

PROPOR 2024

**Proceedings of the 16th International Conference on Computational
Processing of the Portuguese Language, PROPOR 2024 - vol. 2**

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

ISBN 979-8-89176-062-2

PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays

Organizers:

Rafael Ferreira Mello, Universidade Federal Rural de Pernambuco/CESAR (Brazil)
Hilário Tomaz Alves de Oliveira, Instituto Federal do Espírito Santo (Brazil)
Moésio Wenceslau, Universidade Federal Rural de Pernambuco (Brazil)
Hyan Batista, Universidade Federal Rural de Pernambuco (Brazil)
Thiago Cordeiro, Universidade Federal de Alagoas (Brazil)
Ig Ibert Bittencourt, Universidade Federal de Alagoas / Harvard University (Brazil/UK)
Seiji Isotani, Universidade de São Paulo / Harvard University (Brazil/UK)

First Workshop on NLP for Indigenous Languages of Lusophone Countries

Organizers:

Aline Paes, Federal Fluminense University (Brazil)
Aline Villavicencio, University of Sheffield and University of Exeter (UK)
Claudio Pinhanez, IBM Research Brazil and University of São Paulo (Brazil)
Paulo Rodrigo Cavalin, IBM Research Brazil (Brazil)
Edward Gow-Smith, University of Sheffield (UK)

Program Committee:

Arnaldo Candido Junior, UTFPR (Brazil)
Leonel Figueiredo de Alencar, UFC (Brazil)
Lílian Teixeira De Sousa, UNICAMP (Brazil)
Marcelo Finger, USP (Brazil)
Marcely Zanon Boito, NAVER Labs
Rodrigo Wilkens, KU Leuven (Germany)

Third Workshop on Digital Humanities and Natural Language Processing

Organizers:

Maria José Bocorny Finatto, Universidade Federal do Rio Grande do Sul, PPG-LETRAS (Brazil)
Leonardo Zilio, Friedrich-Alexander-Universität Erlangen-Nürnberg, CCL (Germany)
Diana Santos, Faculty of Humanities, Linguatca/University of Oslo (Norway)
Renata Vieira, CIDEHUS, Évora University (Portugal)
Valeria de Paiva, Topos Institute (USA)

Programme Committee:

Álvaro Iriarte Sanromán, Minho University (Portugal)
Cassia Trojahn, Toulouse University (France)
Daniel Alves, NOVA University of Lisbon (Portugal)
David Semedo, Nova School of Science and Technology (Portugal)

Denise Nauderer Hogetop, Arquivo Público do RS, (Brazil)
Emanoel Pires, State University of Maranhão (Brazil)
Fátima Farrica, Évora University (Portugal)
Fernanda Olival, Évora University (Portugal)
Helena Freire Cameron, Polytechnic Institute of Portalegre (Portugal)
Idaete da Silva Dias, Minho University (Portugal)
Leandro Krug Wives, UFRGS (Brazil)
Paulo Quaresma, Évora University (Portugal)
Raquel Amaro, NOVA University of Lisbon (Portugal)
Rebeca Schumacher, Linguateca (Portugal)
Sandro Marengo Drumond, UFS (Brazil)
Suemi Higuchi, Getúlio Vargas Foundation/PUC-Rio (Brazil)

Demos

Demo Chairs:

Iria de-Dios-Flores, Universitat Pompeu Fabra (Spain)
Marlo Souza, Universidade Federal da Bahia (Brazil)

Programme Committee:

Clarissa Castellã Xavier, SiDi (Brazil)
Evandro Fonseca, Blip (Brazil)
Iria de-Dios-Flores, Universitat Pompeu Fabra (Spain)
Larissa Astrogildo de Freitas, Universidade Federal de Pelotas (Brazil)
Marlo Souza, Universidade Federal da Bahia (Brazil)
Susana Sotelo Docío, Universidade de Santiago de Compostela (Spain)

Table of Contents

PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays

PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays	2
<i>Rafael Ferreira Mello, Hilário Oliveira, Moésio Wenceslau, Hyan Batista, Thiago Cordeiro, Ig Ibert Bittencourt and Seiji Isotanif</i>	
AESVoting: Automatic Essay Scoring with Bert and Voting Classifiers	7
<i>Tiago Barbosa de Lima, Elyda Freitas and Valmir Macario</i>	
PiLN at PROPOR: A BERT-Based Strategy for Grading Narrative Essays	11
<i>Rogério F. de Sousa, Jeziel C. Marinho, Francisco A. R. Neto, Rafael T. Anchiêta and Raimundo S. Moura</i>	
Exploring the Automated Scoring of Narrative Essays in Brazilian Portuguese using Transformer Models	15
<i>Eugénio Ribeiro, Nuno Mamede and Jorge Baptista</i>	

First Workshop on NLP for Indigenous Languages of Lusophone Countries (ILLC-NLP)

Building a Language-Learning Game for Brazilian Indigenous Languages: A Case of Study	20
<i>Gustavo Polleti</i>	
Computational Model for Yoruba Aroko Communication System	25
<i>Adéwuyì Adétáyò Adégbíté and Odétúnjí Àjàdí Odéjobí</i>	
Human Evaluation of the Usefulness of Fine-Tuned English Translators for the Guarani Mbya and Nheengatu Indigenous Languages	34
<i>Claudio Pinhanez, Paulo Cavalin and Julio Nogima</i>	
A Universal Dependencies Treebank for Nheengatu	39
<i>Leonel Figueiredo de Alencar</i>	
Network-based Approach for Stopwords Detection	57
<i>Felermimo D. M. A. Ali, Gabriel de Jesus, Henrique Lopes Cardoso, Sérgio Nunes and Rui Sousa-Silva</i>	
Grammar Induction for Brazilian Indigenous Languages	66
<i>Diego Pedro Gonçalves da Silva and Thiago Alexandre Salgueiro Pardo</i>	
NLP Tools for African Languages: Overview	75
<i>Joaquim Mussandi and Andreas Wichert</i>	

Third Workshop on Digital Humanities and Natural Language Processing (3rdDHandNLP)

Can rules still beat neural networks? The case of automatic normalisation for 18th-century Portuguese texts	86
<i>Leonardo Zilio, Rafaela R. Lazzari and Maria José B. Finatto</i>	
Revealing Public Opinion Sentiment Landscape: Eurovision Song Contest Sentiment Analysis	96
<i>Klara Kozolic, Gaurish Thakkar and Nives Mikelic Preradovic</i>	
Could Style Help Plagiarism Detection? - A Sample-based Quantitative Study of Correlation between Style Specifics and Plagiarism	106
<i>Adile Uka and Maria Berger</i>	
Authorship attribution in translated texts: a stylometric approach to translator style	112
<i>Ana Pagano, Carlos Perini, Evandro Cunha and Adriana Pagano</i>	
Support Verb Constructions in Medieval Portuguese: Evidence from the CTA Corpus	121
<i>Maria Inês Bico, Esperança Cardeira, Jorge Baptista and Fernando Baptista</i>	

Semantic Exploration of Textual Analogies for Advanced Plagiarism Detection	133
<i>Elyah Frisco Andriantsialo, Volatiana Marielle Ratianantitra and Thomas Mahatody</i>	
Creating datasets for emergent contact languages preservation	137
<i>Dalmo Buzato and Átila Vital</i>	
Psychoanalytic Studies in the Digital Humanities: Employing Topic Modeling with an LLM to Decode Dreams During the Brazilian Pandemic	144
<i>João Pedro Campos, Natalia Resende, Ricardo de Souza and Gilson Iannini</i>	
Decoding Sentiments about Migration in Portuguese Political Manifestos (2011, 2015, 2019)	152
<i>Erik Bran Marino, Renata Vieira, Jesus Manuel Benitez Baleato, Ana Sofia Ribeiro and Katarina Laken</i>	
Analysing entity distribution in an annotated 18th-century historical source	163
<i>Daniel De Los Reyes, Renata Vieira, Fernanda Olival, Helena Freire Cameron and Fátima Farrica</i>	
Roda Viva boundaries: an overview of an audio-transcription corpus	168
<i>Isaac Souza de Miranda Jr., Gabriela Wick-Pedro, Cláudia Dias de Barros and Oto Vale</i>	
Demos	
GiDi: A Virtual Assistant for Screening Protocols at Home	174
<i>Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, María del Carmen López-Pérez and Ignacio Novo-Veleiro</i>	
FazGame: A Game Based Platform that Uses Artificial Intelligence to Help Students to Improve Brazilian Portuguese Writing Skills	178
<i>Jéssica Soares Dos Santos, Gabriel Coelho, Sidney Melo, Oniram Atila and Carla Zeltzer</i>	
Indexing Portuguese NLP Resources with PT-Pump-Up	182
<i>Rúben Almeida, Ricardo Campos, Alípio Jorge and Sérgio Nunes</i>	
plain X –AI Supported Multilingual Video Workflow Platform	186
<i>Carlos Amaral, Catarina Lagrifa, Mirko Lorenz, Peggy van der Kreeft and Tiago Veiga</i>	
Perfil Público: Automatic Generation and Visualization of Author Profiles for Digital News Media	190
<i>Nuno Guimarães, Ricardo Campos and Alípio Jorge</i>	
Exploring Open Information Extraction for the Portuguese language: An integrated monolithic approach in Cloud environment	194
<i>Augusto Barreto and Daniela Claro</i>	
Blip Copilot: a smart conversational assistant	198
<i>Evandro Fonseca, Tayane Soares, Dyovana Baptista, Rogers Damas and Lucas Avanco</i>	
GalicianNeural Machine Translation System	201
<i>Sofía García González</i>	
Nós-TTS: a Web User Interface for Galician Text-to-Speech	204
<i>Carmen Magariños, Alp Öktem, Antonio Moscoso Sánchez, Marta Vázquez Abuín, Noelia García Díaz, Adina Ioana Vladu, Elisa Fernández Rei and María Baqueiro Vidal</i>	
Autopilot: a smart sales assistant	208
<i>Amanda Oliveira, João Alvarenga, Evandro Fonseca and William Colen</i>	

**PROPOR'24 Competition on Automatic Essay
Scoring of Portuguese Narrative Essays**

PROPOR’24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays

Rafael Ferreira Mello^{a,b} Hilário Oliveira^c Moésio Wenceslau^a Hyan Batista^a
Thiago Cordeiro^d Ig Ibert Bittencourt^{d,e} and Seiji Isotani^{f,e}

^aUniversidade Federal Rural de Pernambuco ^bCESAR

^cInstituto Federal do Espírito Santo ^dUniversidade Federal de Alagoas

^eHarvard University ^fUniversidade de São Paulo

{rafael.mello, moesio.wenceslau}@ufrpe.br hilario.oliveira@ifes.edu.br

Abstract

The PROPOR’24 Competition on Automatic Essay Scoring (AES) of Portuguese Narrative Essays evaluated the performance of four participating systems and three baselines for the task of estimating the individual score of four competencies (Formal Register, Narrative Rhetorical Structure, Thematic Coherence, and Cohesion) in narrative essays written by mid-school students in Brazil. The corpus comprises 1,235 essays, divided into train, validation and test sets, and the competitors were evaluated using Cohen’s Kappa coefficient and the weighted F1-score. Most submitted systems and baselines leveraged pre-trained language models, particularly Albertina and BERTimbau, demonstrating fair to moderate agreement with human evaluator scores based on the Kappa coefficient. These results highlight the challenge of AES for Portuguese narrative essays while demonstrating the promise of pre-trained language models for future improvement.

1 Introduction

Integrating Artificial Intelligence (AI) into education presents an opportunity to transform the teaching and learning process, offering innovative solutions that enhance efficiency, personalization, and accessibility (Chen et al., 2020). AI holds the potential to impact various educational facets, ranging from content adaptation based on student profiles to delivering personalized, real-time feedback (Cavalcanti et al., 2021). Among the promising applications of AI in education, automatic scoring of textual production, especially essays, stands out (Ferreira-Mello et al., 2019; Chen et al., 2020). AI algorithms can automatically analyze and assess various aspects of an essay, including grammar, cohesion, coherence, argumentative structure, and originality.

