Ricardo Marcacini - Universidade de São Paulo  
Rodrigo Wilkens - University of Exeter  
Roney Santos - Universidade Federal da Bahia  
Sandra Avila - Universidade Estadual de Campinas  
Sergio Antonio A. Freitas - Universidade de Brasília  
Thiago A. Salgueiro Pardo - Universidade de São Paulo  
Tiago Timponi Torrent - Universidade Federal de Juiz de Fora  
Valéria Feltrim - Universidade Estadual de Maringá  
Valeria de Paiva - Topos Institute USA

### **Additional Reviewers**

Aline Ioste - Universidade de São Paulo  
Andre Coneglian - Federal University of Minas Gerais  
Annie Amorim - Universidade Federal Fluminense  
Arthur Scalercio - Universidade Federal Fluminense  
Babacar Mane - Universidade Federal da Bahia  
Bruno Cabral - Universidade Federal da Bahia  
Eduardo Gonçalves - Escola Nacional de Ciências Estatísticas (ENCE/IBGE)  
Elisa Marchioro Stumpf - Universidade Federal do Rio Grande do Sul  
Felipe Serras - Universidade de São Paulo  
Fernando Moraes - Universidade Federal da Bahia  
Izabela Müller - Universidade do Algarve / INESC-ID Lisboa  
Jonnathan Carvalho - Instituto Federal Fluminense – RJ  
Larissa D. Xavier da Silva - Universidade Federal da Paraíba  
Laura A. Costa Ribeiro - Universidade Federal Fluminense  
Lilian Teixeira de Sousa - Universidade Federal da Bahia  
Marcos Treviso - Universidade de São Paulo  
Reginaldo Santos - Universidade Federal do Pará  
Rogerio Sousa - Instituto Federal do Piauí  
Sidney Leal - Universidade de São Paulo

# Contents

## STIL

### **A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos**

*Paula Cardoso, Jackson Souza, Roana Rodrigues, Ewerson Dantas, Larissa Santa Bárbara, Mateus Araújo, Naira Gama, Tobias Almeida, Gabriel Cruz*.....1-10

### **Synthetic AI Data Pipeline for Domain-Specific Speech-to-Text Solutions**

*Anderson Luiz Karl, Guilherme Sales Fernandes, Leonardo Augusto Pires, Yvens R. Serpa, Carlos Caminha*.....11-21

### **Automatic Annotation of Enhanced Universal Dependencies for Brazilian Portuguese**

*Elvis A. de Souza, Magali S. Duran, Maria das Graças V. Nunes, Gustavo Sampaio, Giovanna Belasco, Thiago A. S. Pardo*.....22-31

### **Biases in GPT-3.5 Turbo model: a case study regarding gender and language**

*Fernanda Malheiros Assi, Helena de Medeiros Caseli*.....32-43

### **Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português?**

*Gabriel Assis, Arthur Vasconcelos, Lívia de Azevedo, Mariza Ferro, Aline Paes*.....44-54

### **TableRAG: A Novel Approach for Augmenting LLMs with Information from Retrieved Tables**

*Elvis A. de Souza, Patricia F. da Silva, Diogo Gomes, Vitor Batista, Evelyn Batista, Marco Pacheco*.....55-64

### **PropBank e anotação de papéis semânticos para a língua portuguesa: O que há de novo?**

*Cláudia Freitas, Thiago Alexandre Salgueiro Pardo*.....65-75

### **LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis**

*Bryan K. S. Barbosa, Claudio E. C. Campelo*.....76-85

### **Evaluating Federated Learning with Homomorphic Encryption for Medical Named Entity Recognition Using Compact BERT Models**

*Marcos F. Pontes, Rodrigo C. Pedrosa, Pedro H. Lopes, Eduardo J. Luz*.....86-94

### **A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach**

*Ariani Di Felippo, Maria das Graças V. Nunes, Bryan K. da Silva Barbosa*.....95-104

### **Geração Automática de Perguntas em Português do Brasil Usando os Modelos PTT5 e FLAN-T5**

*Tiago Felipe V. Braga, Bruno Cardoso Coutinho, Hilário Tomaz Alves de Oliveira*.....105-114

