

Evandro Eduardo Seron Ruiz
Tiago Timponi Torrent

STIL 2021

**XIII Brazilian Symposium in Information
and Human Language Technology and
Collocated Events**

Proceedings of the Conference

Online event from
November 29th to December 3rd, 2021

© 2021 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. Republication of material from this volume requires permission by the copyright owners.

Editors' addresses:

Departamento de Computação e Matemática
Programa de Pós-Graduação em Computação Aplicada
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Universidade de São Paulo
evandro@usp.br

Departamento de Letras
Programa de Pós-Graduação em Linguística
Faculdade de Letras
Universidade Federal de Juiz de Fora
tiago.torrent@ufjf.br

XIII Brazilian Symposium in Information and Human Language Technology

STIL is the bi-annual Language Technology event supported by the Brazilian Computer Society (SBC) and by the Brazilian Special Interest Group on Natural Language Processing (CE-PLN).

In 2021, STIL will be held as an online event collocated with BRACIS 2021 (The 10th Brazilian Conference on Intelligent Systems), ENIAC 2021 (The 18th National Meeting on Artificial and Computational Intelligence), and KDD-BR2021 (the 5th Brazilian Competition on Knowledge Discovery In Databases).

STIL will feature the following collocated events:

- VII Workshop on Portuguese Description (JDP); the
- VII Student Workshop on Information and Human Language Technology (TILic); and the
- OpenCor 2021 – Latin American and Iberian Languages Open Corpora Forum (OpenCor).

The conference is multidisciplinary and covers a broad spectrum of disciplines related to Human Language Technology, such as Linguistics, Computer Science, Psycholinguistics, and Information Science. It aims at bringing together both academic and industrial participants working on those areas.

We received 45 submissions for the main conference. Each paper was reviewed by at least two members of the Program Committee, which had 58 members from 8 countries and various institutions. After a rigorous reviewing process, 31 papers were selected for oral presentation. In JDP we had a total of 16 submissions; out of each 15 were accepted. Each paper was reviewed by at least two program committee members featuring 15 reviewers from public and private institutions from Brazil, Portugal, and France. For TILic, 5 out of 10 submissions were accepted after reviewing by at least two separate reviewers. TILic's program committee was composed of 20 researchers from Brazilian and international institutions. We thank the authors for their submissions, the program committee for their hard work, invited speakers, SBC staff, and the JDP, TILic, OpenCor, and BRACIS chairs.

November, 2021

Evandro Eduardo Seron Ruiz
Tiago Timponi Torrent

Acknowledgments

The STIL chairs acknowledge the financial support to the conference provided by the Brazilian Computer Society (SBC) and by the sponsors: Americanas S.A., Banco Itaú, NVidia, Ambev Tech, Google and Loggi. We thank the Program Committees of the XIII Brazilian Symposium in Information and Human Language Technology and Collocated Events for their reviews. We offer our sincere thanks to our colleagues, Jackson Wilke da Cruz Souza (UNIFAL, MG), Magnun Rochel Madruga (UFMG) and Roana Rodrigues (UFS), chairs of the VII Workshop on Portuguese Description, to colleagues Cláudia Dias de Barros (IFSP) and Daniela Barreiro Claro (UFBA), chairs of the VII Student Workshop on Information and Human Language Technology, TILic, and also to Livy Real, chair of the Latin American and Iberian Languages Open Corpora Forum, OpenCor. We also thank Ely Matos for his valuable help in the publication of these proceedings. Last but not least, we are grateful to Reinaldo Bianchi, BRACIS 2021 General Chair, and his team of collaborators.