Automatic essay scoring (AES) is the task of automatically assigning a grade score to an essay

based on a predefined grading rubric (Ramesh and Sanampudi, 2022). The manual correction of essays written by students is a labor-intensive process that places significant demands on teachers and evaluators in terms of time and effort (Costa et al., 2020; de Lima et al., 2023). Moreover, the assessment procedures may be susceptible to individual examiners’ personal biases regarding a given topic, resulting in inconsistencies in their evaluations. Developing computer systems capable of automatically evaluating essays based on established criteria can help deal with time demands and consistency challenges in evaluation (Ferreira-Mello et al., 2019). These systems can assist teachers in the classroom by enhancing formative feedback strategies, enabling them to focus on specific areas of writing that require improvement among their students (Ramesh and Sanampudi, 2022).

In recent years, AES systems have experienced advances, particularly in extensively studied languages like English (de Lima et al., 2023). However, progress in low-resource languages like Portuguese still needs to be improved. Most research on Portuguese AES systems concentrates on dissertative-argumentative essays within the high school context (Oliveira et al., 2023a,b), with few studies exploring other domains, such as narrative textual productions commonly utilized in early basic education (Filho et al., 2023).

This shared task aims to contribute to the progression of Portuguese AES systems. In particular, the emphasis is on assessing narrative essays written in Portuguese by students within the Brazilian basic education system. The evaluation was carried out using a corpus comprising 1,235 narrative essays authored by primary school students. Human examiners assessed each essay based on four correction criteria: textual cohesion, thematic coherence, textual typology, and spelling and grammatical errors. The competition involved the participation of four competitors from Brazil and Portu-

gal. Additionally, three commonly used approaches from the literature served as baselines for comparison purposes. The competitors and baselines assessment relied on two automatic evaluation measures: Cohen’s Kappa coefficient and weighted F-1 score.

2 Dataset Description

The dataset used in this competition contains 1,235 essays written by students in Brazil’s 5th to 9th year of public schools. The students were instructed to write a narrative essay based on a pre-defined prompt given by the teachers. All essays were manually transcribed and anonymized by teachers selected based on their competence with the students in the selected grades.

Afterward, the essays were analyzed by two human evaluators who assessed different aspects of each essay using a pre-defined correction rubric. Given the complexity and subjectivity involved in evaluating this process, disagreements between annotators are common. To mitigate this problem, a third human evaluator with more experience in the task was included to join the annotation team and solve the divergences with the first two annotators. The rubric provides instructive guidance for educators to consider four required competencies:

- **Formal Register:** Appropriate use of the Portuguese language. Aspects such as misspelling words, inadequate use of nominal/verbal agreement and nominal/verbal re-gency, and inappropriate usage of punctuation symbols are considered.
- **Thematic Coherence:** Adequate understanding of the text production proposal and its development associated with knowledge from different areas, according to the requested proposal, i.e., the plausibility of the text developed concerning the motivating text.
- **Narrative Rhetorical Structure:** Conformity of the text production proposal regarding a Narrative textual typology, articulating ideas, facts, and information in a sequenced and logical way, presenting the constituent elements of this type of textual structure: narra-tor, place/space, temporal organization, multi-ple or single characters performing actions.
- **Textual Cohesion:** Correct use of linguis-tic mechanisms to interconnect text elements,

such as words, sentences, and paragraphs.

For each of the four previous competencies, the human evaluators assigned a level ranging from *I* to *V*, with **Level I** demonstrating a complete lack of knowledge in the competency domain and **Level V** a complete mastery of the competency.

Figure 1 shows the essay distribution of the full corpus by level for each evaluated competency. The final dataset was divided into three subsets with the following division: 60% (740) for training, 10% (125) for validation, and 30% (370).

The Cohen’s Kappa agreement score between annotators 1 and 2 for the four competencies was 0.2475. The overall agreement between the first and third annotators and the second and third anno-tators was 0.5405 and 0.5650, respectively. Despite the low level of agreement between the first two an-notators, it was observed that most of the disagree-ments were at adjacent levels. For instance, an an-notator assigned level III for the essay, whereas the other set level II or IV. This divergence is consid-ered normal in assessments of textual productions, given the subjectivity of the items in the correction rubrics. The final dataset is available at: <https://doi.org/10.34740/kaggle/ds/4464018>.

3 Competition Participants

This section describes the approaches adopted by the participants in the competition. Five teams ini-tially registered for the competition, but one team did not submit their system’s source code. Conse-quently, evaluating this system on the private test dataset was not feasible. As a result, the competi-tion proceeded with the following four participants:

AESVoting This approach proposes an ensem-ble of three classifiers (Random Forest, Gaussian Naive Bayes, and Logistic Regression) with a vot-ing/majority rule. Specifically, four separate mod-els are trained, each dedicated to one competency. As input, these models receive an encoded mul-tidimensional contextual representation of the es-say extracted using the BERTimbau (Souza et al., 2020). Additionally, during training, the SMOTE technique (Chawla et al., 2002) is employed to ad-dress the effects of imbalanced data.

INESC-ID This approach adopted different pre-trained language models based on the Trans-formers architecture for the Portuguese language. The authors investigated different versions of Al-bertina PT-* (Rodrigues et al., 2023) model and the BERTimbau-large architecture (Souza et al.,

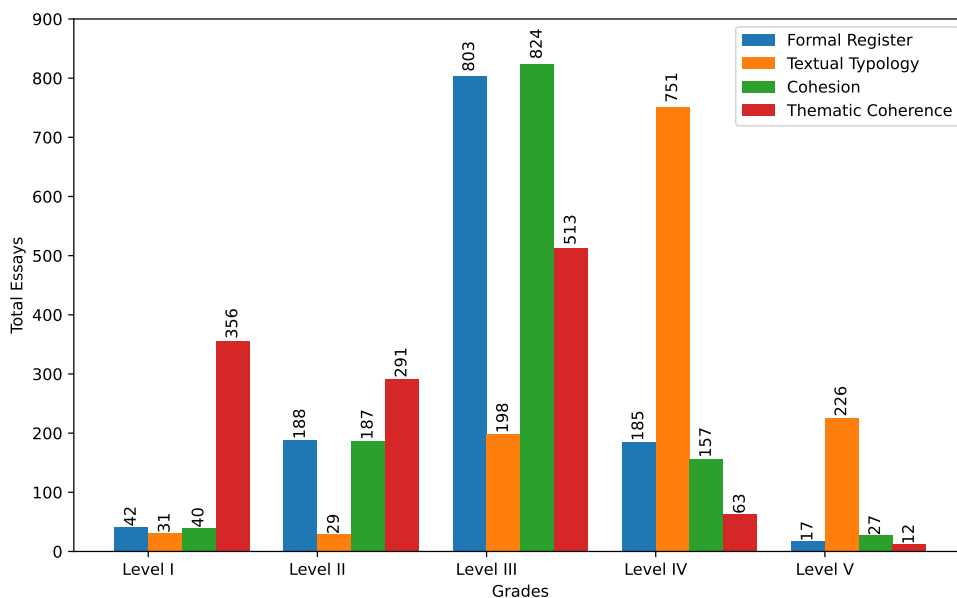


Figure 1: Essays distribution by level for each assessed competency on the complete corpus.

2020). The competitors performed several fine-tuning training steps to estimate the scoring grades of the competencies as a classification problem. The best model for each competency was selected based on the validation set.

PiLN This participant system explored the BERTimbau (base and large) (Souza et al., 2020) with an additional linear layer to predict the scores of the competencies as a regression problem. As the scores predicted by the regression model are continuous values, as post-processing, the estimated scores were rounded to the nearest exact grades (1 to 5). As input, the proposed model receives the essay and the motivating text. The authors tested several configurations and achieved the results using the BERTimbau-large architecture.

Ocean Team In this approach, a two-stage encoding strategy was used as input to various classical machine learning classifiers. First, word embeddings were generated using the BERTimbau model (Souza et al., 2020). Afterward, the Term Frequency – Inverse Document Frequency (TF-IDF) measure was applied to capture term importance. Both representations were used as input to train different classifiers, including Random Forest, XGBoost, Support Vector Machine, AdaBoost, and an ensemble of Extremely Randomized Trees. Four separate models were trained for each competency.

4 Baseline Methods

In addition to the competition participants, the following approaches were evaluated as baselines. These approaches modeled the AES task as a supervised Machine Learning (ML) classification problem and were selected because they present encouraging results in the literature for the AES task in Portuguese in high school essays (Oliveira et al., 2023a,b).

TF-IDF + ML This approach relies on the conventional text representation using the TF-IDF measure. The initial step involved pre-processing the essays by segmenting the text into words, eliminating punctuation symbols, and discarding words with a single character. The words retained after pre-processing from the training corpus were used to build a vocabulary. Each essay was then represented by a vector containing the TF-IDF value for each word in the vocabulary. Finally, these essays were used to train traditional ML algorithms to estimate scores for each competency. The performance of several algorithms available in the scikit-learn¹, eXtreme Gradient Boosting (XGBoost)² and Light Gradient Boosting Machine (LGBM)³, were examined. The algorithm with the best performance for each competence was used as a baseline for

¹<https://scikit-learn.org/>

²<https://github.com/dmlc/xgboost/>

³<https://github.com/Microsoft/LightGBM/>

comparison with competing systems.

BERT Embedding + ML This approach involves encoding essays using a multidimensional contextual representation (contextual word embeddings). The representations were obtained through the pre-trained language model of the BERTimbau-base architecture (Souza et al., 2020). The essays were truncated to a maximum of five hundred and twelve (512) tokens, the maximum token sequence length of the BERT model. Each essay is then represented by a vector of seven hundred and sixty-eight (768) values⁴, enabling the exploration of syntactic and semantic patterns within the essays. The encoded essays were employed to train traditional ML algorithms for predicting scores across assessed competencies. Then, similar to the TF-IDF approach, several ML models were analyzed, and the top-performing model for each competence was selected as the baseline.

BERT Classifier In this approach, the BERTimbau-base architecture (Souza et al., 2020) was employed, along with an additional dense linear layer, to estimate the scores of the essays' competencies. The model underwent fine-tuning training using the AdamW optimizer with decoupled weight decay and an initial learning rate of $5 * 10^{-5}$. The fine-tuning process was conducted over five training epochs. A BERTimbau model was fine-tuned for each competency considered in the competition.

5 Evaluation Measures

Two automatic evaluation measures commonly used to evaluate AES systems were adopted to assess the baselines and participating competitors (Ramesh and Sanampudi, 2022; de Lima et al., 2023).

Cohen's Kappa coefficient It is a statistical measure used to evaluate the agreement or reliability between two or more annotators or algorithms regarding classifying items into mutually exclusive categories (Cohen, 1960). The Kappa coefficient ranges from -1 to 1 , where 1 indicates perfect agreement between, 0 indicates agreement that could be achieved by randomly guessing, and negative values suggest disagreement beyond guessing. The other Kappa values can be interpreted as follows (Landis and Koch, 1977): (i) values higher than 0.81 are considered indicative of a very high level of agreement, (ii) values between 0.61 and

0.80 suggest a good level of agreement, (iii) values between 0.41 and 0.6 indicate moderate agreement, (iv) values between 0.21 and 0.4 indicate fair (reasonable) agreement and (v) values below 0.21 indicate poor agreement.

Weighted F1-score This evaluation metric is commonly used in machine learning classification problems, especially when significant class imbalances exist. It combines precision and recall metrics into a single score reflecting a model's precision and ability to correctly identify positive examples of each class. This measure computes the weighted average of the traditional F1-score for each class, with weights assigned based on the frequency of each class in the dataset. Therefore, less represented classes have less influence on the overall score, while more represented classes carry higher weight. This metric is particularly valuable in unbalanced multi-class classification scenarios, where simple averaging of the F1-score would not adequately represent the model's effectiveness across all classes. The closer the weighted F1-score value is to 1 , the better the model's performance in classifying the different categories.

6 Results

A hold-out strategy was employed to assess both participant competitors and baselines. The dataset was partitioned into three subsets: **training** (60%, 740 essays), **validation** (10%, 125 essays), and **test** (30%, 370 essays) sets. The training and validation sets were provided to competitors to develop their systems, while the test set remained reserved for the final evaluation.

