

PROPOR 2024

**Proceedings of the 16th International Conference on Computational  
Processing of the Portuguese Language, PROPOR 2024 - vol. 1**

12–15 March, 2024  
Universidade de Santiago de  
Compostela, Galicia, Spain



©2024 The Association for Computational Linguistics

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 979-8-89176-062-2

# Introduction

These Proceedings present the 16th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2024), held for the first time in Galicia, at the Universidade de Santiago de Compostela, Spain, during March 12–15, 2024.

PROPOR continues to be a key forum bringing together researchers dedicated to the computational processing of Portuguese, promoting the exchange of experiences in the development of methodologies, language resources, tools, applications, and innovative projects. In this edition, the conference was held in Galicia, the birthplace of the Portuguese language, and was extended to Galician (considered as a local variety of the former). PROPOR 2024 is, thus, the main scientific event in the area of natural language processing that is focused on theoretical and technological issues of written and spoken Portuguese and Galician.

This year’s edition includes an Industry track, alongside the General track. We received a total of 111 submissions — 81 long papers and 30 short papers —, from 152 different authors working on Portuguese and Galician around the world, in countries such as Portugal, Brazil, Spain, Norway, France, and Germany.

The Proceedings of PROPOR 2024 are divided in two volumes. The first volume presents the 51 long papers accepted (reflecting an acceptance rate of 63%), and the 20 short papers accepted (an acceptance rate of 66,6%), plus the papers presenting the two Best Dissertations, selected by the jury from a set of 12 candidates. All submissions were peer-reviewed by 104 reviewers from a total of 60 academic institutions and companies, coming from 14 countries. Full papers are organized thematically and show research and developments in resources and evaluation, natural language processing tasks, natural language processing applications, speech processing and applications, and lexical semantics. The second volume of the Proceedings is dedicated to the demos and to the works selected and presented in the four workshops hosted by PROPOR 2024, namely the *PROPOR’24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays*, the *5th OpenCor: Latin American and Iberian Languages Open Corpora Forum*, the *Third Workshop on Digital Humanities and Natural Language Processing*, and the *First Workshop on NLP for Indigenous Languages of Lusophone Countries*.

The invited speakers of PROPOR 2024 are experts from different fields and with diverse research and professional experiences, showing the range, relevance and applicability of the computational processing of Portuguese and Galician:

- Elias Feijó, Lecturer of Portuguese language, literature, and culture at the Universidade de Santiago de Compostela and director of the Galabra Group, will reflect on the question “É este galego latim em pó?”.
- Gemma Boleda, ICREA Research Professor of the Department of Translation and Language Sciences at the Universitat Pompeu Fabra, will talk about ‘Pressures on the lexicon and their effects’.
- Marta Ruiz Costa-jussà, currently research scientist at FAIR, Meta, and recipient of an ERC Starting Grant and two Google Faculty Research Awards, will present “Beyond Semantic Evaluation in Seamless Speech Translation Models”.

PROPOR 2024 is a three and a half days conference that encompasses one full-day of workshops and shared tasks and two and a half days of communications, demos and community meetings.

Our sincere thanks go to every person and institution involved in the complex organization of this event, especially to the members of the Program Committee of the main event, the dissertations contest and the associated workshops chairs, the invited speakers, and the general organization staff. We are also grateful to the agencies and

organizations that supported and promoted the event.

Thank you all for participating and we wish you an enjoyable and inspiring conference!

Pablo Gamallo

Daniela Claro

António Teixeira

Livy Real

Marcos Garcia

Hugo Gonçalo Oliveira

Raquel Amaro

March 2024

**Organization:**

*General Chairs:* Pablo Gamallo and Daniela Claro

*Program Chairs:* António Teixeira, Livy Real, and Marcos Garcia

*Editorial Chairs:* Hugo Gonçalo Oliveira and Raquel Amaro

*Demo Chairs:* Iria de Dios and Marlo Souza

*Workshop/Tutorial Chairs:* Alberto Abad, Alberto Simões and Helena Caseli

*Best Dissertation Chairs:* António Branco, Paulo Quaresma and Renata Vieira

*Industry Track Chairs:* José Ramom Pichel and Luís Trigo

*Local Organizers:* Daniel Bardanca and José Ramom Pichel

*Consultants:* Carolina Scarton and Fernando Batista

**Program Committee:**

Adina Valdu, Universidade de Santiago de Compostela (Spain)

Adriana Pagano, Federal University of Minas Gerais (Brazil)

Alberto Abad, INESC-ID and Instituto Superior Técnico, Universidade de Lisboa (Portugal)

Alberto Simões, 2Ai Lab - IPCA (Portugal)

Alexandre Rademaker, IBM Research and FGV/EMAp (Brazil)

Aline Paes, Fluminense Federal University (Brazil)

Amália Mendes, University of Lisbon (Portugal)

Ana Isabel Mata, University of Lisbon (Portugal)

Ana Luísa V. Leal, University of Macau (Macau)

Antonio Bonafonte, Amazon (Spain)

António Branco, University of Lisbon (Portugal)

António Teixeira, University of Aveiro (Portugal)

Ariani Di Felippo, Federal University of São Carlos (Brazil)

Arnaldo Candido Junior, Federal University of Technology - Paraná (Brazil)

Berthold Crysmann, CNRS and University of Paris (France)

Brett Drury, Liverpool Hope University (UK)

Bruno Martins, INESC-ID and Instituto Superior Técnico, Universidade de Lisboa (Portugal)

Carlos A. Prolo, Federal University of Rio Grande do Norte (Brazil)

Carlos Ramisch, Aix-Marseille University (France)

Carmen Magariños, Universidade de Santiago de Compostela (Spain)

Catarina Oliveira, University of Aveiro (Portugal)

Christopher Shulby, Kin AI (Brazil)

Cristiane Namiuti, Universidade Estadual do Sudoeste da Bahia (Brazil)

Cristina Carbajal, Universidade de Santiago de Compostela (Spain)

Daniela Claro, Federal University of Bahia (Brazil)

Daniel Beck, University of Melbourne (Australia)

David Martins de Matos, INESC-ID and Instituto Superior Técnico, Universidade de Lisboa (Portugal)

Diamantino Freitas, University of Oporto (Portugal)

Diana Santos, LINGUATECA and University of Oslo (Norway)

Eloize Rossi M. Seno, Federal Institute of São Paulo (Brazil)