November, 2021

Evandro Eduardo Seron Ruiz
Tiago Timponi Torrent

Conference chairs

Evandro Eduardo Seron Ruiz (Universidade de São Paulo)
Tiago Timponi Torrent (Universidade Federal de Juiz de Fora)

Program Committee

Alessandra Alaniz Macedo, FFCLRP-USP
Alexandre Rademaker, IBM Research
Aline Evers, UFRGS
Andre Adami, Universidade de Caxias do Sul
Ariani Di Felippo, UFSCar
Arnaldo Candido Junior, UTFPR
Arthur Lorenzi, Universidade Federal de Juiz de Fora
Bento Dias da Silva, UNESP
Carlos Prolo, Universidade Federal do Rio Grande do Norte
Carlos Ramisch, Aix Marseille Université
Cassia Trojahn dos Santos, IRIT & UTM2
Christopher Shulby, University of São Paulo
Clarissa Xavier, UFRGS
Claudia Barros, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP
Daniela Barreiro Claro, Federal University of Bahia
Diana Santos, Linguatca, Universidade de Oslo
Diego Amancio, USP
Diogo Cortiz, PUC-SP
Ely Matos, Universidade Federal de Juiz de Fora
Eraldo Fernandes, Universidade Federal de Mato Grosso do Sul
Eric Laporte, Université Gustave Eiffel
Erick Fonseca, Real Digital
Erick Maziero, Universidade Federal de Lavras
Evandro Eduardo Seron Ruiz, USP
Francis Bond, Nanyang Technological University
Geraldo Xexéo, UFRJ
Gustavo Paetzold, University of Sheffield
Helena Caseli, UFSCar
Heliana Mello, Universidade Federal de Minas Gerais
Horacio Saggion, Universitat Pompeu Fabra
Hugo Gonçalo Oliveira, Universidade de Coimbra
Isabel Trancoso, INESC-ID, IST
Ivandré Paraboni, USP Leste
Jorge Baptista, University Algarve
Leandro Mendonça de Oliveira, Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)
Livy Real, B2W Digital/GLiC
Lucelene Lopes, USP/ICMC
Marcelo Barros Custodio, Universidade Federal de Juiz de Fora
Marcelo Finger, USP/IME
Maria das Graças Nunes, USP/ICMC

Maria José Finatto, Universidade Federal do Rio Grande do Sul
Marlo Souza, Universidade Federal da Bahia – UFBA
Mateus Machado, USP/ICMC
Nelson Neto, Federal University of Pará (UFPA)
Norton Roman, USP/EACH
Oto Vale, UFSCar
Paulo Cavalin, IBM Research Brazil
Paulo Quaresma, Universidade de Évora
Renata Vieira, PUCRS
Sergio Freitas, Universidade de Brasília
Stella Tagnin, USP
Thiago Pardo, USP/ICMC
Tiago Timponi Torrent, Universidade Federal de Juiz de Fora
Valéria Feltrim, Universidade Estadual de Maringá
Vithor Gomes Bertalan, USP
Vladia Pinheiro, Universidade de Fortaleza
Zheng Xin Yong, Brown University

CE-PLN Steering Committee

Leandro Henrique Mendonça de Oliveira (coordenador), EMBRAPA
Evandro Eduardo Seron Ruiz, USP
Marlo Souza, UFBA
Oto Vale, UFSCar
Thiago Alexandre Salgueiro Pardo, USP/São Carlos