Table 1 shows the results on the test set for each competency based on Cohen's Kappa coefficient and the weighted F-1 score. The competitors shared the source code, which was then used to train and evaluate the system on the blind test set.

The first point to highlight is that the performance of the three baselines remained competitive with the participating systems across all competencies. Specifically, the **TF-IDF + ML** approach for Narrative Rhetorical Structure and the **BERT Embedding + ML** method for cohesion yielded the best results based on the Kappa coefficient values. The **BERT Classifier** demonstrated superior performance regarding the weighted F-1 score in the two previous competitions. The **BERT Embedding + ML** also achieved the best results on the Formal Register competence.

⁴Default representation size defined.

Approach	Cohesion		Formal Register		Narrative Rhetorical Structure		Thematic Coherence	
	Kappa	Weighted F1	Kappa	Weighted F1	Kappa	Weighted F1	Kappa	Weighted F1
AESVoting	0.192	0.567	0.274	0.593	0.219	0.513	0.355	0.552
INESC-ID	0.356	0.691	0.375	0.668	0.284	0.607	0.548	0.666
PiLN	0.366	0.692	0.414	0.702	0.250	0.616	0.548	0.679
Ocean Team	0.225	0.647	0.237	0.640	0.187	0.591	0.485	0.621
TF-IDF + ML	0.281	0.650	0.280	0.652	0.286	0.623	0.526	0.667
BERT Embedding + ML	0.367	0.701	0.407	0.708	0.232	0.606	0.448	0.604
BERT Classifier	0.355	0.702	0.413	0.704	0.283	0.626	0.495	0.643

Table 1: Results of evaluations of participating competitors and baselines in the private test set.

The PiLN achieved the top performance for evaluation measures in Thematic Coherence and the best Kappa coefficient for the Formal Register competency. Also, INESC-ID attained an identical Kappa value to PiLN for Thematic Coherence.

The best outcomes exhibit a reasonable to moderate Kappa coefficient compared to the grades assigned by human evaluators for each competency. These findings show that there is still much room for future progress and highlight the complexity of the task. It is particularly noteworthy the superior performance of participant systems and baselines integrating pre-trained language models such as Albertina and BERTimbau. Such results suggest that leveraging these models holds promise for developing more precise Portuguese AES systems.

References

- Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.
- N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. **Smote: Synthetic minority over-sampling technique**. *Journal of Artificial Intelligence Research*, 16:321–357.
- Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Luciana Costa, Elaine Oliveira, and Alberto Castro Júnior. 2020. **Corretor automático de redações em língua portuguesa: um mapeamento sistemático de literatura**. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1403–1412, Porto Alegre, RS, Brasil. SBC.
- Tiago Barbosa de Lima, Ingrid Luana Almeida da Silva, Elyda Laisa Soares Xavier Freitas, and Rafael Ferreira Mello. 2023. Avaliação automática de redação: Uma revisão sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.
- Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.
- Moésio Silva Filho, André Nascimento, Pérciles Miranda, Luiz Rodrigues, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Rafael Mello. 2023. **Automated formal register scoring of student narrative essays written in portuguese**. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 1–11, Porto Alegre, RS, Brasil. SBC.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023a. Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.
- Hilário Oliveira, Rafael Ferreira Mello, Pérciles Miranda, Bruno Alexandre, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotani. 2023b. Classificação ou regressão? avaliando coesão textual em redações no contexto do enem. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1226–1237. SBC.
- Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.
- João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. **Advancing neural encoding of portuguese with transformer albertina pt-***.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

AESVoting: Automatic Essay Scoring with Bert and Voting Classifiers

Tiago Barbosa de Lima

Departamento de Computação –
Universidade Federal Rural
de Pernambuco

Rua Dom Manuel de Medeiros, s/n,
Dois Irmãos - Recife – PE – Brazil
tiago.blima@ufrpe.br

Elyda Freitas

Departamento de
Sistemas de Informação
Universidade de Pernambuco

Caruaru, PE – Brazil
elyda.freitas@upe.br

Valmir Macario

Departamento de Computação –
Universidade Federal Rural
de Pernambuco

Rua Dom Manuel de Medeiros, s/n,
Dois Irmãos - Recife – PE – Brazil
valmir.macario@upe.br

Abstract

In this work, we explore the use of pre-trained models to extract features to automatic essay-scoring tasks using Multinomial Logistic Regression, Random Forest and Gaussian Naïve Bayes. We further utilise instance oversampling to mitigate the scarcity of instances to some classes. The results suggest that the addition of synthetic examples turns the model biased and worsens the final result. Therefore, we make use of a voting classifier to mitigate bias which improves the final overall result.

1 Introduction

In Brazil, the National High School Exam (ENEM) work as an evaluation entry exam for many universities (de Lima et al., 2023). One of the requirements is to write an essay in a dissertative argumentative style as proven by academic proficiency (de Lima et al., 2023). The exam produces a demand for the evaluation of millions of essays which is a manually costly operation every year (de Lima et al., 2023). Besides, Automatic Essay scoring (AES) aims to support this task by automatically attributing a score to a textual production often written by a student (de Lima et al., 2023; Sharma and Goyal, 2020; da Silva Filho et al., 2023). Then, several works propose the automatic correction of those essay styles in the literature using different means like pre-trained machine learning models (de Lima et al., 2023; Akio Matsuoka, 2023). Algorithms used for AES such as Logistic Regression, Naïve Bayes and others rely on feature extraction systems to be used as classifiers in several settings (Rudner and Liang, 2002; Kumar et al., 2019a; Sharma and Goyal, 2020; Ludwig et al., 2021; Kumar et al., 2019b). One method to achieve AES is to extract features automatically using a pre-trained model. The work (Beseiso and Alzahrani, 2020) combines manually generated features with model extract features from BERT and Long Short Term Memory to improve AES.

Bidirectional Encoder Representation (BERT) is a widely used pre-trained encoder-only model in several tasks like sentiment analysis, question answer and others (Souza et al., 2020). The work (Akio Matsuoka, 2023) used the Portuguese version of BERT known as BERTimbau developed by (Souza et al., 2020), to automatically score ENEM essays in different categories such as adherence to them, cohesion and coherence, grammatical correction and others obtaining state of the art results.

Despite all the advantages of AES, current methods developed for the Portuguese language in some cases focus on the dissertative argumentative style. On the other hand, works such as (da Silva Filho et al., 2023) evaluate the narrative essays produced by elementary school students in the aspect of a formal register that measures the correct use of linguistics rules for the students. The results show that is possible to achieve good agreement with one of the annotators showing the potential of the application.

Therefore, we explore the use of BERTimbau to extract features to automatically classify students' essays in narrative written style. We use the features from the BERT model as inputs to Multinomial Logistic Regression, Random Forest and Gaussian Naïve Bayes.

2 Materials and Methods

We used 1,235 essays from elementary school students in Brazil proposed by the PROPOR'24. Each essay is written according to a prompt in a narrative style and any personal information is removed automatically. For the competition the texts are evaluated according to four different aspects: a) formal register which evaluates the grammatical aspects of the texts (da Silva Filho et al., 2023); b) thematic coherence which evaluates the if the written text follows the same theme as the motivation

Table 1: The table shows the distribution of grades according to each rubric.

	Formal Register	Rhetorical Structure	Cohesion	Thematic Coherence
1	27	20	26	204
2	111	13	109	175
3	475	123	484	317
4	116	437	108	39

prompt (da Silva Filho et al., 2023); c) rhetorical structure evaluate the uses of discourse marks by students along the essays (Lu et al., 2023) d) cohesion which evaluate the use of linkers and connective ideas along the essay (Oliveira et al., 2023). In each aspect, the text is evaluated in the range from 1 to 5 where 1 is the worst and 5 is the best possible grade. From all 1,235, essays 740 were used for training, 75 for validation and 370 for final test.

We extracted features using the BERTimbau large model similar to what was proposed by (Beiseiso and Alzahrani, 2020), but we didn't increment with any other feature. We classified the features using Logistic Regression, Gaussian Naïve Bayes and Random forest algorithms. Since most of the classes are imbalanced, we decided to use the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to mitigate the problem. Further, we used the model's implementation from sci-kit-learn and Huggingface (Wolf et al., 2020; Pedregosa et al., 2011).

The metrics used were the same used by the competition, which are the weighted average f1-score and Cohen kappa metric from scikit-learn library (Cohen, 1960; Pedregosa et al., 2011). Further, instead of considering only the means for all aspects of evaluation we also highlighted which model performance in each aspect.

3 Experiments

We performed preliminary experiments without considering any pre-processing of the original text. The table 2 shows the results for all rubrics evaluated. The results show the Logistic regression model outperforms the others when considering thematic coherence and rhetorical structure and has the best overall result. Meanwhile, random forest is the best model when considering cohesion and formal register.

In a further analysis of the public scores and private, the voting classifiers achieve better results than their counterparts. The results were 0.495 for the public score 0.486 for a private score for logistic regression, and 0.403 e 0.529 for Random

Forest for the same metrics. The voting classifiers achieved better results with 0.517 and 0.509 for private and public scores.

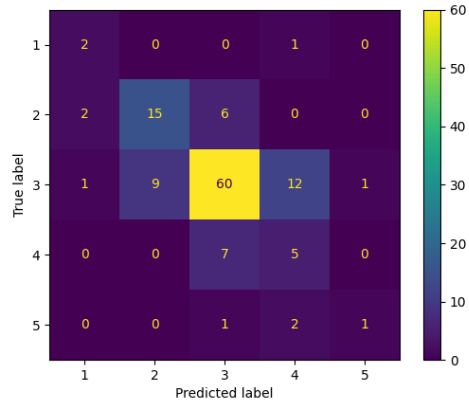


Figure 1: The confusion matrix shows the results of voting classifier for each grade. The expected grade is at the y axis and the predict grade in the x axis.

The confusion matrix showed by the figure 1 suggests that the classifiers performs better on the grades that have more non-synthetic examples.

4 Conclusion and Discussion

In a first analysis, in each rubric the algorithms are bias to produce a lower score as more synthetic data is added. It mostly happens in the rubric of rhetorical structure and cohesion where the imbalance dataset is more evident. Therefore, as more synthetic data is add more bias the models performs what reduces the precision of single model. Furthermore, since the voting class achieves better overall result in the final test set, it turns what that is less bias to the synthetic data of the training set corroborated by the figure 1. Furthermore, the results shows that one of the biggest challenges is to handle imbalance dataset for each rubric in automatic essay scoring task what might be mitigate by the use of Large Language models such as GPT-3 with few shot learning technique (Brown et al., 2020; Touvron et al., 2023).

Table 2: The table shows the result for each model according to each rubric and the result of the voting classifier.