Eric Laporte, Université Gustave Eiffel (France)  
Evelin Amorim, INESC TEC (Portugal)  
Fernando Batista, INESC-ID and ISCTE-IUL (Portugal)  
Fernando Perdigão, Instituto de Telecomunicações and University of Coimbra (Portugal)  
Gaël Dias, University of Caen Normandy (France)  
Helena Bermúdez, JinnTec GmbH (Germany)  
Helena Caseli, Federal University of São Carlos (Brazil)  
Hugo Gonçalo Oliveira, CISUC and University of Coimbra (Portugal)  
Irene Rodrigues, University of Évora (Portugal)  
Iria de Dios Flores, Universidade de Santiago de Compostela (Spain)  
Isabel Falé, Universidade Aberta and University of Lisbon (Portugal)  
Isabel Trancoso, INESC-ID and Instituto Superior Técnico, Universidade de Lisboa (Portugal)  
Ivandré Paraboni, University of São Paulo (Brazil)  
Ivanovitch Silva, Universidade Federal do Rio Grande do Norte (Brazil)  
Javier González Corbelle, Universidade de Santiago de Compostela (Spain)  
João Balsa, University of Lisbon (Portugal)  
João Silva, University of Lisbon (Portugal)  
Jorge Baptista, University of Algarve and INESC-ID (Portugal)  
José Ramom Pichel, Universidade de Santiago de Compostela (Spain)  
José Saias, University of Évora (Portugal)  
Leandro Oliveira, Embrapa Informatica Agropecuaria (Spain)  
Leonardo Zilio, Friedrich-Alexander-Universität Erlangen-Nürnberg (Germany)  
Livy Real, QuintoAndar Inc. (Brazil)  
Luciana Benotti, National University of Córdoba (Argentina)  
Magali Duran, University of São Paulo (Brazil)  
Marcelo Finger, University of São Paulo (Brazil)  
Marcos Fernández Pichel, Universidade de Santiago de Compostela (Spain)  
Marcos Garcia, Universidade de Santiago de Compostela (Spain)  
Marcos Spalenza, UFES (Brazil)  
Marcos Treviso, Instituto de Telecomunicações (Brazil)  
Maria das Graças Volpes Nunes, University of São Paulo (Brazil)  
Maria José B. Finatto, Federal University of Rio Grande do Sul (Brazil)  
Mario Ezra Aragon, Universidade de Santiago de Compostela (Spain)  
Mário Rodrigues, ESTGA/IEETA - University of Aveiro (Portugal)  
Marlo Souza, Federal University of Bahia (Brazil)  
Martín Pereira-Fariña, University of Santiago de Compostela (Spain)  
Mikel L. Forcada, University of Alacant (Spain)  
Nádia Silva, University of São Paulo (Brazil)  
Nelson Neto, Federal University of Pará (Brazil)  
Norton Roman, University of São Paulo (Brazil)  
Oto Vale, Federal University of São Carlos (Brazil)  
Pablo Faria, University of Campinas (Brazil)  
Palmira Marrafa, University of Lisbon (retired) (Portugal)  
Paula Cardoso, Federal University of Lavras (Brazil)  
Paulo Quaresma, University of Évora (Portugal)  
Plinio Barbosa, University of Campinas (Brazil)  
Prakash Poudyal, Kathmandu University (Nepal)

Raquel Amaro, NOVA University Lisbon (Portugal)  
Renata Vieira, CIDEHUS and University of Évora (Portugal)  
Ricardo Ribeiro, INESC-ID and ISCTE (Portugal)  
Ricardo Rodrigues, CISUC and Polytechnic Institute of Coimbra (Portugal)  
Sandra Aluísio, University of São Paulo (Brazil)  
Sara Mendes, University of Lisbon (Portugal)  
Saullo Haniell, Pontifícia Universidade Católica de Campina (Brazil)  
Sheila Castilho, Dublin City University (Ireland)  
Sofía García, UPV/EHU (Basque Country/Spain)  
Susana Duarte Martins, NOVA University Lisbon (Portugal)  
Susana Sotelo Docío, University of Santiago de Compostela (Spain)  
Tamás Gábor Csapó, Budapest University of Technology and Economics (Hungary)  
Teresa Gonçalves, University of Évora (Portugal)  
Thiago Ferreira, aiXplain inc. (USA)  
Thiago Pardo, University of São Paulo (Brazil)  
Valeria de Paiva, Topos Institute (USA)  
Valeria Feltrim, Universidade Estadual de Maringá (Brazil)  
Violeta Quental, Pontifical Catholic University of Rio de Janeiro (Brazil)  
Vládía Pinheiro, Universidade de Fortaleza (Brazil)

#### **Additional Reviewers:**

Erik Bran Marino, Universidade de Évora (Portugal)  
Márcio Lima Inácio, University of Coimbra (Portugal)  
Isabel Carvalho, University of Coimbra (Portugal)  
Rodrigo Wilkens, Université Catholique de Louvain (Belgium)  
André Santos, INESC TEC (Portugal)  
Felipe Ribas Serras, Universidade de São Paulo (Brazil)  
Guido Ivorra, Universidad Nacional de Córdoba (Argentina)