Contents

I Main conference	xi
Porttinari - a Large Multi-genre Treebank for Brazilian Portuguese <i>Thiago Pardo and Magali Duran and Lucelene Lopes and Ariani Felippo and Norton Roman and Maria Nunes</i>	1
Utilizando um dicionário morfológico para expandir a cobertura lexical de uma gramática do português no formalismo HPSG <i>Ana Nunes and Alexandre Rademaker and Leonel Alencar</i>	11
Explorando a revisão de corpora por meio da comparação de regras gramaticais em padrões sintáticos <i>Wellington Silva and Alexandre Rademaker and Leonel Alencar</i>	19
PetroGold ? Corpus padrão ouro para o domínio do petróleo <i>Elvis Souza and Aline Silveira and Tatiana Cavalcanti and Maria Castro and Cláudia Freitas</i>	29
Lexicalidade biomédica e sua mensuração em um corpus sobre COVID-19 em língua portuguesa <i>Karhyme Assis and Camila Silva and Janaína Leite and Wellington Nogueira and Kenji Nose Filho and André Takahata and Margarethe Steinberger-Elias</i>	39
Análise de polaridade e de tópicos em tweets no domínio da política no Brasil <i>Leonardo Capellaro and Helena Caseli</i>	47
Utilizando BERTimbau para a Classificação de Emoções em Português <i>Luiz Hammes and Larissa Freitas</i>	56
Sentiment Analysis in Portuguese Texts from Online Health Community Forums: Data, Model and Evaluation <i>Yohan Gumiel and Isabela Lee and Tayane Soares and Thiago Ferreira and Adriana Pagano</i>	64
A Weakly Supervised Dataset of Fine-Grained Emotions in Portuguese <i>Diogo Cortiz and Jefferson Silva and Newton Calegari and Ana Freitas and Ana Soares and Carolina Botelho and Gabriel Rêgo and Waldir Sampaio and Paulo Boggio</i>	73
Learning rules for automatic identification of implicit aspects in Portuguese <i>Mateus Machado and Thiago Pardo and Evandro Ruiz and Ariani Felippo</i>	82
Text Mining for Cyberbullying Detection: a Brazilian Portuguese Evaluation <i>Carolina Eberhart and Luciano Ignaczak and Márcio Martins</i>	92

CONTENTS

Relation extraction in structured and unstructured data: a comparative investigation on smartphone titles in the e-commerce domain <i>João Barbirato and Livy Real and Helena Caseli</i>	101
Classificação multimodal para detecção de produtos proibidos em uma plataforma marketplace <i>Alan Romualdo and Livy Real and Helena Caseli</i>	111
Measuring Brazilian Portuguese Product Titles Similarity using Embeddings <i>Alan Romualdo and Livy Real and Helena Caseli</i>	121
Augmenting Customer Support with an NLP-based Receptionist <i>André Barbosa and Alan Godoy</i>	133
Audio MFCC-gram Transformers for respiratory insufficiency detection in COVID-19 <i>Marcelo Gavy and Marcelo Finger</i>	143
DP-Symptom-Identifier: uma estratégia para classificar sintomas de depressão utilizando um conjunto de dados textuais na língua portuguesa <i>Vinicius Casani and Alinne Souza and Rafael Mantovani and Francisco Souza</i>	153
Identificando sintomas de depressão em postagens do Twitter em português do Brasil <i>Augusto Mendes and Rafael Passador and Helena Caseli</i>	162
Detecção de desinformação sobre Covid-19 no Twitter <i>Ana Mota and Wellington Franco and César Mattos</i>	172
A Long Texts Summarization Approach to Scientific Articles <i>Cinthia Souza and Renato Vimieiro</i>	182
A Preliminary Study for Literary Rhyme Generation based on Neuronal Representation, Semantics and Shallow Parsing <i>Luis-Gil Moreno-Jiménez and Juan-Manuel Torres-Moreno and Roseli Wedemann</i>	190
Structural Characterization and Graph-based Detection of Fake News in Portuguese <i>Roney Santos and Thiago Pardo</i>	199
ReVera Framework: Um Framework para rastreabilidade em fact-checking automático <i>João Souza and Elias Assis and Fabrício Mendonça and Jairo Souza</i>	209
An Empirical Study of Information Retrieval and Machine Reading Comprehension Algorithms for an Online Education Platform <i>Eduardo Montesuma and Lucas Carneiro and Adson Damasceno and João Sampaio and Romulo Férrer Filho and Paulo Maia and Francisco Oliveira</i>	217
Assessing the Impact of Stemming Algorithms Applied to Brazilian Legislative Documents Retrieval <i>Ellen Souza and Gyovana Moriyama and Douglas Vitória and André Carvalho and Nádia Félix and Hidelberg Albuquerque and Adriano Oliveira</i>	227
verBERT: Automating Brazilian Case Law Document Multi-label Categorization Using BERT <i>Felipe Serras and Marcelo Finger</i>	237