	Formal Register	Rhetorical Structure	Cohesion	Thematic Coherence
Logistic Regression	0.550	0.425	0.480	0.518
Random Forest	0.610	0.368	0.482	0.468
Gaussian Naïve Bayes	0.486	0.285	0.384	0.414
Voting Classifier	0.560	0.315	0.472	0.474

References

- Felipe Akio Matsuoka. 2023. Automatic essay scoring in a brazilian scenario. *arXiv e-prints*, pages arXiv–2401.
- Majdi H. Beseiso and Saleh Alzahrani. 2020. [An empirical analysis of bert embedding for automated essay scoring](#). *International Journal of Advanced Computer Science and Applications*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Moésio Wenceslau da Silva Filho, André CA Nascimento, Pércles Miranda, Luiz Rodrigues, Thiago Cordeiro, Seiji Isotani, Ig Ibert Bittencourt, and Rafael Ferreira Mello. 2023. Automated formal register scoring of student narrative essays written in portuguese. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 1–11. SBC.
- Tiago Barbosa de Lima, Ingrid Luana Almeida da Silva, Elyda Laisa Soares Xavier Freitas, and Rafael Ferreira Mello. 2023. Avaliação automática de redação: Uma revisão sistemática. *Revista Brasileira de Informática na Educação*, 31:205–221.
- A Kumar, P Sharma, and R Singh. 2019a. Ensemble learning approach for predictive modeling using random forest. *Journal of Big Data Analytics in Healthcare*, 4(2):1–11.
- Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019b. Get it scored using autosas—an automated system for scoring short answers. In *Proceedings of the AAI conference on artificial intelligence*, volume 33, pages 9662–9669.
- Hayden Lu, Iel Lykha Dahunog, Jenny Rose Morales, and Norhynie Ranain. 2023. The writing makers of college students: A discourse analysis. *International Journal of Research*, 12(1):15–19.
- Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021. [Automated essay scoring using transformer models](#). *Psych*, 3(4):897–915.
- Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023. Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes’ theorem. *The Journal of Technology, Learning and Assessment*, 1(2).
- Shakshi Sharma and Anjali Goyal. 2020. Automated essay grading: An empirical analysis of ensemble learning techniques. In *Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 2*, pages 343–362. Springer.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

PiLN at PROPOR: A BERT-Based Strategy for Grading Narrative Essays

Rogério F. de Sousa

IFPI, Picos
rogerio.sousa@ifpi.edu.br

Jeziel C. Marinho

IFMA, Barra do Corda
jeziel.marinho@ifma.edu.br

Francisco A. R. Neto

IFPI, Teresina
farn@ifpi.edu.br

Rafael T. Anchiêta

IFPI, Picos
rta@ifpi.edu.br

Raimundo S. Moura

UFPI, Teresina
rsm@ufpi.edu.br

Abstract

This paper describes the participation of the PiLN team in the PROPOR’24 shared task on Automatic Essay Scoring of Portuguese Narrative Essays. The task aimed to develop methods for automatically evaluating essays to assist teachers in the classroom by enhancing formative feedback strategies, offering a more efficient and cost-effective alternative to human assessment. We approached this task by developing a strategy based on a BERT model; specifically, we fine-tuned a pre-trained BERT model of Portuguese - BERTimbau Large - to calculate scores for each assessed competency, incorporating both the prompt text and the essay text as input. Our simple approach achieved a reasonable result, reaching 4th place with an average score of 0.53985.

1 Introduction

Automated Essay Scoring (AES) is the computer technology that evaluates and scores the written prose (Shermis and Barrera, 2002). It aims to provide computational models for automatically grading essays or with minimal involvement of humans. This research area began with Page (Page, 1966) in 1966 with the Project Essay Grader system, which, according to Ke and Ng (Ke and Ng, 2019) remains since then.

AES is one of the most important educational applications of Natural Language Processing (NLP) (Ke and Ng, 2019; Beigman Klebanov et al., 2016). It encompasses some other fields, such as Cognitive Psychology, Education Measurement, Linguistics, and Written Research (Shermis and Burstein, 2013). Together, they aim to study methods to assist teachers in automatic assessments, providing a cheaper, faster, and more deterministic approach than humans when scoring an essay.

For Portuguese, this area has gained the attention of the community for grading ENEM essays due to publicly available corpora (Marinho et al.,

2021, 2022a). ENEM (High School National Exam - *Exame Nacional do Ensino Médio*) consists of an objective assessment and an essay test. The latter comprises a topic (prompt), usually a current problem in Brazilian society, and requires an intervention proposal from the participants. Besides, the text must be written in the argumentative type and not exceeding thirty lines. To grade an essay, ENEM adopts five specific traits that analyze different aspects of an essay¹.

Unlike the ENEM essays, this shared task adopted narrative-type essays and four traits (competencies): formal register, thematic coherence, narrative rhetorical structure, and cohesion. The objective was to develop a computational system capable of estimating a grade for an input essay for each specified trait of interest following the established grading rubric. The task used the average between the weighted F1 score and Cohen’s Kappa score, which are widely used in the literature for this task. To deal with this task, we developed a strategy based on BERT (Devlin et al., 2019); specifically, we fine-tuned the BERTimbau model (Souza et al., 2020) for predicting the four traits of narrative essays. With this strategy, we achieved 0.53985 on average and ranked 4th.

The rest of the paper is organized as follows. Section 2 describes the corpus of the shared task. In Section 3, we detail the developed approach and learned lessons. In Section 4, we present the achieved results. Finally, Section 5 outlines the limitations and future work.

2 Corpus

The dataset in this competition contains 1,235 essays written by students in Brazil’s 5th to 9th year of public schools. The students were instructed

¹<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>

to write a narrative essay based on a motivating text. All essays were manually digitized and anonymized. Afterward, the essays were analyzed by two human evaluators, who assessed different aspects of the essay based on a pre-defined correction rubric. This rubric provides instructive guidance for educators to consider four required competencies: Formal Register, Thematic Coherence, Narrative Rhetorical Structure, and Cohesion. Each dimension was assessed using integer levels ranging from 1 to 5, with higher levels indicating better text quality and language proficiency and lower levels demonstrating a lack of proficiency.

For that, this task made a training set and testing set available. The first one has 740 samples, while the second has 135 samples, where each row in the files contains the essay and a score for each competency. The final ranking was decided based on a blind testing set with 370 essays, according to Figure 1.

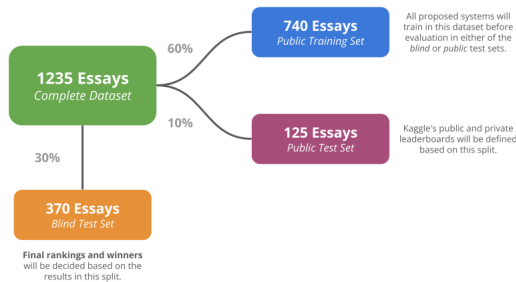


Figure 1: Corpus of the shared task.

In what follows, we detail our proposed strategy.

3 Proposed Method

The adopted method to address this task is based on the work of Matsuoka (2023), which uses a BERT model to evaluate ENEM essays, with some adaptations. The author developed a specialized model called *BERT_ENEM_Regression*, specifically designed for regression tasks, including evaluating essays according to the five competencies established by ENEM. Their study suggested employing the pre-trained Portuguese language model BERTimbau Base (Souza et al., 2020), and they enhanced Marinho et al. (2022a) findings by incorporating essay prompt text as input with the essay in the modeling process.

We employed the same strategy of incorporating the prompt text alongside the essay text as input to

the method. However, after testing several configurations, including adjusting the pre-trained model size (Base or Large) and varying the parameters for the fine-tuning process, the results improved when employing the pre-trained BERTimbau Large model. The optimal parameter set used for fine-tuning the model is presented in Table 1.

Parameter	Value
Dropout Layer	0.4
Linear Layer	(1024,4)
Epochs	6
Batch Size	8
Learning Rate	4×10^{-5}
Optimizer	AdamW
Loss	MSE

Table 1: Training parameters for the BERTimbau Large model.

The training corpus is divided into two parts to support the training process and, consequently, the discovery of the best parameters, with 90% for training and 10% for validation/development. The validation set selects the best model during the model fine-tuning process. We chose the model with the lowest loss.

It is worth noting that, as shown in Table 1, we employed a linear layer to make predictions. This layer takes as input, which aligns with BERT’s hidden size (1024 inputs) and has an output size of 4. This directly correlates to the four competencies in the essays, with each value corresponding to the score of the respective competency, enabling precise score predictions for each essay.

We also developed a hybrid model, using the features set of the (Marinho et al., 2022b). Although this model improved the results, we decided not to submit it to the shared task.

4 Results

We evaluate our method on the public test set of the shared task. The results for each competency are detailed in Table 2, where FR is Formal Register, TC is Thematic Coherence, NRS is Narrative Rhetorical Structure, and Co is Cohesion. As we can see, the best results are from the F-score metric; the best average was in the second competency, Thematic Coherence.

For the blind test set, we achieved 0.539 in the average between F-score and Kappa metrics. Ta-

Competency	F-score	Kappa	Avg
FR	0.68	0.45	0.56
TC	0.65	0.53	0.59
NRS	0.54	0.19	0.36
Co	0.66	0.34	0.50
Average	0.50		

Table 2: Results for each competency in the public test set.

ble 3 presents the final results available from the shared task organizers. One can see that our approach was ranked 4th.

Team	Score
INESC-ID	0.61
nlpr	0.55
Baseline - BERT Classifier	0.54
Ours (PiLN)	0.539
Baseline - BERT Embeddings+DT	0.532
Tiago de Lima	0.51
Ocean Team	0.46
Baseline - TFIDF	0.44

Table 3: Final result of the shared task.

To better understand the behavior of the model, we generated the confusion matrix for all four competencies, and upon observing it, we can highlight the following insights. Figure 2 presents all matrices. Concerning the Formal Register competency (Figure 2a), the model appears to encounter difficulties distinguishing intermediate scores, as a significant number of essays with intermediate scores mistakenly classified as score 1. Additionally, no essay received the maximum score, indicating that the model also faces challenges recognizing features corresponding to high-quality formal writing. For Thematic Coherence (Figure 2b), it is noticeable that there is a predominance of correct classifications for the lowest score, but on the other hand, almost no essays were classified with the highest score. This suggests that the model may recognize a lack of thematic coherence but has a limited ability to identify more sophisticated thematic coherence.

In Narrative Rhetorical Structure (Figure 2c), the confusion matrix reveals challenges distinguishing between scores 3 and 4, indicating difficulty recognizing excellent narrative structure. Furthermore,

there is a relatively high frequency of essays, with the minimum score being classified as higher scores (3) than they should deserve, indicating the need for improvements. Finally, regarding cohesion (Figure 2d), the model appears to have a good ability to distinguish between essays with intermediate levels of cohesion, but, on the other hand, it can be observed that there is a low quantity of essays classified correctly with the maximum score, indicating limitations in identifying advanced cohesion elements.

The source code is publicly available at: <https://github.com/lplnufpi/aes-propor>.

5 Limitations and Future Works

Although our approach reached good results, there is still room for improvement. For example, an error analysis would help to understand why the result in the Narrative Rhetorical Structure was not good. Moreover, a statistical analysis of the essay texts would show insights for incorporating other features into the BERT model.

For future work, we intend to use large language models (e.g., Albertina (Rodrigues et al., 2023)) as data augmentation to balance the corpus.

References

- Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. [Topicality-based indices for essay scoring](#). In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72, San Diego, CA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zixuan Ke and Vincent Ng. 2019. [Automated essay scoring: a survey of the state of the art](#). In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308, Macao, China. AAAI Press.
- Jeziel C. Marinho, Rafael T. Anchieta, and Raimundo S. Moura. 2021. [Essay-br: a brazilian corpus of essays](#). In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBDD 2021*, pages 53–64, Online. SBC.

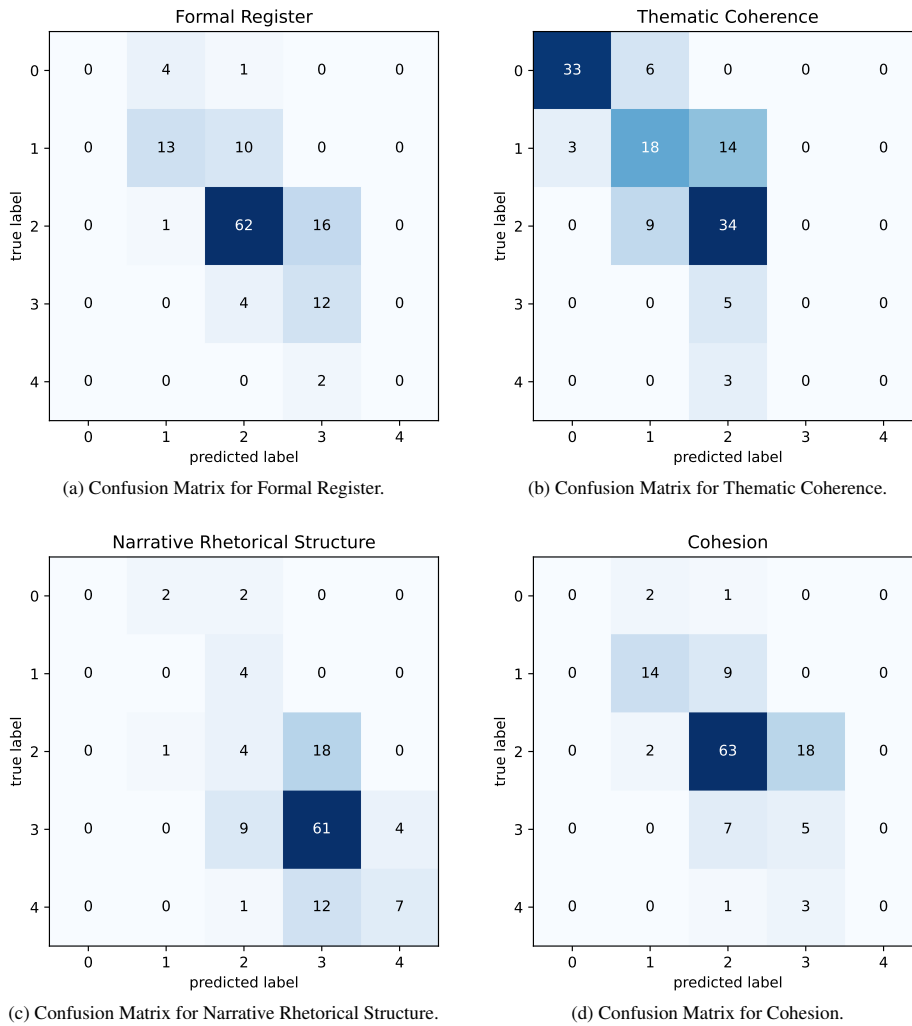


Figure 2: Confusion Matrix for Each Competency.

Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. 2022a. [Essay-br: a brazilian corpus to automatic essay scoring task](#). *Journal of Information and Data Management*, 13(1):65–76.

Jeziel C. Marinho, Fábio C., Rafael T. Anchiêta, and Raimundo S. Moura. 2022b. [Automated essay scoring: An approach based on enem competencies](#). In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60, Campinas, Brazil. SBC.

Felipe Akio Matsuoka. 2023. [Automatic essay scoring in a brazilian scenario](#).

Ellis B Page. 1966. [The imminence of... grading essays by computer](#). *The Phi Delta Kappan*, 47(5):238–243.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. [Advancing neural encoding of portuguese with transformer albertina pt-*](#).

Mark D Shermis and Felicia D Barrera. 2002. [Exit assessments: Evaluating writing ability through automated essay scoring](#). In *Annual Meeting of the*

American Educational Research Association, pages 1–30, New Orleans, LA. ERIC.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, USA.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. [Bertimbau: Pretrained bert models for brazilian portuguese](#). In *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Exploring the Automated Scoring of Narrative Essays in Brazilian Portuguese using Transformer Models

Eugénio Ribeiro¹ and Nuno Mamede^{1,2} and Jorge Baptista^{1,3}

¹ INESC-ID Lisboa, Portugal

² Instituto Superior Técnico, Universidade de Lisboa, Portugal

³ Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
{eugenio.ribeiro,nuno.mamede,jorge.baptista}@inesc-id.pt

Abstract

The automated scoring of narrative essays written by students according to different competences can assist teachers in their evaluation process and help them to focus on specific areas of writing that require improvement among their students. In this paper, we explore the fine-tuning of Portuguese foundation models to automatically score student essays according to four competences: formal register, thematic coherence, narrative rhetorical structure, and cohesion. The results of our experiments show that the agreement between these models and human graders varies between fair and substantial. Thus, although they can provide cues for essay scoring, significant research is still required towards their improvement, especially for the more complex competences.

1 Introduction

Automated Essay Scoring (AES) has garnered significant attention due to its potential to revolutionize the assessment of written language, particularly in educational settings. The PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays addresses this problem in the context of the Brazilian basic education system by focusing on scoring essays according to four competences: formal register, thematic coherence, narrative rhetorical structure, and cohesion.

In this paper, which describes our approach to the competition, we explore how Transformer-based foundation models for Portuguese perform when fine-tuned for scoring essays in terms of each of the target competences. With this study, we aim to assess whether these models are sufficiently robust for the task and can help teachers in their evaluation process or whether additional information is required to make them useful.

2 Related Work

AES has evolved considerably since its inception (Ifenthaler, 2022). The work by Haswell (2006) provides comprehensive insights into the development and history of AES. Recent reviews (e.g., Uto, 2021; Ramesh and Sanampudi, 2022; Vijaya Shetty et al., 2022), offer up-to-date perspectives on the state-of-the-art techniques and challenges in AES. Ethical considerations surrounding AES implementation, including economic pressures and validity concerns, have been extensively discussed (Jones, 2006; McAllister and White, 2006; Hannah et al., 2023). Furthermore, studies have explored the quality assessment of AES systems, aiming to maximize agreement between human and machine evaluations (Chen and He, 2013). Recent advancements in deep learning have propelled AES, with Transformer models and multimodal machine learning approaches gaining traction (Zhu and Sun, 2020; Kumar and Boulanger, 2021; Ludwig et al., 2021). Evaluation campaigns on AES (Mathias and Bhattacharyya, 2020) signal advancements in the area and can potentially enhance the efficiency and effectiveness of essay assessment processes.

In Brazilian Portuguese, research on AES has mainly focused on automatically grading the Exame Nacional do Ensino Médio (ENEM), which serves as an admission test for most universities in Brazil. Recent advances on this subject were mainly based on the development of the Essay-BR corpus (Marinho et al., 2022) and the fine-tuning of foundation models (Matsuoka, 2023). However, this problem had already been explored using both frequency-based (Bazelato and Amorim, 2013) and manually engineered features (Amorim and Veloso, 2017; Fonseca et al., 2018) paired with classical machine learning approaches. Additional studies focused on specific competences or aspects of the essays, such as thematic coherence (Passero et al., 2019; Pacheco et al., 2023), punctuation (de Lima

et al., 2023), formal register (Filho et al., 2023), and cohesion (Oliveira et al., 2023).

3 Experimental Setup

3.1 Dataset

The dataset used in the competition consists of 1,235 essays written by 5th to 9th-year students of public schools in Brazil. Each essay is based on a motivating text that accompanies it in the dataset. The essays were annotated by two human evaluators in terms of four competences: formal register, thematic coherence, narrative rhetorical structure, and cohesion. Each competence is scored in a scale of 1 to 5, with higher values indicating better text quality and language proficiency.

For the purpose of the competition, the dataset was split into a training set with 740 samples, a public test set with 125 samples, and a blind test set with 370 samples. The experiments in this study focus on the public test set. The distribution of scores is similar across the training and test sets. However, it is highly unbalanced and, with the exception of the thematic coherence competence, biased towards a single value: 3 for formal register and cohesion and 4 for narrative rhetorical structure.

3.2 Foundation Models

In this study, we explore the use of several foundation models for Portuguese. More specifically, we use the large version of BERTimbau (Souza et al., 2020), which is the most used of such models, and multiple versions of the Albertina PT-* model (Rodrigues et al., 2023), which achieved the state-of-the-art performance on multiple Natural Language Processing (NLP) tasks in Portuguese. We use the two large versions of the Albertina PT-BR model, one trained on brWaC (Wagner Filho et al., 2018) and the other on the OSCAR (Suárez et al., 2019) corpus. Additionally, we use the base version of the Albertina PT-BR model to assess the impact of using a smaller foundation model and the large version of the Albertina PT-PT model to assess the impact of using a foundation model dedicated to a different variety of the language.

3.3 Training & Evaluation

We address the scoring of the essays according to each competence independently as a 5-class classification problem. For each competence, each foundation model is fine-tuned on the training data for 20 epochs. The best epoch is then selected ac-

ording to the sum of the two evaluation metrics used in the context of the competition: weighted F₁ score and Cohen’s κ .

Considering the ordinal nature of the scores, the problem could also be approached as a regression task. However, preliminary experiments revealed a decrease in performance in comparison to the classification approach. Nonetheless, during the prediction phase, in addition to the traditional approach of selecting the class with highest probability, we also explore computing the weighted average of the class probability distribution. This approach, which we refer to as softmax regression, led to more robust predictions in a task of similar nature (Ribeiro et al., 2024).

To account for the non-deterministic aspects of neural approaches and enhance robustness, we performed six independent experimental runs. The evaluation metrics are reported as the average across these runs. All non-error metrics are reported in percentage form.

4 Results

Table 1 shows the average results of our experiments. Comparing the results for the different competences, we can see that the scoring performance is significantly worse for the narrative rhetorical structure than the remaining competences. This was expected, as it presents a more complex problem. Furthermore, the fact that the best results for this competence were achieved using the smaller foundation models suggests that the larger models are overfitting and additional training data is required to capture that complexity.

Looking into the results for the other competences, we can see that using the large Albertina PT-BR model trained on brWaC consistently led to better performance than both BERTimbau, which was trained on the same corpus, and the version of the Albertina PT-BR model that was trained on the OSCAR corpus. Furthermore, we observed significant drops in performance when using the base version of the Albertina PT-BR model, which has one ninth of the parameters of the large versions. For thematic coherence and cohesion, we also observed a drop in performance when using the Albertina PT-PT model. However, it outperformed all the other models for scoring in terms of formal register, in spite of being dedicated to a different Portuguese variety. This is probably due to the fact that it was trained on a large amount

Foundation Model		Formal Register		Thematic Coherence		Rhetorical Structure		Cohesion	
		F ₁	κ	F ₁	κ	F ₁	κ	F ₁	κ
BERTimbau Large	CL	70.32	.4434	69.70	.5886	56.83	.2587	68.69	.3909
	SR	69.83	.4375	69.39	.5842	56.22	.2442	68.69	.3909
Albertina PT-BR	CL	69.88	.4508	68.66	.5834	53.53	.1777	68.72	.4080
	SR	69.53	.4475	69.70	.5982	54.37	.1920	68.80	.4098
Albertina PT-BR brWaC	CL	72.39	.5115	69.78	.5956	55.37	.2328	69.88	.4306
	SR	72.24	.5075	70.29	.6079	55.30	.2265	68.97	.4096
Albertina PT-BR Base	CL	67.79	.4210	66.39	.5464	56.86	.2283	67.96	.3814
	SR	65.85	.3971	66.89	.5534	56.93	.2361	67.69	.3776
Albertina PT-PT	CL	73.64	.5222	68.19	.5763	56.20	.2339	67.66	.3738
	SR	74.07	.5308	67.67	.5720	56.37	.2353	68.15	.3857

Table 1: Average results across the multiple runs. CL stands for classification and SR for softmax regression.

of parliament data, which is typically more formal and better written than generic web-crawled data.

Regarding the prediction approach, the results reveal no clear advantage in using softmax regression, as its impact varies across models. Still, it led to the highest average performance in terms of F₁ for narrative rhetorical structure and both metrics for formal register and thematic coherence.

Finally, it is important to refer that we relied on the models with best performance across all runs to enter the competition. In comparison to the average performance, these represent an improvement between 1 and 4 percentage points in terms of F₁ and between .03 and .09 in terms of agreement.

5 Conclusion

Overall, the results of our experiments show that the agreement between the fine-tuned models for AES and human graders varies between fair and substantial. Thus, although these models can provide cues for essay scoring, significant research is still required towards their improvement, especially for the more complex competences. In this context, as future work, we intend to explore the use of hybrid models that combine the strengths of foundation models with those of manually engineered features specific to each of the competences.

Acknowledgments

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-

CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

References

- Evelin Amorim and Adriano Veloso. 2017. *A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese*. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL): Student Research Workshop*, pages 94–102.
- Bruno Smarsaro Bazelato and ECF Amorim. 2013. *A Bayesian Classifier to Automatic Correction of Portuguese Essays*. In *Conferência Internacional sobre Informática na Educação (TISE)*, pages 779–782.
- Hongbo Chen and Ben He. 2013. *Automated Essay Scoring by Maximizing Human-Machine Agreement*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1741–1752.
- Tiago de Lima, Luiz Rodrigues, Valmir Macario, Elyda Freitas, and Rafael Mello. 2023. *Automatic Punctuation Verification of School Students’ Essay in Portuguese*. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 58–70.
- Moésio Silva Filho, André Nascimento, Péricles Miranda, Luiz Rodrigues, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Rafael Mello. 2023. *Automated Formal Register Scoring of Student Narrative Essays Written in Portuguese*. In *Anais do Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (WAPLA)*, pages 1–11.
- Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. *Automatically Grading Brazilian Student Essays*. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 170–179.