# Table of Contents

## Long Papers

<b>A Multilingual Dataset for Investigating Stereotypes and Negative Attitudes Towards Migrant Groups in Large Language Models</b> . . . . .	1
<i>Danielly Sorato, Carme Colominas Ventura and Diana Zavala-Rojas</i>	
<b>Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? A Preliminary Study</b> . . . . .	13
<i>Gladson de Araujo, Tiago de Melo and Carlos Maurício S. Figueiredo</i>	
<b>A Galician Corpus for Misogyny Detection Online</b> . . . . .	22
<i>Lucía M. Álvarez-Crespo and Laura M. Castro</i>	
<b>Simple and Fast Automatic Prosodic Segmentation of Brazilian Portuguese Spontaneous Speech</b> . . . . .	32
<i>Giovana Meloni Craveiro, Vinicius Gonçalves Santos, Gabriel Jose Pellisser Dalalana, Flaviane R. Fernandes Svartman and Sandra Maria Aluísio</i>	
<b>LLMs and Translation: different approaches to localization between Brazilian Portuguese and European Portuguese</b> . . . . .	45
<i>Eduardo G. Cortes, Ana Luíza Vianna, Mikaela Martins, Sandro Rigo and Rafael Kunst</i>	
<b>SPARQL can also talk in Portuguese: answering natural language questions with knowledge graphs</b> . . . . .	56
<i>Elbe Miranda, Aline Paes and Daniel de Oliveira</i>	
<b>Exploring Pre-Trained Transformers for Translating Portuguese Text to Brazilian Sign Language</b> . . . . .	67
<i>Jose Mario De Martino and Dener Stassun Christinele</i>	
<b>NLP for historical Portuguese: Analysing 18th-century medical texts</b> . . . . .	76
<i>Leonardo Zilio, Rafaela Radünz Lazzari and Maria Jose Bocorny Finatto</i>	
<b>Text Summarization and Temporal Learning Models Applied to Portuguese Fake News Detection in a Novel Brazilian Corpus Dataset</b> . . . . .	86
<i>Gabriel Lino Garcia, Pedro Henrique Paiola, Danilo Samuel Jodas, Luis Afonso Sugi and João Paulo Papa</i>	
<b>Automatic Text Readability Assessment in European Portuguese</b> . . . . .	97
<i>Eugénio Ribeiro, Nuno Mamede and Jorge Baptista</i>	
<b>Toxic Speech Detection in Portuguese: A Comparative Study of Large Language Models</b> . . . . .	108
<i>Amanda da Silva Oliveira, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas and Eduardo José da Silva Luz</i>	
<b>Named entity recognition specialised for Portuguese 18th-century History research</b> . . . . .	117
<i>Joaquim Santos, Helena Freire Cameron, Fernanda Olival, Fátima Farrica and Renata Vieira</i>	
<b>Exploring Open Information Extraction for Portuguese Using Large Language Models</b> . . . . .	127
<i>Bruno Cabral, Daniela Claro and Marlo Souza</i>	
<b>Bringing Pragmatics to Porttinari - Adding Speech Acts to News Texts</b> . . . . .	137
<i>Nataly L. Patti da Silva, Norton Trevisan Roman and Ariani Di Felippo</i>	
<b>Authorship Attribution with Rejection Capability in Challenging Contexts of Limited Datasets</b> . . . . .	146
<i>Pedro Oliveira and Joaquim Silva</i>	
<b>Using Large Language Models for Identifying Satirical News in Brazilian Portuguese</b> . . . . .	156
<i>Gabriela Wick-Pedro, Cássio Faria da Silva, Marcio Lima Inácio, Oto Araújo Vale and Helena de Medeiros Caseli</i>	
<b>Semantic Permanence in Audiovisual Translation: a FrameNet approach to subtitling</b> . . . . .	168



<i>Mairon Samagaio, Tiago Torrent, Ely Matos and Arthur Almeida</i>	
<b>Hurdles in Parsing Multi-word Adverbs: Examples from Portuguese</b>	177
<i>Izabela Muller, Nuno Mamede and Jorge Baptista</i>	
<b>Portal NURC-SP: Design, Development, and Speech Processing Corpora Resources to Support the Public Dissemination of Portuguese Spoken Language</b>	187
<i>Ana Carolina Rodrigues, Alessandra A. Macedo, Arnaldo Candido Jr, Flaviane R. F. Svartman, Giovana M. Craveiro, Marli Quadros Leite, Sandra M. Aluísio, Vinícius G. Santos and Vinícius M. Garcia</i>	
<b>TransAlign: An Automated Corpus Generation through Cross-Linguistic Data Alignment for Open Information Extraction</b>	196
<i>Alan Rios, Bruno Cabral, Daniela Claro, Rerisson Cavalcante and Marlo Souza</i>	
<b>BATS-PT: Assessing Portuguese Masked Language Models in Lexico-Semantic Analogy Solving and Relation Completion</b>	207
<i>Hugo Gonçalo Oliveira, Ricardo Rodrigues, Bruno Ferreira, Purificação Silvano and Sara Carvalho</i>	
<b>Towards the automatic creation of NER systems for new domains</b>	218
<i>Emanuel Matos, Mário Rodrigues and António Teixeira</i>	
<b>A New Benchmark for Automatic Essay Scoring in Portuguese</b>	228
<i>Igor Cataneo Silveira, André Barbosa and Denis Deratani Mauá</i>	
<b>Predicting the Age of Emergence of Consonants</b>	238
<i>Luís Jesus and Jihen Trabelsi</i>	
<b>Applying event classification to reveal the Estado da Índia</b>	247
<i>Gonçalo C. Albuquerque, Marlo Souza, Renata Vieira and Ana Sofia Ribeiro</i>	
<b>Exploring Computational Discernibility of Discourse Domains in Brazilian Portuguese within the Carolina Corpus</b>	255
<i>Felipe Ribas Serras, Mariana Sturzeneker, Miguel de Mello Carpi, Mayara Feliciano Palma, Maria Clara Ramos Morales Crespo, Aline Silva Costa, Vanessa Martins Do Monte, Cristiane Namiuti, Maria Clara Paixão de Souza and Marcelo Finger</i>	
<b>Identification of Types of Event-Time Temporal Relations in Portuguese Using a Rule-Based Approach</b>	266
<i>Dárcio S. Rocha, Marlo Souza and Daniela B. Claro</i>	
<b>A Corpus of Stock Market Tweets Annotated with Named Entities</b>	276
<i>Michel Monteiro Zerbinati, Norton Trevisan Roman and Ariani Di Felippo</i>	
<b>Frequency, overlap and origins of palatal sonorants in three Iberian languages</b>	285
<i>Carlos Silva and Luís Trigo</i>	
<b>A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning</b>	290
<i>Rafael Oleques Nunes, Dennis Giovani Balreira, André Suslik Spritzer and Carla Maria Dal Sasso Freitas</i>	
<b>Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models?</b>	301
<i>Gabriel Assis, Annie Amorim, Jonnatahn Carvalho, Daniel de Oliveira, Daniela Vianna and Aline Paes</i>	
<b>Aspect-based sentiment analysis in comments on political debates in Portuguese: evaluating the potential of ChatGPT</b>	312
<i>Eloize Seno, Lucas Silva, Fábio Anno, Fabiano Rocha and Helena Caseli</i>	
<b>CLSJR.BR - A Model for Abstractive Summarization of Legal Documents in Portuguese Language based on Contrastive Learning</b>	321

Alex Aguiar Lins, Cecilia Silvestre Carvalho, Francisco Das Chagas Jucá Bomfim, Daniel de Carvalho Bentes and Vlândia Pinheiro