### **Sumarização Automática de Artigos de Notícias em Português: Da Extração à Abstração com Abordagens Clássicas e Modelos de Neurais**

*Marcio Alves Sarmento, Hilário Tomaz Alves de Oliveira*.....115-124

**Boosting not so Large Language Models by using Knowledge Graphs and Reinforcement Learning**

*William Jones Beckhauser, Renato Fileto*.....125-135

**Genipapo – a Multigenre Dependency Parser for Brazilian Portuguese**

*Ariani Di Felippo, Norton T. Roman, Bryan K. S. Barbosa, Thiago A. S. Pardo*..... 136-145

**Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?**

*Amanda S. Oliveira, Pedro H. L. Silva, Valéria de C. Santos, Gladston Moreira, Vander L. S. Freitas, Eduardo J. S. Luz*..... 146-155

**Disfluency Detection and Removal in Speech Transcriptions via Large Language Models**

*Pedro L. S. de Lima , Cláudio E. C. Campelo*..... 156-164

**Detection and Censorship of Offensive Language in Extended Texts in Portuguese**

*Lucas Lenocho de Souza, Franciele Beal, André Roberto Ortoncelli, Marlon Marcon*..... 165-174

**EyetrackingMOS: Proposta de um método de avaliação online para modelos de síntese de fala**

*Gustavo E. Araújo, Julio C. Galdino, Rodrigo de F. Lima, Leonardo Ishida, Gustavo W. Lopes, Miguel Oliveira Jr., Arnaldo Candido Jr., Sandra M. Aluísio, Moacir A. Ponti*..... 175-184

**Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers**

*Mirelle Bueno , E. Seiti de Oliveira , Rodrigo Nogueira , Roberto Lotufo, Jayr Pereira*..... 185-195

**Mineração de Argumentos em Textos de Redes Sociais no Idioma Português**

*Vitor Domingos Balduino dos Santos, Livia Alabarse dos Santos, Orlando B. Coelho (in memoriam), Renata Mendes de Araujo, Ivan Carlos Alcântara de Oliveira*..... 196-206

**A Hybrid Machine Learning Method to Author Name Disambiguation**

*Natan S. Rodrigues, Célia G. Ralha*.....207-216

**Adapting LLMs to New Domains: A Comparative Study of Fine-Tuning and RAG strategies for Portuguese QA Tasks**

*Leandro Yamachita da Costa, João Baptista de Oliveira e Souza Filho*.....217-227

**A Change in Perspective: The Trade-Off Between Perspective API and Custom Models in Classifying Hate Speech in Portuguese**

*Arthur Buzelin, Yan Aquino, Pedro Bento, Samira Malaquias, Wagner Meira Jr, Gisele L. Pappa*.....228-236

**No Argument Left Behind: Overlapping Chunks for Faster Processing of Arbitrarily Long Legal Texts**

*Israel Fama, Bárbara Bueno, Alexandre Alcoforado, Thomas Palmeira Ferraz, Arnold Moya, Anna Helena Reali Costa*.....237-246

**Syntactic parsing: where are we going?**

*Lucelene Lopes, Thiago Alexandre Salgueiro Pardo, Magali S. Duran*.....247-254

**Segmentação Textual Baseada em Tópicos em Português Utilizando BERTimbau**

*Luciano A. C. da Silva, Maiara S. F. Rodrigues, Adriana P. Archanjo, Luis Pessoa, Miguel L. Silva, Thiago F. de Almeida, Leonardo Silveira,*..... 255-259

**Avaliação de modelos para detecção de ataques de replay usando diferentes bases de dados**

*Giovana Y. Nakashima, Higor D. C. Santos, Jone W. M. Soares, Mário Uliani Neto, Fernando O. Runstein, Ricardo P. V. Violato, Marcus Lima*.....260-265

**Avaliação de arquiteturas de síntese de fala generativa com abordagens de espectrograma e fim-a-fim em cenários low-resource para clonagem de voz**