- Liam Hannah, Eunice Eunhee Jang, Maitree Shah, and Vinayak Gupta. 2023. [Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits](#). *Language Assessment Quarterly*, 20(4–5):399–420.
- Richard H. Haswell. 2006. [A Bibliography of Machine Scoring of Student Writing, 1962–2005](#). In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences*, chapter 17, pages 234–243. Utah State University Press.
- Dirk Ifenthaler. 2022. [Automated Essay Scoring Systems](#). In Olaf Zawacki-Richter and Insung Jung, editors, *Handbook of Open, Distance and Digital Education*, pages 1–15. Springer Nature Singapore.
- Edmund Jones. 2006. [ACCUPLACER's Essay-Scoring Technology: When Reliability Does Not Equal Validity](#). In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences*, chapter 6, pages 93–113. Utah State University Press.
- Vivekanandan S. Kumar and David Boulanger. 2021. [Automated Essay Scoring and the Deep Learning Black Box: How are Rubric Scores Determined?](#) *International Journal of Artificial Intelligence in Education*, 31:538–584.
- Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021. [Automated Essay Scoring using Transformer Models](#). *Psych*, 3(4):897–915.
- Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. 2022. [Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task](#). *Journal of Information and Data Management*, 13(1).
- Sandeep Mathias and Pushpak Bhattacharyya. 2020. [Can Neural Networks Automatically Score Essay Traits?](#) In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 85–91.
- Felipe Akio Matsuoka. 2023. [Automatic Essay Scoring in a Brazilian Scenario](#). *Computing Research Repository*, arXiv:2401.00095.
- Ken S. McAllister and Edward M. White. 2006. [Interested Complicities: The Dialectic of Computer-Assisted Writing Assessment](#). In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences*, chapter 1, pages 8–27. Utah State University Press.
- Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023. [Towards Explainable Prediction of Essay Cohesion in Portuguese and English](#). In *Proceedings of the International Learning Analytics and Knowledge Conference (LAK)*, page 509–519.
- Rafael Pacheco, Luiz Rodrigues, Lucas Lins, Péricles Miranda, Valmir Macário, Seiji Isotani, Thiago Cordeiro, Ig Bittencourt, Diego Dermeval, Dragan Gašević, and Rafael Mello. 2023. [Automated Thematic Coherence Scoring of Student Essays Written in Portuguese](#). In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1086–1097.
- Guilherme Passero, Rafael Ferreira, and Rudimar Luís Scaranto Dazzi. 2019. [Off-Topic Essay Detection: A Comparative Study on the Portuguese Language](#). *Revista Brasileira de Informática na Educação*, 27(3):177–190.
- Dadi Ramesh and Suresh Kumar Sanampudi. 2022. [An Automated Essay Scoring Systems: a Systematic Literature Review](#). *Artificial Intelligence Review*, 55(3):2495–2527.
- Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. [Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches](#). In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*.
- João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. [Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*](#). *Computing Research Repository*, arXiv:2305.06721.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. [BERTimbau: Pretrained BERT Models for Brazilian Portuguese](#). In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures](#). In *Workshop on the Challenges in the Management of Large Corpora (CMLC)*, pages 9–16.
- Masaki Uto. 2021. [A Review of Deep-Neural automated Essay Scoring Models](#). *Behaviormetrika*, 48(2):459–484.
- S. Vijaya Shetty, K. R. Guruvyas, Pranav P. Patil, and Jeevan J. Acharya. 2022. [Essay Scoring Systems using AI and Feature Extraction: A Review](#). In *Proceedings of the International Conference on Communication, Computing and Electronics Systems (ICC-CES)*, pages 45–57.
- Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. [The brWaC Corpus: a New Open Resource for Brazilian Portuguese](#). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4339–4344.
- Wilson Zhu and Yu Sun. 2020. [Automated Essay Scoring System using Multi-Model Machine Learning](#). In *Proceedings of the International Conference on Machine Learning Techniques and NLP (MLNLP)*, pages 109–117.

**First Workshop on NLP for Indigenous
Languages of Lusophone Countries
(ILLC-NLP)**

Building a Language-Learning Game for Brazilian Indigenous Languages: A Case of Study

Gustavo Polleti

Universidade de São Paulo
gustavo.polleti@usp.br

Abstract

In this paper we discuss a first attempt to build a language learning game for Brazilian indigenous languages and the challenges around it. We present a design for the tool with gamification aspects. Then we describe a process to automatically generate language exercises and questions from a dependency treebank and a lexical database for Tupian languages. We discuss the limitations of our prototype highlighting ethical and practical implementation concerns. Finally, we conclude that new data gathering processes should be established in partnership with indigenous communities and oriented for educational purposes.

1 Introduction

Language learning games are key tools to vitalize endangered languages (Thomason, 2015; Xu et al., 2022; Neubig et al., 2020). LARA (Akhlaghi et al., 2019), a multi language learning assistant, is an example that has been key to support actions related to endangered languages protection (Rayner and Wilmoth, 2023; Bédi et al., 2022; Zuckermann et al., 2021). Despite the necessity of language learning tools to vitalize endangered languages, they are typically restricted to high-resource languages, such as English, and require significant effort to be extended to languages with few spoken and written resources. For example, “7000 languages”¹, a non-profit organization dedicated to build online courses for endangered languages, require 1 to 2 years to build a language course. As a result, despite the immediate need, language learning tools are expensive and, as they are developed today, are hard to scale to cover all the 2,680 languages in risk of being extinct by the end of this century (Wurm, 2001). In particular, Brazil hosts approximately 270 indigenous languages, all of which are endangered. Brazilian indigenous languages, collectively known as Brazilian Indigenous

Languages (BILs) henceforth, are spoken by at most 30 thousand people, few of which are young children and teenagers. Brazilian indigenous communities require language learning tools that can teach their native language to Portuguese speakers. Since Portuguese is a low-resource language, when compared to English for example, and due to the lack of data resources on BIL, Brazilian indigenous communities are underserved by current learning tools.

In this work, we describe the process of building a language learning tool for BIL, which we will refer as “BILingo”. BILingo is a language learning game app, heavily inspired by industry leaders on language learning apps (Duolingo² and Busuu³). We discuss in detail the challenges of building a language learning tool for BIL, such as the lack of written and phonetical resources, ethical concerns on available treebanks and databases used for exercise generation, and provide some suggestions on steps forward. We managed to build a minimal proof of concept course for Guajajara language divided in two sections. We employed dependency treebanks and a lexical database on BIL as source for exercise generation. The main contribution of this work is to present a case of study on building a language learning tool for BIL and, we hope, it will serve as a starting point for the development of an actual fully fledged language learning app that can be used to strengthen the culture of indigenous communities in Brazil.

The paper is organized as follows. Section 2 presents BILingo’s design and describe its development process, including their data sources and exercise’s format. In Section 3 we discuss the challenges and limitations of our prototype, we analyse our processes and resources from both a practical implementation and ethical perspective. Finally, in

¹<https://www.7000.org/>

²<https://www.duolingo.com/>

³<https://www.busuu.com/>

Section 4 we offer concluding remarks.

2 Method and Results

BILingo’s design follows the gamified language learning structure found in apps available in the industry (e.g. Duolingo) and in the literacy (Lightbown, 2021; von Ahn, 2006; Katinskaia et al., 2017). It has an initial course page as depicted in Figure 1. The student will progress linearly in the course by completing lessons. To complete a lesson the student needs to pass a series of exercises. Every time they make a mistake, they lose a “red gem”, if the student is out of red gems, they have to wait 5 minutes before trying a lesson again. Once the student completes a lesson, they can advance to the next one. These are gamification aspects typically found in language learning apps in the industry.

When the student engages with a lesson, they can find three different types of language exercises: (1) “translate sentence” TS1, (2) “translate sentence in the target language” TS2 or (2) “concept match” CM, see Figure 2. In the TS1 exercise, depicted in Figure 2a, the student is presented with a sentence in portuguese and asked to select the tokens from the Guajajara language in the correct order to form the translated sentence; TS2 is the same but the initial sentence is presented in Guajajara and the student is asked to translate it to portuguese. CM exercises present a word in Guajajara and images of possible concepts that are represented by that word, then, the student is asked to select which one of the images correspond to the given word, see Figure 2b. Popular language learning assistants, such as LARA, employ phonetical exercises that are absent in our prototype. At this point, we focused on written exercises only, with the purpose of simplify the setup and because we couldn’t find any phonetical databases readily available.

Now that we presented our tool, we describe the details of its implementation. In order to build BILingo’s question database for BIL, we used a simple exercise generation method based on available treebanks and lexical databases. In this work we used TuLeD (Gerardi et al., 2022b) and TuDeT (Gerardi et al., 2022a), which are respectively a lexical database and a dependency treebank for several Tupian languages, including Guajajara. These databases compile several resources from the literature on Tupian languages and structure them into a single format suitable for analytical pur-

poses. TuDeT treebank offers several dependency trees in the original indigenous language with some correspondent sentences in other mainstream languages, such as english, spanish, portuguese and french. For example, the token list (“oho”, “kara”, “ipiaromo”) corresponds to the sentence “Ela foi buscar inhamé” in portuguese. On the other hand, TuLeD offers a lexical database with the ontological concept associated to each term. So, in our example, you will find that the word “kara” means “yam”. In order to link both databases, we first conducted a search of all the concepts available in TuLeD on TuDeT sentences, so that we could tell which concepts were present in each dependency tree. Here, we considered a hit if the exact same form that appeared in each dependency tree was present in the lexical database. For example, we could tell that the sentence “oho kara ipiaromo” refers to the concept “yam” on the word “kara”.

As we presented, BILingo consists in 3 concepts: “course section”, “lesson” and “exercise”. First, we had to determine with the available resources at hand, our unified TuLeD and TuDeT database, which topics or subjects we could cover in our prototype. To select the topics, we grouped the available dependency trees for each language by concept, and then manually inspected which concepts were suitable for building course sections. For example, concepts like “yam”, “pineapple” and “pepper” appear in a significant number of sentences, e.g. more than 10, we can tell that “food” can be a good candidate as course subject. Once we selected the subjects for a given course section, we can filter only the sentences related to the listed concepts. Now we can generate CM exercises by sampling from the listed concepts within the same section. Furthermore, we can also use the dependency trees, with their correspondent translation in portuguese, to build class TS1 and TS2 exercises. To generate tokens other than the correct ones for the TS1 and TS2 exercises, we can simply select a sample set of dependency trees from the same language, shuffle their tokens and apply a random sampling. The lessons were generated by randomly sampling a predefined number of exercises, e.g. 4, from our set, always ensuring to have two of each kind. In this work, we were able to generate exercises for two course sections, each one comprising 4 lessons. Table 1 presents an example list of exercises generated through our process.