Investigating the Generalizability of Portuguese Readability Assessment Models Trained Using Linguistic Complexity Features . . . . .	332
<i>Soroosh Akef, Amália Mendes, Detmar Meurers and Patrick Rebuschat</i>	
UlyssesNERQ: Expanding Queries from Brazilian Portuguese Legislative Documents through Named Entity Recognition . . . . .	342
<i>Hidelberg Albuquerque, Ellen Souza, Tainan Silva, Rafael P. Gouveia, Flavio Junior, Douglas Vitorio, Nádia F. F. da Silva, André C.P.L.F. de Carvalho, Adriano L.I. Oliveira and Francisco Edmundo de Andrade</i>	
Across the Atlantic: Distinguishing Between European and Brazilian Portuguese Dialects . . . . .	353
<i>David Preda, Tomás Osório and Henrique Lopes Cardoso</i>	
Accent Classification is Challenging but Pre-training Helps: a case study with novel Brazilian Portuguese datasets . . . . .	364
<i>Ariadne Matos, Gustavo Araújo, Arnaldo Candido Junior and Moacir Ponti</i>	
RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese . . . . .	374
<i>Eduardo A. S. Garcia, Nadia F. F. Silva, Felipe Siqueira, Hidelberg O. Albuquerque, Juliana R. S. Gomes, Ellen Souza and Eliomar A. Lima</i>	
Evaluating Pre-training Strategies for Literary Named Entity Recognition in Portuguese . . . . .	384
<i>Mariana O. Silva and Mirella M. Moro</i>	
Brazilian Portuguese Product Reviews Moderation with AutoML . . . . .	394
<i>Lucas Nildaimon dos Santos Silva, Livy Real, Fernando Rezende Zagatti, Ana Claudia Bianchini Zandavalle, Tatiana da Silva Gama and Carolina Francisco Gadelha Rodrigues</i>	
Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework . . . . .	401
<i>Lucelene Lopes and Thiago Pardo</i>	
Study of the State of the Art Galician Machine Translation: English-Galician and Spanish-Galician models . . . . .	411
<i>Sofía García González and German Rigau Claramunt</i>	
Bartoli's areal norms revisited: an agent-based modeling approach . . . . .	422
<i>Dalmo Buzato and Evandro Cunha</i>	
RePro: a benchmark for Opinion Mining for Brazilian Portuguese . . . . .	432
<i>Lucas Nildaimon dos Santos Silva, Livy Real, Ana Claudia Bianchini Zandavalle, Carolina Francisco Gadelha Rodrigues, Tatiana da Silva Gama, Fernando Guedes Souza and Phillipe Derwich Silva Zaidan</i>	
Glória: A Generative and Open Large Language Model for Portuguese . . . . .	441
<i>Ricardo Lopes, Joao Magalhaes and David Semedo</i>	
Evaluating large language models for the tasks of PoS tagging within the Universal Dependency framework . . . . .	454
<i>Mateus Machado and Evandro Ruiz</i>	
Question Answering for Dialogue State Tracking in Portuguese . . . . .	461
<i>Francisco Pais, Patricia Ferreira, Catarina Silva, Ana Alves and Hugo Gonçalo Oliveira</i>	
Toxic Content Detection in online social networks: a new dataset from Brazilian Reddit Communities . . . . .	472
<i>Luiz Henrique Quevedo Lima, Adriana Silvina Pagano and Ana Paula Couto da Silva</i>	
A Natural Language Text to Role-Playing Game Animation Generator . . . . .	483
<i>Caio F. Oliveira, Artur Franco, Wellington Franco and José G. R. Maia</i>	

<b>From Random to Informed Data Selection: A Diversity-Based Approach to Optimize Human Annotation and Few-Shot Learning</b> . . . . .	492
<i>Alexandre Alcoforado, Lucas Hideki Takeuchi Okamura, Israel Campos Fama, Bárbara Fernandes Dias Bueno, Arnold Moya Lavado, Thomas Palmeira Ferraz, Bruno Veloso and Anna Helena Reali Costa</i>	
<b>Enhancing Stance Detection in Low-Resource Brazilian Portuguese Using Corpus Expansion generated by GPT-3.5</b> . . . . .	503
<i>Dyonnatan Maia and Nádia Félix Felipe da Silva</i>	
<b>Short Papers</b>	
<b>A Bag-of-Users approach to mental health prediction from social media data</b> . . . . .	509
<i>Rafael Oliveira and Ivandré Paraboni</i>	
<b>Semi-automatic corpus expansion: the case of stance prediction</b> . . . . .	515
<i>Camila Pereira and Ivandré Paraboni</i>	
<b>Sequence-to-sequence and transformer approaches to Portuguese text style transfer</b> . . . . .	521
<i>Pablo Costa and Ivandré Paraboni</i>	
<b>Comparative Analysis of Intentional Gramatical Error Correction Techniques on Twitter/X</b> . . . . .	527
<i>Thainá Marini and Taffarel Brant-Ribeiro</i>	
<b>Towards a Syntactic Lexicon of Brazilian Portuguese Adjectives</b> . . . . .	532
<i>Ryan Martinez, Jorge Baptista and Oto Vale</i>	
<b>Literary similarity of novels in Portuguese</b> . . . . .	539
<i>Diana Santos</i>	
<b>An evaluation of Portuguese language models' adaptation to African Portuguese varieties</b> . . . . .	544
<i>Diego Fernando Válio Antunes Alves</i>	
<b>Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches</b> . . . . .	551
<i>Eugénio Ribeiro, Nuno Mamede and Jorge Baptista</i>	
<b>Is it safe to machine translate suicide-related language from English to Galician?</b> . . . . .	558
<i>John E. Ortega and Annika Marie Schoene</i>	
<b>First assessment of Graph Machine Learning approaches to Portuguese Named Entity Recognition</b> . . . . .	563
<i>Gabriel Silva, Mário Rodrigues, António Teixeira and Marlene Amorim</i>	
<b>Exploring Multimodal Models for Humor Recognition in Portuguese</b> . . . . .	568
<i>Marcio Inácio and Hugo Gonçalo Oliveira</i>	
<b>RecognaSumm: A Novel Brazilian Summarization Dataset</b> . . . . .	575
<i>Pedro Henrique Paiola, Gabriel Lino Garcia, Danilo Samuel Jodas, João Vitor Mariano Correia, Luis Afonso Sugi and João Paulo Papa</i>	
<b>A Speech-Driven Talking Head based on a Two-Stage Generative Framework</b> . . . . .	580
<i>Brayan Bernardo and Paula Costa</i>	
<b>Increasing manually annotated resources for Galician: the Parallel Universal Dependencies Treebank</b> . . . . .	587
<i>Xulia Sánchez-Rodríguez, Albina Sarymsakova, Laura Castro and Marcos Garcia</i>	
<b>CorpusNÓS: A massive Galician corpus for training large language models</b> . . . . .	593
<i>Iria de-Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia and Pablo Gamallo</i>	