*Bruno C. dos S. Ribeiro, Gustavo H. dos S. Figueiredo, Leonardo H. da S. Correia, Mário Uliani Neto, Fernando O. Runstein, Ricardo P. V. Violato, Marcus Lima* .....266-270

**Leveraging Structured Data Input for Effective Chatbot Integration in Enterprises**

*Caio Siqueira, Orlando Guilarte, Giuliano Ferreira, Omar Leiva*.....271-275

**Anomaly Detection in Text Data: A Semi-Supervised Approach Applied to the Portuguese Domain**

*Fabio Masaracchia Maia, Anna Helena Reali Costa*.....276-281

**Identificação de aspectos explícitos e implícitos em críticas gastronômicas em português: avaliando o potencial dos LLMs**

*Luiz H. N. Silva, Eloize R. M. Seno, Rozane R. Rebechi, Helena M. Caseli Fabiano M. Rocha Júnior, Guilherme A. Faller*.....282-287

**Avaliação de Algoritmos de Clusterização para Agrupamento de Descrições de Produtos em Notas Fiscais Eletrônicas**

*Jonas Gabriel L. de Araújo, Thaís G. do Rêgo, Yuri de A. M. Barbosa*.....288-293

**Unified Knowledge-Graph for Brazilian Indigenous Languages: An Educational Applications Perspective**

*Gustavo Polleti, Fabio Cozman, Fabricio Gerardi*.....294-299

**A Robustness Analysis of Automated Essay Scoring Methods**

*Rafael T. Anchiêta, Rogério F. de Sousa, Raimundo S. Moura*.....300-305

**Classificação de Notícias em Português Utilizando Modelos Baseados em Transferência de Aprendizagem e Transformers**

*Wagner Narde, João Mendanha, Henrique Barbosa, Frederico Coelho, Bruno Santos, Luiz Torres*.....306-310

**Beyond Single Models: Leveraging LLM Ensembles for Human Value Detection in Text**

*Diego Dimer Rodrigues, Mariana Recamonde-Mendoza, Viviane P. Moreira*.....311-316

**JORNADA DE DESCRIÇÃO DO PORTUGUÊS**

**Performance in a dialectal profiling task of LLMs for varieties of Brazilian Portuguese**

*Raquel M. Ko Freitag, Túlio Sousa de Gois* .....317-326

**Mini-glossário do Tucumã do Pará no Município de Acará: olhares, significados e cultura da Amazônia**

*Eliene da S. Alves, Brayna C. dos S. Cardoso*.....327-335

**Desambiguação de lema e atributos morfológicos na anotação do corpus Porttinari-base**

*Lucelene Lopes, Magali S. Duran, Thiago Alexandre Salgueiro Pardo*.....336-345

**An NLP approach to impersonal –se in Brazilian Portuguese**

*Elvis A. de Souza, Magali S. Duran, Adriana S. Pagano* .....346-355

**Inferências baseadas em sintaxe: a anotação de sujeitos implícitos**

*Magali Sanches Duran, Maria das Graças Volpe Nunes, Thiago Pardo*.....356-364

**Notes on variation and lexical diachrony in the Parish Memories-Alentejo collection (1758)**

*Helena Freire Cameron, Fernanda Olival, Renata Vieira*..... 365-374

**A sílaba e a composicionalidade em emakhuwa (P31): análise de empréstimos do português**

*Francelino Wilson, Vasco Magona, Felermino Ali* ..... 375-385

**Linguistic and emotional dynamics in satirical vs. real news: a psycholinguistic analysis**

*Gabriela Wick-Pedro, Roney Lira de Sales Santos, Oto Araújo Vale*..... 386-392

**Complementos de eco de adjetivos com completiva-sujeito em português do Brasil**

*Ryan Saldanha Martinez, Jorge Baptista, Oto Vale*..... 393-402

**Modelagem baseada em frames para identificação do léxico da Violência de Gênero**

*Lorena Tasca Larré, Tiago Timponi Torrent*..... 403-412

**WORKSHOP DE IC EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA**

**Relações Retóricas de List e Sequence em textos jornalísticos**

*Tobias J. A. Almeida, Patrícia V. Almeida, Paula C.F. Cardoso* ..... 413-417

**Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários**

*Naira Silva Gama, Jackson Wilke da Cruz Souza* ..... 418-423

**Relações de coerência do português brasileiro: um estudo bibliográfico-documental da RST e seus sinalizadores discursivos**