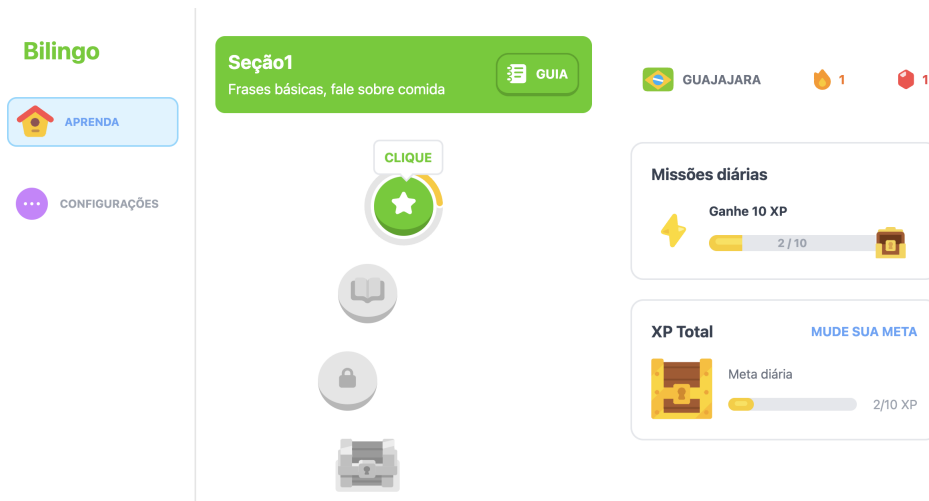
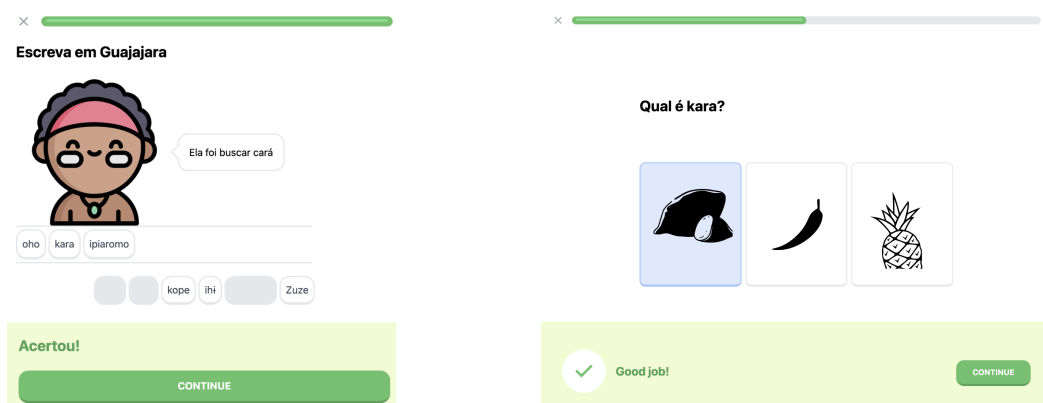


Figure 1: Landing page of BILingo. It displays a linear progress trajectory of lessons and incorporates gamification aspects such as daily quests and daily streak.



(a) Translate sentence exercise type example. The student was asked to form the sentence “She went to look for yam” using the Guajajara word set.

(b) Concept match exercise type example. The student was asked to tell which of the images corresponded to the Guajajara word “kara” that means yam.

Figure 2: Exercise types in our learning language tool. Depicts scenarios where the student has answered correctly.

Table 1: BILingo example exercises.

Section	Question (translate sentence)	Answer
food	ela foi buscar cará	oho kara ipiaromo
food	a mãe de josé foi a roça para buscar carã	oho zuze ihi kope kara ipiaromo
food	tem abacaxi na roça de josé	heta nana zuze kope
food	ele colhe cacau	opo?o aka?u a?e
animal	a mulher envolveu o peixe	owan kuza pira a?e
animal	foi o queixado	tazahu ru?u
animal	o que foi que o queixado comeu na roça	ma?e tazahu u?u kope ra?e
animal	o homem alimentou o peixe	opoz awa pira a?e

Table 2: Statistics on sources for exercise generation. For each language, we have the count of sentences that had at least a single concept associated (has_concept) and their respective translations to portuguese (pt) and english (en).

language	has_concept	pt	en
Tupinamba	True	0	140
Tupinamba	False	0	409
Teko	True	0	19
Teko	False	0	95
Munduruku	True	22	22
Munduruku	False	155	156
Makurap	True	0	15
Makurap	False	0	37
Karo	True	0	260
Karo	False	0	664
Kaapor	True	0	58
Kaapor	False	0	83
Guajajara	True	719	487
Guajajara	False	1172	806
Akuntsu	True	25	186
Akuntsu	False	36	325

3 Challenges and Limitations

Our prototype falls short in several aspects, from the difficulties of working with limited sources of data for exercise generation to ethical concerns, now we examine all the learnings and challenges to actually build a working system for BIL. As we discussed, our work relies on TuLeD and TuDeT as source for exercise generation. Both databases were developed by compiling several sources from the literature, so there was no structured data gathering process and, thus, the data may be seem questionable in many senses. First, we could observe that it is severely incomplete, notably when we consider coverage of dependency trees with translation to portuguese. Since many works that were used as source for TuDeT were carried out by foreigner research groups, most of the translations are in english, see Table 2. The lack of portuguese translations hinder their application for language learning purposes targeting brazilian people. In fact, we only had portuguese translations for “Guajajara”, “Munduruku” and “Akuntsu” out of the 8 languages available. Additionally, we applied a data cleaning process to fix spelling and remove citations in the sentences themselves. Often the translated sentence would include a citation to its original work. For example, the sentence “opo?o

aka?u a?e” was escorted by its portuguese translation “ele colhe cacau (harrison, 2013:12)”, where there is a citation to the Portuguese-Guajajara dictionary (Harrison and Harrison, 2013). The link between the treebank and TuLeD also face issues related to coverage, see Table 2. Finally, it is worth to note that the material available in the treebank was not necessarily designed for educational purposes and, thus, require moderation if ever properly applied in practice.

Besides the practical limitations of our data sources, it is worth to comment on some ethical concerns. BILingo prototype and its underlying data sources were designed without substantial indigenous community involvement (Pinhanez et al., 2023). First, since TuLeD and TuDeT compile plenty sources from the literacy, it is hard to ensure that their data gathering procedures were compliant with ethical guidelines (Lewis et al., 2020), for example Los Pinos Declaration⁴, or even if all their translations are validated by actual indigenous speakers. Here we should note that any language learning tool on BIL should rely on data sources that were carefully designed in partnership with indigenous communities.

4 Conclusion

In this case of study, we explored the development of a language learning tool for BIL. We described how such a tool could work by detailing the student progression through course sections, lessons and exercises. We managed to use an existing dependency treebank (TuDeT) and a lexical database (TuLeD) to generate exercises. We were able to produce a working prototype and validate the potential of using dependency trees associated with lexical database to automatically generate exercises. Finally, we discussed the challenges and limitations of such system from practical and ethical perspectives.

Future work should develop data gathering protocols for creating treebanks and lexical databases with indigenous communities, and oriented for educational purposes so that we can have sufficient and reliable data sources to build effective learning tools in practice. Additionally, once it is possible to release such a system, research should be conducted to evaluate the engagement on indigenous communities and optimize the learning system so that students stay engaged.

⁴<https://unesdoc.unesco.org/ark:/48223/pf0000374030>

References

- Elham Akhlaghi, Branislav Bédi, Matt Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. 2019. [Overview of LARA: A Learning and Reading Assistant](#). In *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 99–103.
- Branislav Bédi, Hakeem Beedar, Belinda Chiera, Nedelina Ivanova, Christele Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, John Sloan, and Ghil'ad Zuckermann. 2022. [Using lara to create image-based and phonetically annotated multimodal texts for endangered languages](#). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, COMPUTEL 2022 - 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Proceedings of the Workshop, pages 68–77. Association for Computational Linguistics (ACL). Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages : Language Diversity: from Low-Resource to Endangered Languages, COMPUTEL-5 2022 ; Conference date: 26-05-2022 Through 27-05-2022.
- Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2022a. [TuDeT: Tupían Dependency Treebank](#). Zenodo.
- Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Tim Wientzek, Johann-Mattis List, and Robert Forkel. 2022b. [TuLeD. Tupían Lexical Database](#). Zenodo.
- Carl Harrison and Carole Harrison. 2013. *Dicionário Guajajara-Português*. SIL.
- Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. [Revita: a system for language learning and supporting endangered languages](#). In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 27–35, Gothenburg, Sweden. LiU Electronic Press.
- Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. [Indigenous protocol and artificial intelligence position paper](#). Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI. Edited by Jason Edward Lewis. English Language Version of "Ka?ina Hana ?Ōiwi a me ka Waihona ?Ike Hakuha Pepa Kūlana" available at: <https://spectrum.library.concordia.ca/id/eprint/990094/>.
- Spada N. Lightbown, P. M. 2021. *How Languages Are Learned (5 ed.)*. Oxford University Press.
- Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan S Sharma, and Patrick Littell. 2020. [A summary of the first workshop on language technology for language documentation and revitalization](#). In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.
- Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. [Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Manny Rayner and Sasha Wilmoth. 2023. [Using LARA to rescue a legacy Pitjantjatjara course](#). In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 13–18, Remote. Association for Computational Linguistics.
- S. G. Thomason. 2015. *Endangered Languages: An Introduction*. Cambridge: Cambridge University Press.
- L. von Ahn. 2006. [Games with a purpose](#). *Computer*, 39(6):92–94.
- S.A. Wurm. 2001. *Atlas of the world's languages in danger of disappearing*. Unesco Pub.
- Liang Xu, Elaine Uí Dhonnchadha, and Monica Ward. 2022. [Faoi gheasa an adaptive game for Irish language learning](#). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 133–138, Dublin, Ireland. Association for Computational Linguistics.
- Ghil'Ad Zuckermann, Sigurður Vigfússon, Manny Rayner, Neasa Ní Chiaráin, Nedelina Ivanova, Hanieh Habibi, and Branislav Bédi. 2021. [LARA in the service of revivalistics and documentary linguistics: Community engagement and endangered languages](#). In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 13–23, Online. Association for Computational Linguistics.

Computational Model for *Yorùbá Àrokò* Communication System

Adéwuyi Adétáyò ADÉGBÌTÉ

Department of Computer Science Department of Computer Science & Engineering

Adekunle Ajasin University

Akungba Akoko, Nigeria.

adewuyi.adegbite@aaua.edu.ng

Ọdétúnjí Àjàdí ỌDÉJOBÍ

Obafemi Awolowo University

Ile-Ife, Nigeria.

oodejobi@oauife.edu.ng

Abstract

This research interrogates the *Yorùbá Àrokò* System (YAS) from a computing perspective. This is with the view to determining the extent to which the communication elements and structure of *Àrokò* can be precisely formulated as computing artefacts. *Àrokò* is a message encoding system used to serve a variety of communication functions in *Yorùbá* societies. The operational terminologies for the concept of *Àrokò* were gathered from experts using a combination of observation, consultation, interview and documented materials and the structure formalised. The elicited information was modelled using formal language and automata theory based techniques. The model and the system designed was validated by demonstrating its use with selected *Àrokò* communication. The grammar of *Àrokò* Language was formulated. The study provides a computational perspective and explanation to the aspects of the process underlining the *Àrokò* system, which in return established a feasible and digital resource for information communication. The study developed a computational model to formally and explicitly represent the concepts embedded in *Yorùbá Àrokò* communication system.

Keywords: Computational model, *Yorùbá*, Communication System, *Àrokò*, Symbol-object

1 Introduction

Àrokò is a collection of objects which are usually packaged and parcelled together. It is a symbol-object that is sometimes sent by the means of a messenger to another person with a purpose of proper message decoding at the receiving end and conveyance of message from the source (Ogundeji, 1997). It is the indigenous system (Mundy and Compton, 1991; Mundy and Lloyd-Laney, 1992) of shared meaning for communication between acquaintances and adversaries in the *Yorùbá*

tradition (Bascom, 1973; Elúyemí, 1987). This communication method employs a set of symbolic objects and signs with mutually understood reasons for communication (Dima et al., 2014). Message representation in the *Yorùbá* cultural heritage can take various forms including visually in artworks as seen in various designs on ancient decorated doors, orally in folklores as seen in different stories relating to tortoise and elephants (Elegbe and Nwachukwu, 2017; Uzochukwu and Ekwugha, 2015), and implicitly in belief and value systems as seen in the use of *é* for elderly people or more than one person and *ó* for someone younger (Ayan-tayo, 2010; Akpabio, 2003). The various heritage in *Yorùbá* land is not fully represented computationally, but various efforts are recently been put in place in recent times to affect this (Folorunso et al., 2010). *Àrokò* is a message exchange instrument that engages the sender, the carrier, and the receiver, which sometimes is in the context of the interaction that has been taking place between the two parties. *Àrokò* can be a continuation of a discussion that has started beforehand, so it can be based and interpreted in the context of an ongoing discourse, though it could also kickstart a discussion between two or more individuals. *Àrokò*, which is a non-verbal means of communication (Metaxas and Zhang, 2013) in *Yorùbá*, is vast and heterogeneous. This makes it an interesting area for research in message representation with an excellent underlying computational possibility in particular (Alimi, 2013; Aziken and Emeni, 2010).