Exploring the effects of vocabulary size in neural machine translation: Galician as a target language . . . . .	600
<i>Daniel Bardanca Outeirinho, Pablo Gamallo Otero, Iria de-Dios-Flores and José Ramon Pichel Campos</i>	
A Reproducibility Analysis of Portuguese Computational Processing Conferences: A Case of Study . . . . .	605
<i>Daniel Leal, Anthony Luz and Rafael Anchiêta</i>	
Automated admissibility of complaints about fraud and corruption . . . . .	610
<i>Thiago De Paula, André Do Amaral, Andre Victor, Luis Alberto Sales, Rodrigo Moreira, Thiago Meirelles and Rafael Basso</i>	
Natural Language Processing Application in Legislative Activity: a Case Study of Similar Amendments in the Brazilian Senate . . . . .	614
<i>Diany Pressato, Pedro Lucas Castro de Andrade, Flávio Rocha Junior, Felipe Alves Siqueira, Ellen Polliana Ramos Souza, Nádia Félix Felipe da Silva, Márcio de Souza Dias and André Carlos Ponce de Leon Ferreira de Carvalho</i>	
Spatial Information Challenges in English to Portuguese Machine Translation . . . . .	620
<i>Rafael Fernandes, Rodrigo Souza, Marcos Lopes, Paulo Santos and Thomas Finbow</i>	
Compilation and tagging of a corpus with Celpe-Bras texts . . . . .	627
<i>Juliana Schoffen, Elisa Stumpf, Deise Amaral, Luiza Divino, Isadora Hanauer, Isabel Lisboa, Amanda Raupp and Brenda Xavier</i>	
<b>Best Dissertations</b>	
TTS applied to the generation of datasets for automatic speech recognition . . . . .	633
<i>Edresson Casanova, Sandra Aluísio and Moacir Antonelli Ponti</i>	
Text clustering applied to unbalanced data in legal contexts (Final Version) . . . . .	639
<i>Lucas José Gonçalves Freitas</i>	

# Evaluating large language models for the tasks of PoS tagging within the Universal Dependency framework

**Mateus Tarcinalli Machado**

Dep. of Computing and Mathematics  
FFCLRP, University of São Paulo  
mateusmachado@usp.br

**Evandro Eduardo Seron Ruiz**

Dep. of Computing and Mathematics  
FFCLRP, University of São Paulo  
evandro@usp.br

## Abstract

Large language models (LLMs) have emerged as a valuable tool for a variety of natural language processing tasks. This study focuses on assessing the capabilities of three language models in the context of part-of-speech tagging using the Universal Dependency (UPoS) tagset in texts written in Brazilian Portuguese. Our experiments reveal that LLMs can effectively leverage prior knowledge from existing tagged datasets and can also extract linguistic structure with arbitrary labels. Furthermore, we present results indicating an accuracy of 90% in UPoS tagging for a multilingual model, while smaller monolingual models achieve an accuracy of 48%.

## 1 Introduction

The rapid advancements in information and communication technologies have ignited significant interest in Natural Language Processing (NLP) tools. Consequently, this has led to the creation of a multitude of diverse NLP tools (Green, 2017). However, numerous challenges persist in the development of efficient and reliable NLP tools for accurate natural language processing. One tool addressing these challenges is Part-of-Speech (PoS) tagging, which involves assigning appropriate and unique grammatical categories (PoS tags) to words in a sentence (Inoue et al., 2017). Despite significant research endeavors, PoS tagging still faces challenges in improving accuracy, reducing false-positive rates, and efficiently handling the tagging of unknown words (Chiche and Yitagesu, 2022).

Supervised learning tasks have traditionally been prevalent in Natural Language Processing until recently. These tasks include question answering (Roy et al., 2023), machine translation (Wei et al., 2022), reading comprehension (Ouyang and Fu, 2022), and sentiment analysis (Shah et al., 2022), and they are typically tackled using specific datasets. Nevertheless, the landscape has evolved

as large language models (LLMs) have started to learn these tasks without explicit supervision (Min et al., 2023). This shift has occurred as these models are trained on vast datasets consisting of billions of words. The core concept is to acquire a universal, underlying language representation from a general task initially and subsequently apply it to various NLP tasks. Language modeling functions as the general task, given its ample availability of self-supervised text for extensive training.

In a paper by Radford et al., 2019, it was demonstrated that GPT-2, which was a 1.5-billion-parameter Transformer model at the time, achieved state-of-the-art results on 7 out of 8 tested language modeling datasets in a zero-shot learning setting. Later, Perez et al., 2021 conducted a study to evaluate pretrained LLMs in true few-shot learning scenarios, where held-out examples were unavailable. Their study highlighted the overestimation of LLMs’ true few-shot capabilities in previous work, due to the challenges in selecting effective models which were cross-validation and minimum description length, for LLM prompts and hyperparameter selection. More recently, Qin et al., 2023 have shown the rapid adoption of tools like ChatGPT in various NLP tasks. Going a step further, (Kuzman et al., 2023) postulate the hypothesis of the ‘beginning of the end of corpus annotation tasks’ with the advent of large language models.

As we have seen, LLM have made extraordinary progress in many NLP tasks. But, in the unsupervised PoS tagging task of texts written in Portuguese, works utilizing the language models are few and even fewer if we consider the state-of-the-art (SotA) tags from the Universal Dependency (De Marneffe et al., 2021) framework for grammar annotation.

The contributions of this work can be summarized as follows: (1) We conducted an evaluation of the SotA LLMs for the task of part-of-speech (PoS) tagging in Portuguese within the Universal Depen-

dencies (UD) model, here called UPoS tagging. (2) We discovered that UPoS tagging using LLMs, which may leverage prior knowledge from existing tagged datasets, can also extract linguistic structure with arbitrary labels. (3) We presented an analysis to measure the impact of this practical labeling process. In essence, our findings provide valuable insights into the proficiency of these generalized LLMs in excelling at specialized tasks and shed light on the effectiveness of the teaching process for these language models.

The remainder of this paper is organized as follows: In the subsequent section (Section 2), we provide a literature review on part-of-speech (PoS) tagging using Large Language Models (LLMs). Section 3 outlines the corpora utilized and the methodology adopted. Section 4 presents preliminary findings concerning the task of few-shot Universal Part-of-Speech (UPoS) tagging. Finally, Section 5 concludes the paper.

## 2 Related work

Part-of-Speech (PoS) tagging is a challenging task that involves classifying words to label their morphosyntactic information within a sentence. Accurate and dependable PoS tagging is essential for numerous natural language processing (NLP) tasks. Typically, extensive annotated corpora are required to achieve the desired accuracy of PoS taggers. However, recently, Large Language Models (LLMs) have emerged as valuable tools for a wide range of exciting NLP applications, such as Named Entity Recognition (NER), Relation Classification, Natural Language Inference (NLI), Question Answering (QA), Common Sense Reasoning (CSR), Summarization, and, of course PoS tagging (Qin et al., 2023).