*Larissa Jesus Santa Bárbara, Roana Rodrigues, Jackson Wilke da Cruz Souza*.....424-428

**Subsídios Linguísticos para classificação automática de textos de User-Generated Content**

*Mateus Araújo Pereira, Jackson Wilke da Cruz Souza*.....429-433

**Relações de coerência do espanhol peninsular: Um estudo bibliográfico-documental da Rhetorical Structure Theory**

*Ewerson Dantas, Roana Rodrigues Jackson Wilke da Cruz Souza*.....434-439

**Estratégias automáticas para análise da concordância da anotação de Sinalizadores Discursivos**

*Gabriel Sizinio Bomfim Cruz, Jackson W. C. Souza, Paula C. F. Cardoso.....440-444*

**Mineração de Emoções Multirrótulo Em Textos Curtos**

*Ramon N. Mendes, Syanne. K. M. Tavares, Luiz Nicollas M. Campos, Fabíola P. O. Araújo.....445-450*

**Classificação automática de textos de User-Generated Content utilizando Aprendizagem de Máquina Supervisionado**

*Iolanda Victoria Morais Ramos, Jackson Wilke da Cruz Souza.....451-456*

**PLN e Segurança Jurídica Identificação de divergências jurisprudenciais com Processamento de Linguagem Natural**

*Marcella Queiroz de Castro; Ana Régia Mendonça.....457-462*

**Um Pipeline de Pré-Processamento de Dados Textuais em Português para Análise de Redes Sociais**

*Livia A. dos Santos, Orlando B. Coelho (in memoriam)<sup>1</sup>, Renata Araujo, Ivan Carlos A. Oliveira.....463-468*

**Especulação Mística. Uma abordagem de Clusterização e Busca Semântica na aproximação de preço em cartas de Magic: The Gathering**

*Rodrigo Marques Duarte, André de Lima Salgado, Paula Figueira Cardoso.....469-473*

**Comparação de Ferramentas para Análise de Sentimentos Aplicada no Contexto Educacional**

*Benjamin G. Moreira, Luiz C. Camargo, Ricardo J. Pfitscher, Tatiana R. Garcia.....474-478*

**Modelo de Linguagem Quantizados na Área da Saúde: Um Enfoque em Perguntas e Respostas com Base na Técnica DPO**

*Mário Pinto Freitas Filho, João Dallyson Sousa de Almeida, Anselmo C. Paiva.....479-483*

**PORTUGUESE DATA AUGMENTATION WORKSHOP**

**LLM-SEMREL: Towards a Better Coreference Resolution for Portuguese**

*Evandro Fonseca, Joaquim Neto..... 484-492*

**Automated Topic Annotation in Brazilian Product Reviews: A Case Study of Adversarial Examples with Sabia-3**

*Lucas Nildaimon dos Santos Silva, Livy Real..... 493-501*

**Text extraction from Knowledge Graphs in the Oil and Gas Industry**

*Laura P. Navarro, Elvis A. de Souza, Marco A. C. Pacheco ..... 502-507*

**Getting Logic From LLMs Annotating Natural Language Inference with Sabiá**

*Fabiana Avais, Marcos Carreira, Livy Real ..... 508-517*

**Augmenting Data to Improve the Performance of Recommender Systems**

*Leticia Freire de Figueiredo, Joel Pinho Lucas, Aline Paes ..... 518-521*

**Brazilian Consumer Protection Code: a methodology for a dataset to Question-Answer (QA) Models**

*Aline Athaydes, Lucas Bulcao, Caio Sacramento, Babacar Mane, Daniela Barreiro Claro, Marlo Souza, Robespierre Pita..... 522-529*