CONTENTS

Annotation Difficulties in Natural Language Inference <i>Aikaterini-Lida Kalouli and Livy Real and Annebeth Buis and Martha Palmer and Valeria Paiva</i>	247
A machine learning approach to literary genre classification on Portuguese texts: circumventing NLP's standard varieties <i>Dionéia Monte-Serrat and Mateus Machado and Evandro Ruiz</i> .	255
Evaluation of Synthetic Datasets Generation for Intent Classification Tasks in Portuguese <i>Robson Paula and Décio Aguiar Neto and Davi Romero and Paulo Guerra</i>	265
Tackling neural machine translation in low-resource settings: a Portuguese case study <i>Arthur Estrella and João Souza Filho</i>	275
Uma revisão breve sobre perguntas complexas em bases de conhecimento para sistemas de perguntas e respostas <i>Jorão Gomes Jr. and Rômulo Mello and Ana Reis and Victor Ströele and Jairo Souza</i>	283
II VII Workshop on Portuguese Description	295
Efeitos da variação linguística na decisão lexical <i>Victor Souza and Raquel Freitag</i>	297
Palatalização na fala e na leitura de universitários sergipanos <i>Lucas Silva and Raquel Freitag</i>	307
A propósito do verbo falar no português brasileiro: uma análise em corpus e em bases de dados verbais <i>Isaac Miranda Junior and Marcela Couto and Francimeire Coelho and Roana Rodrigues and Oto Vale</i>	315
Provérbios portugueses usuais: distribuição em corpora <i>Sônia Reis and Jorge Baptista and Nuno Mamede</i>	325
Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies <i>Ariani Felippo and Caroline Postali and Gabriel Ceregatto and Laura Gazana and Emanuel Silva and Norton Roman and Thiago Pardo</i>	335
Descrição de numerais segundo modelo Universal Dependencies e sua anotação no português <i>Magali Duran and Lucelene Lopes and Thiago Pardo</i>	344
Construções de Estrutura Argumental com Argumento Preposicionado: uma modelagem linguístico-computacional na FrameNet Brasil <i>Vânia Almeida and Tiago Torrent</i>	353
Modelagem de Construções Interrogativas QU- no Constructicon da FrameNet Brasil <i>Natália Marção and Tiago Torrent</i>	363
Banco de dados VerboWeb: um panorama do léxico verbal do PB <i>Márcia Cançado and Luana Amaral and Letícia Meirelles and Thaís Bechir and Amanda Amanda</i>	372

CONTENTS

Engenharia de features linguísticas para classificação de triplas relacionais <i>Elían Luz and Camilla Silva and Daniela Claro</i>	381
Descrição de uma metodologia desenvolvida para revisão de um léxico de palavras de emoção <i>Barbara Ramos</i>	389
Respostas emocionais da variação linguística: Análise exploratória de rastreamento ocular <i>Raquel Freitag and Julian Tejada and René Almeida and Paloma Cardoso and Victor Souza and Vanesca Leal</i>	398
Complexidade textual em notícias satíricas: uma análise para o português do Brasil <i>Gabriela Wick-Pedro and Roney Santos</i>	409
Constituintes Frasais com Função de Sujeito em Sentenças Judiciais <i>Ester Motta and Maria Finatto</i>	416

III VII Student Workshop on Information and Human Language Technology 425

Compilação de um corpus etiquetado da Língua Geral Amazônica <i>Dominick Alexandre and Juliana Gurgel and Leonel Araripe</i>	427
Ferramenta linguístico-computacional como facilitadora para o ensino de gramática na escola <i>Livia Dutra and Natália Sigiliano</i>	432
Criação e Anotação do corpus de resumos científicos de Ciências Sociais Aplicadas <i>Sabrina Taniwaki and Jackson Souza</i>	437
Avaliação de parsers na detecção de relações essenciais do modelo Universal Dependencies para o português <i>Luana Belisário and Thiago Pardo</i>	442
Utilizando Pistas Linguística para Detectar Conteúdo Enganoso em Português <i>Rodrigo Rodrigues and Larissa Freitas</i>	447