In their study, Blevins et al., 2022 tackled the question of whether pretrained language models (PLMs) primarily rely on generalizable linguistic comprehension or surface-level lexical patterns when applied to a wide array of language tasks. To investigate this, they introduced a structured prompting approach designed for linguistic structured prediction tasks, which facilitated zero- and few-shot sequence tagging using autoregressive PLMs. The researchers extended their evaluation to UPoS for the English language. They executed structured prompting using GPT-3 models via the OpenAI API<sup>1</sup>, specifically employing the

base GPT-Curie (approximately 6 billion parameters) and GPT-Davinci (approximately 175 billion parameters). The results showed an accuracy of 66.27% for GPT-Curie and 65.9% for GPT-Davinci.

Lai et al., 2023 recently conducted tests on ChatGPT across seven different tasks, spanning 37 diverse languages with varying levels of resources, including high, medium, low, and extremely low resource languages. In their experiments, they employed the XGLUE-POS dataset (Liang et al., 2020) from Huggingface Datasets<sup>2</sup>, which encompasses 17 languages, excluding Portuguese. ChatGPT’s evaluation was carried out with both English (en) and language-specific (spc) task descriptions, achieving accuracies of 88.5% and 89.6%, respectively. Additionally, they utilized ChatGPT for PoS tagging in 16 other languages, obtaining an average accuracy of 84.5% and 79.8% (spc).

Our literature review has identified CamemBERT (Martin et al., 2020) as the pioneering monolingual Large Language Model (LLM) utilized for Part-of-Speech (PoS) tagging tasks. It is worth mentioning some previous works analyzing how linguistic information (including PoS) is encoded in the different layers of a (monolingual) transformer (Tenney et al., 2019; Liu et al., 2019). In their paper, the researchers examine the feasibility of training monolingual Transformer-based language models for languages other than English, using French as an illustrative case. In their study, the researchers assess the performance of language models across multiple language-related tasks, encompassing UPoS tagging, dependency parsing, named entity recognition, and natural language inference. In the case of the fine-tuned CamemBERT model, its UPoS data reached an impressive accuracy of 98.18%.

Finally, Chang’s belief, as mentioned in Chang et al., 2023, is that evaluation should be considered an essential discipline in order to better support the development of Large Language Models (LLMs).

In the following section, we will introduce the datasets and methods examined in this paper.

## 3 Data and methods

### Dataset and resources

In line with our objective to investigate SotA LLMs for PoS tagging in Portuguese within the Universal

<sup>1</sup><https://openai.com/blog/openai-api>

<sup>2</sup><https://huggingface.co/datasets/xglue>



Dependencies (UD) framework, we opted to employ the recently released Porttinari (Duran et al., 2023). Porttinari (which stands for ‘PORTuguese Treebank’) is a substantial and diverse treebank for Brazilian Portuguese, encompassing various genres. For our study, we specifically focused on the journalistic segment of the Porttinari treebank. This resource has been thoughtfully designed to serve as a versatile asset for NLP tasks in Brazilian Portuguese, with a special emphasis on the human-revised section, which comprises a total of 8,418 sentences.

### 3.1 UD PoS tags

Universal PoS tags (UPoS) are standardized grammatical labels utilized in Universal Dependencies (UD), a project aimed at creating consistent treebank annotations across multiple languages. The UPoS tagset comprises 17 tags designed to mark the core part-of-speech categories. These tags are categorized into three main groups, as outlined below:

**Open class words** ADJ, ADV, INTJ, NOUN, PROPN, and VERB;

**Closed class words** ADP, AUX, CCONJ, DET, NUM, PART, PRON, and SCONJ;

**Other** PUNCT, SYM, X

### Large Language Models

In this experiment, we employed the following Large Language Models (LLMs):

- LLaMA is a series of LLMs introduced by Meta AI<sup>3</sup>, released in February 2023. LLaMA language models have parameter counts ranging from 7 to 65 billion. These models were trained on trillions of tokens, demonstrating the possibility of achieving SotA model performance using only publicly available datasets (Touvron et al., 2023). Since we installed the models locally, we chose to use the LLaMA-7B version;
- Maritaca<sup>4</sup> represents a collection of LLMs that have undergone training using text written in Portuguese. The available documentation does not provide clear information regarding the specific LLM that the API utilizes.

<sup>3</sup><https://ai.meta.com/>

<sup>4</sup><https://www.maritaca.ai/>

However, it is known that Sabiá, a monolingual Large Language Model, was introduced in April 2023 with a primary focus on the Portuguese language. Notably, research conducted by Pires et al., 2023 exemplifies the significant and favorable impact of pretraining Sabiá specifically in the target language on models that have previously undergone extensive training on diverse corpora. Lastly;

- GPT, referenced as GPT-3 (OpenAI, 2020) or Generative Pre-trained Transformer 3, is a LLM introduced by OpenAI<sup>5</sup> in 2020. Notably, GPT-3 stands as one of the most extensive language models to date, equipped with an impressive 175 billion parameters, allowing it to tackle a diverse array of language-related tasks. It’s important to note that its knowledge extends only up to January 2022.

### Experiments

We chose the initial 1,010 sentences from the Porttinari-base, specifically the journalistic section of the Porttinari treebank. Among these, the first ten sentences were employed as a query example for the selected Large Language Model (LLM).

As an example, below, one may observe the prompt utilized to direct the LLM in performing UPoS tagging<sup>6</sup>:

Atuando como linguista e sem efetuar correções ou alterações no texto, faça a análise morfossintática das frases seguindo a anotação UD (Universal Dependencies) conforme os exemplos abaixo:

Entrada: A Odebrecht pagou 300 \% a mais pelo por o direito de explorar o aeroporto do de o Galeão .

Saída: A/DET Odebrecht/PROPN pagou/VERB 300/NUM %/SYM a/ADP mais/ADV pelo/None por/ADP o/DET direito/NOUN de/ADP explorar/VERB o/DET aeroporto/NOUN do/None de/ADP o/DET Galeão/PROPN ./PUNCT

<sup>5</sup><https://openai.com/>

<sup>6</sup>Prompt in English: ‘Acting as a linguist and without making any corrections or changes to the text, perform the morphosyntactic analysis of the sentences following the Universal Dependencies (UD) annotation as shown in the examples below.’

To enhance clarity and precision for the language model, we chose to represent prepositional contractions by separating the preposition and definite article. For instance, the word ‘pelo’ was retained as is, and then you added the preposition ‘por’ followed by the article ‘o’. These components were appropriately tagged as ‘ADP’ (adposition) and ‘DET’ (determiner), respectively. To avoid any potential confusion for the language model, the contracted word ‘pelo’ was tagged as ‘None’. This consistent approach was also applied to combined words, such as ‘de’ + ‘o’.

The output sentence was initially examined to ensure that the number of output tokens matched the input. If they did not match, the query was resubmitted for a maximum of ten iterations.

## 4 Results

In the context of a LLM, ‘temperature’ is a hyperparameter that governs the degree of randomness in the model’s responses. A higher temperature setting promotes greater diversity and randomness in the model’s responses, whereas a lower temperature setting leads to more deterministic and focused responses. The temperature parameter serves as a tool for adjusting the balance between randomness and determinism in the model’s generated outputs.

Initially, our objective was to fine-tune the temperature parameter to achieve the optimal balance between precision and recall, as measured by the F-measure, for each Large Language Model (LLM). This fine-tuning process was conducted exclusively on the initial 20 sentences. In Figure 1, we illustrate how variations in temperature impact the F-measure for each of the evaluated LLM.

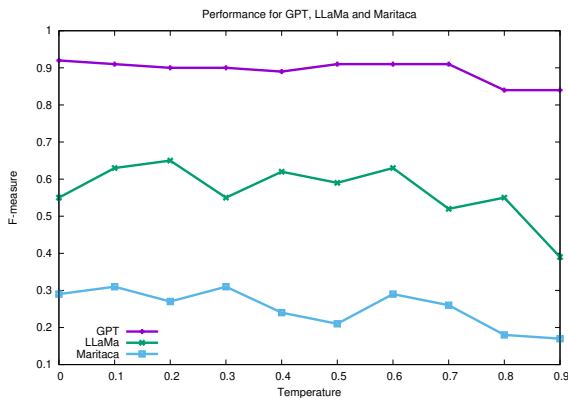


Figure 1: F-measure for GPT, LLaMa and Maritaca.

In Figure 1, a distinct advantage is evident for the GPT model, particularly when compared to

LLaMa-7B and Maritaca. The overall performance of the GPT exhibits only a slight decrease, primarily occurring at a temperature of 0.8.

The optimal temperature for LLaMa was identified at 0.2, while Maritaca exhibited its best performance at an even lower temperature of 0.1. Figure 1 demonstrates a variation of approximately  $\pm 1\%$  for GPT until the temperature reaches 0.8, after which performance begins to deteriorate, the optimal temperature found being 0.

We then repeated the experiment for each model in the remaining set of 1,000 sentences, using the optimal temperature found. We observed that in some cases, the language models (mainly Maritaca) made some incorrect taggings, for example, tags followed by some accentuation (‘VERB,’ , ‘VERB’). In these cases, we carry out a post-processing step making the necessary corrections to the identified tags.

In some cases, the models made some changes to the texts. In these cases, we analyzed the number of changes made, and if this number exceeded a threshold, we asked the model to analyze the sentence in question again, repeating this process a maximum of 10 times. In cases where the model was unable to properly process the sentences after this process, we marked all tags as ‘None’ and counted the error. There were 97 errors with Maritaca, 55 with LLaMa, and none with GPT.

Table 1 presents the values of precision, recall, F-measure, and support for each instance of a UPoS tag in the 1,000 analyzed sentences. Once more, it’s worth emphasizing the extensive utilization of ‘None’, which is not a component of the Universal Dependency PoS tagset. It is employed to label contractions and combinations of words. It’s important to take note of the tags with precision scores below 0.8. The first one, the AUX tag, pertains to auxiliary verbs such as ‘ter’, ‘haver’, ‘estar’, and ‘ser’. The second is the SCONJ tag, which stands for subordinating conjunctions. The issue of mislabeling was discussed by [Lopes et al., 2023](#).

Table 2 presents the outcomes for each UPoS tag after processing the same 1,000 sentences, now using LLaMa-7B. It’s evident that many of LLaMa’s results were in line with GPT, especially for tags such as ADP, CCONJ, and NOUN. However, for other tags, there was a notable discrepancy between the two models.

In our assessment, the results obtained from the Maritaca Large Language Model (LLM) in Table 3 do not exhibit a substantial discrepancy in compar-



TAG	Precision	Recall	F-measure	Support
ADJ	0.82	0.97	0.89	996
ADP	0.85	0.93	0.89	2,943
ADV	0.91	0.87	0.89	759
AUX	0.79	0.89	0.84	592
CCONJ	0.99	0.95	0.97	497
DET	0.91	0.94	0.93	2,880
INTJ	0.75	1.00	0.86	3
NOUN	0.98	0.96	0.97	3,757
NUM	0.96	0.87	0.91	364
None	0.83	0.58	0.68	1,208
PRON	0.81	0.78	0.80	771
PROPN	0.98	0.93	0.95	1,290
PUNCT	1.00	0.92	0.96	222
SCONJ	0.65	0.97	0.78	277
SYM	1.00	0.95	0.97	74
VERB	0.95	0.91	0.93	2,024
X	0.00	0.00	0.00	33
Macro Avg.	0.83	0.85	0.84	18,690
Accuracy			0.90	18,690

Table 1: GPT final experiment with 0.0 temperature.

TAG	Precision	Recall	F-measure	Support
ADJ	0.63	0.61	0.62	996
ADP	0.82	0.73	0.77	2,943
ADV	0.63	0.68	0.65	759
AUX	0.58	0.63	0.60	592
CCONJ	0.98	0.74	0.84	497
DET	0.74	0.81	0.77	2,880
INTJ	0.00	0.00	0.00	3
NOUN	0.92	0.69	0.79	3,757
NUM	0.58	0.70	0.63	364
None	0.20	0.45	0.28	1,208
PRON	0.59	0.39	0.47	771
PROPN	0.73	0.78	0.75	1,290
PUNCT	0.82	0.40	0.53	222
SCONJ	0.49	0.28	0.35	277
SYM	0.98	0.70	0.82	74
VERB	0.84	0.81	0.82	2,024
X	0.00	0.00	0.00	33
Macro Avg.	0.62	0.55	0.57	18,690
Accuracy			0.69	18,690

Table 2: LLaMA final experiment with 0.2 temperature.

ison to the LLaMA LLM. The Maritaca LLM displayed notably low values for CCONJ and SCONJ, which undeniably had a negative impact on the overall performance. On the bright side, tags such as ADJ, INTJ, VERB, and PRON showcased values that were comparable and promising.

We also noticed that all models apply tags that do not belong to the domain of Universal Dependencies. Notably, the models can interpret punctuation marks as synonyms for UD labels (e.g., GPT with labels ‘)’ and ‘(’), as well as LLaMA with the label ‘VERB’). Situations like these were addressed in post-processing and counted as correct annotations. However, the Maritaca model listed 93 labels as possible morphosyntactic markers for UD, rather than the expected 17. Labels such as ‘BE’, ‘BEAR’, ‘BEZ’, ‘EXISTE’, ‘HAS’, and ‘MONTH’ resulted in errors.

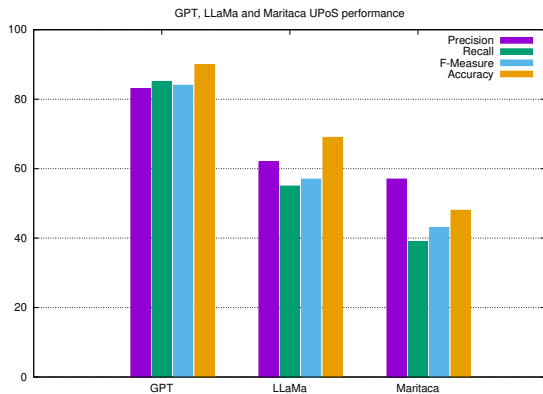


Figure 2: Precision, Recall, F-measure, and Accuracy for GPT, LLaMA and Maritaca.

Finally, Figure 2 offers a summary of the key performance metrics related to the data generated by the three assessed Large Language Models (LLMs). The figure depicts an improvement of around 10 percentage points for F-Measure and accuracy from GPT to LLaMa, as well as a comparable significant increase between LLaMA and Maritaca. Additionally, it highlights a substantial convergence in precision between the LLaMa and Maritaca models.

## 5 Final remarks

It is undeniable that LLMs caused a great revolution, bringing AI into the daily lives of many people. They also provided new ways to process, classify, and extract information through the use of prompts, which facilitated the development of advanced processing using natural language.

We conducted an analysis of the results of part-of-speech (UPoS) tagging in texts written in Brazilian Portuguese using three distinct large language models (LLMs): GPT-3, LLaMA-7B, and Maritaca. We were meticulous in selecting the Porttinar-base treebank, which was released after the aforementioned language models, to reduce the likelihood of these LLMs having the same annotated treebanks as knowledge bases.

GPT-3, a multilingual LLM and purportedly the largest among the three, achieved the highest performance metrics. It was followed by the LLaMA, the LLM from Meta Platforms, Inc., which exhibited a notable disparity in comparison to GPT-3. Lastly, the Maritaca API, which uses

TAG	Precision	Recall	F-measure	Support
ADJ	0.75	0.60	0.67	996
ADP	0.66	0.45	0.53	2,943
ADV	0.48	0.62	0.54	759
AUX	0.50	0.25	0.34	592
CCONJ	0.20	0.09	0.12	497
DET	0.57	0.46	0.51	2,880
INTJ	0.67	0.67	0.67	3
NOUN	0.87	0.66	0.75	3,757
NUM	0.52	0.68	0.59	364
None	0.05	0.28	0.09	1,208
PRON	0.69	0.22	0.34	771
PROPN	0.91	0.42	0.57	1,290
PUNCT	0.75	0.11	0.19	222
SCONJ	0.23	0.01	0.02	277
SYM	1.00	0.45	0.62	74
VERB	0.83	0.61	0.70	2,024
X	0.00	0.00	0.00	33
Macro Avg.	0.57	0.39	0.43	18,690
Accuracy			0.48	18,690

Table 3: Maritaca final experiment with 0.1 temperature.

an undisclosed language model, displayed a similar level of deviation from LLaMA as it did from GPT-3.

The experiments were conducted using a few-shot approach, beginning with exemplifying UPoS tagging with ten annotated sentences before requesting the UPoS task for the eleventh sentence. The GPT-3 API responded with a tagset that closely approximated the Universal Dependencies (UD) tagset. We also encountered some delays and cut-offs when making API calls. The LLaMA model was the most straightforward to execute since it could be downloaded and run locally. The returned tagset was also similar to the UD tagset.

These results were very positive, especially if we take into account that they were obtained using only 20 annotated examples, something that would be unfeasible with traditional machine learning algorithms. However, certain tags presented very low F-measures, such as ‘None,’ ‘NUM,’ ‘PRON,’ and ‘SCONJ,’ which could be attributed to the disparity in model size between LLaMA (7B parameters) and GPT-3 (175B parameters). On the other hand, the Maritaca API exhibited the poorest results. Maritaca returned a PoS tagset consisting of 93 tags, which we believe is the primary reason for its lower performance in PoS tagging.

Annotating data for training AI algorithms is normally expensive, in this case this annotation is even more difficult to carry out, as it requires linguistic knowledge. Another point to highlight is that LLMs are constantly improving, indicating

that even better results may be obtained in a near future.

In conclusion, our findings suggest that specific Large Language Models (LLMs) can function as initial Universal Dependency Part-of-Speech (UPoS) taggers for low-resource languages like Portuguese, especially when supplemented with human review. This proves beneficial even in cases where Universal Dependency (UD) parsers, like PassPort by Zilio et al., 2018, produce comparable outcomes.

## Acknowledgement

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI – <http://c4ai.inova.usp.br/>), with support by the São Paulo Research Foundation (FAPESP Grant #2019/07665-4) and by the IBM Corporation.

## References

- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Prompting language models for linguistic structure. *arXiv preprint arXiv:2211.07830*.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.
- Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25.
- Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. *The Dawn of the Portinari Multigenre Treebank: Introducing its Journalistic Portion*. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.
- Lane Green. 2017. Technology Quarterly: Finding a Voice. *The Economist*.
- Go Inoue, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Joint prediction of morphosyntactic categories for fine-grained Arabic part-of-speech tagging exploiting tag dictionary information. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 421–431.

- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGpt: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. *ArXiv, abs/2303.03953*.
- Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. [XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.
- Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. [Linguistic Knowledge and Transferability of Contextual Representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.
- Lucelene Lopes, Magali Duran, and Thiago Alexandre Pardo. 2023. [Verifica-UD: a Verifier for Universal Dependencies Annotation for Portuguese](#). In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 443–452, Porto Alegre, RS, Brasil. SBC.
- Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamel Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.
- Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.
- OpenAI. 2020. [GPT-3: Language Models for Few-Shot Learning](#). *OpenAI*.
- Jianquan Ouyang and Mengen Fu. 2022. Improving machine reading comprehension with multi-task learning and self-training. *Mathematics*, 10(3):310.
- Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.
- Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Pradeep Kumar Roy, Sunil Saumya, Jyoti Prakash Singh, Snehasish Banerjee, and Adnan Gutub. 2023. Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions on Intelligence Technology*, 8(1):95–117.
- Devansh Shah, Arun Singh, and Sudha Shanker Prasad. 2022. [Sentimental Analysis Using Supervised Learning Algorithms](#). In *2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, pages 1–6.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. [BERT Rediscovered the Classical NLP Pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, Jun Xie, and Rong Jin. 2022. Learning to generalize to more: Continuous semantic augmentation for neural machine translation. *arXiv preprint arXiv:2204.06812*.
- Leonardo Zilio, Rodrigo Wilkens, and Cédric Fairon. 2018. PassPort: a dependency parsing model for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 479–489. Springer.