There is a need for a more extensive understanding of the phenomenon being investigated (that is, *Àrokò*) with an elaborate and extensive explanation that presents a model which captures the *Yorùbá Àrokò* communication system. Some of the ancient means of communication are rich in terms of the structure and composition that make up their methods and means of passing across a message. This work is to study the various concepts in *Àrokò* and

see how it is related and connected to the standard concept of communication. There is a need to study *Àrokò* from a computational concept, the articulate computational model for *Àrokò* by studying the various concepts, and seeing how the concept is related to other communication theories. This research focused on the communication principle and structure of selected *Yorùbá Àrokò* communication system with a focus on the following versions: (i) *Ààlè* (ii) *Ètúfú* and (iii) *Agà*. The work does not include ambiguously interpreted objects while information for objects with ambiguity will assume the most widely accepted and known meaning.

2 The *Àrokò* System

Àrokò has been used to communicate among the indigenous people of *Yorùbá* by conveying messages from one person to another or from one village/community to another. Objects are been packed together and passed from the sender to the receiver. Sometimes, secretive messages are sent (Mehrabian, 1981) using *Àrokò* using animals, messenger, or a friend but mostly sent by hand through a courier or messenger (*ikò* or *iránsé*) (Ogundeji, 1997). The *Àrokò* must be well understood by the receiver. Stone, chewing stick, flywhisk, fruit, parrot, cowrie shell, blood (Nabofa, 1994), fabric, a stick of broom, broom, calabash, kola nut, leaves, and other common things are utilized. *Àrokò*, like many other *Yorùbá* names, appears to be a derivative noun, according to (Ogundeji, 1997). If the term *Àrokò* is seen as the word-formation of two combined verbs *rò* (to think) and *kò* (to agree) preceded by a which is changed to a noun, the result is *Àròkò* rather than *Àrokò*. This usage of *Àrokò* is comparable to how codes and symbols have lately been used. It helps to lessen the usage of spoken words by allowing objects to be utilized to convey information. Some of the objects sent to the receiver are kept for reference purposes (Olómọ̀là, 1979).

In those days, *Yorùbá* used symbols to send warnings, warn a loved one of an imminent threat, alert a partner of a breakup or quarrel (Olatunji, 2013), and inform family members or close relatives of someone's death. *Àrokò* was also utilized in decision-making in *Yorùbá* society if a new king or chief was to be selected and the kingmakers were not in favour of a certain candidate, this was expressed by shaking hands with left hands. Similarly, the ladies of the town wore the wrong sides

of their garments to demonstrate their opposition to the nomination of a new chief or king (Olómọ̀là, 2003). When a couple is looking for a child or children, they employ a variety of methods. When the Ifa priest is consulted and it is discovered that the couple will not have children, eggshells wrapped in cotton wool are packaged and delivered to the parents. The information being passed along implies that the couple will not have children. *Àrokò* is believed to be an idea or thought upon which we have agreed to (Osisanwo, 2009). *Àrokò*, that is, *Àrokò: Ohun tí àjọ̀rò wa kò lé lóri ni à ñ pè ní Àrokò* (Odéjọ̀bí, 2019).

3 Model Design

At the senders' point, which serves as the source as seen in Figure 1, there is the creation of the message to be sent and the packaging of the objects. The objects to be used to get across the exact information intended are carefully chosen. The various objects are also ordered and arranged in the right way to be able to give the right information to the receiver. The objects are packed and sometimes wrapped together with an object which serves as the encoding of the *Àrokò* message system. The materials are released and transmitted with the aim of the receiver getting the information sent. The sender also chooses the person who will send the objects. The interpretation of some messages is influenced by the identity of the courier. At the receiving end, the objects are received and then rightly interpreted. After it is been opened, it is then processed so as to get the true meaning of what the objects sent is as seen in Figure 1, but Figure 2 has creation, packaging, ordering, messenger's choice (this is used when giving a response to the received objects) as parts of its receiving end because of its ability to respond to give a reply back to the sender.

In *Àrokò*, the right interpretation has constraints in the interpretation of its meaning depending on the senders' and the receivers' profession, and existing conversation between the sender and the receiver, and this serves as the caveat. Table 2 has a list of a few of the gathered materials used in the *Àrokò* communication system.

3.1 Àrokò Communication System Model

Definition 1 The Yorùbá Àrokò System(YAS) is formulated by a five-tuple (5-tuple) defined as:
 $YAS ::= \langle P, M, L, C, Ct \rangle$

where:

P is the Package: This is conceptualised as the envelop inside which the message is enclosed. It is used to wrap the actual message being sent.

M represents Media: It is the material that formed the medium through which the message is transmitted.

L represents Language: This is the shared system of meaning for encoding the message. It is a symbolic formulation of the human language, in this case, Standard Yorùbá language.

C represents Content: This is the encoded message. It is the information encoded in the message that is directed and sent to the receiver or audience of the Àrokò. The content in this Àrokò materials is a union of the individual element of materials (E) used and the relations (R), between the individual elements. It is represented by, $C = \cup\{E, R\}$. The relations R, is the union of many non-empty sets R_0, R_1, \dots, R_n .

Ct represents Caveat: This is the restriction or constraint on the interpretation of the content of the message sent. It serves as specific conditions for a particular interpretation of the context of meaning to be derived for the information encoded in the Àrokò object that is sent.

Starting with the basic conception of Àrokò as defined in Section 3.3, an expressively rich computational framework, capable of treating its ontological contents as theoretical objects whose properties and logical components can be clearly defined is developed as seen in Table 3.

The purpose of language is to give material expression to mental objects by rendering them into sensually accessible forms and shapes. From Figure 3, the terminologies used are stated as follows:

$$\begin{aligned}
 M_S &\implies MessageSent \\
 M_E &\implies MessageEncodedInformation \\
 M_T &\implies MessageTransmitted \\
 M_D &\implies MessageDecoded \\
 M_R &\implies MessageReceived \\
 L_1 &\implies LanguageforMessageSent \\
 L_2 &\implies LanguageforMessageEncoded \\
 &\quad Information \\
 L_3 &\implies LanguageforTransmission \\
 L'_2 &\implies LanguageforMessageDecoded \\
 L'_1 &\implies LanguageforMessageReceived \\
 L_0 &\implies MentalLanguage, nativetotheseif \\
 &\quad (formlessandshapeless) \\
 L_1 &\implies Nativelanguageofthecommunityof \\
 &\quad human(HumanLanguage)inirregularforms, \\
 &\quad shapes, norms, format, andstructure. \\
 &\quad (1)
 \end{aligned}$$

A mental language is a linguistic rendering of a mental concept. The rendering gives irregular forms and shapes to the mental concept.

$L_0 = L_1$ - Material

If $M_S = M_R$

Then (1.) There is no contradiction

(2.) There is no ambiguity

If $H_S L_0 = H_R L_0$

Then (1.) This expression is completely interpreted

$L_3 \rightarrow$ Efficiency (Precise)

Precise (i.e. got meanings of materials)

$L_2 \rightarrow$ Effective (Informative and correct)

Objective

$L_1 \rightarrow$ Meaningful

Subjective

From the expression here which relates to how the communication system works. The first line means there is no information in the material and also that symbols are in material objects. The next one states that there is no information in symbols. Symbols can be used to encode sensually accessible information. Information is a mental object. There is no sign in symbols and there are no symbols in signs. A sign is a sensually accessible mental object. Speech is a language encoded sound. The sound is the medium (Material). The language

Table 1: Materials Used in Àrokò

S/N	Material	S/N	Material	S/N	Material	S/N	Material	S/N	Material
1.	<i>Ewé Odán</i>	10.	<i>Èsun Iṣu</i>	19.	<i>Ata</i>	28.	<i>Ọfà</i>	37.	<i>Ètù</i>
2.	<i>Eésan</i>	11.	<i>Oṣé Ṣàngó</i>	20.	<i>Ajè Ìbọn</i>	29.	<i>Ọrun</i>	38.	<i>Àkò</i>
3.	<i>Iyò</i>	12.	<i>Aṣọ Obinrin</i>	21.	<i>Awọ Eran</i>	30.	<i>Ṣìgìdì</i>	39.	<i>Èye</i> <i>Ayékòótó</i>
4.	<i>Koríko</i>	13.	<i>Ọta</i>	22.	<i>Pàkúté</i>	31.	<i>Opa Osugbo</i>	40.	<i>Ewé</i>
5.	<i>Àidan</i>	14.	<i>Ìgbálẹ̀</i>	23.	<i>Ìgò</i>	32.	<i>Omi</i>	41.	<i>Apópó</i> <i>Obì</i>
6.	<i>Apurù Ọdẹ</i>	15.	<i>Ehoro</i>	24.	<i>Àhàyá</i>	33.	<i>Iṣu</i>	42.	<i>Ení</i>
7.	<i>Ṣéééré</i> <i>kekere</i>	16.	<i>Ère Ṣàngó</i>	25.	<i>Efun</i>	34.	<i>Òkúta</i>		
8.	<i>Òwù</i>	17.	<i>Ọmọrí Igbá</i>	26.	<i>Ọmo Ayò</i>	35.	<i>Orí Eye</i>		
9.	<i>Kuùku</i> Àg- <i>bàdo</i>	18.	<i>Aṣọ funfun</i>	27.	<i>Tábà</i>	35.	<i>Obì</i>		

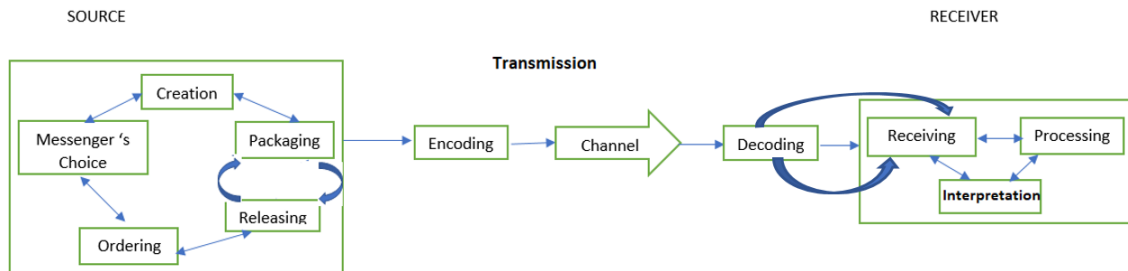


Figure 1: Communication Model of Yorùbá Àrokò Communication System

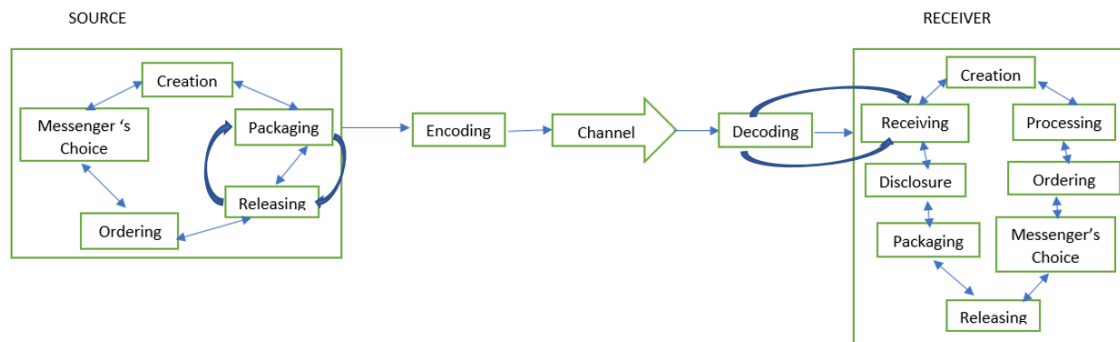
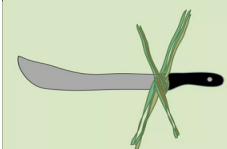
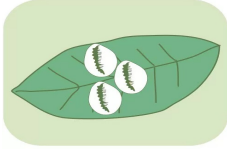

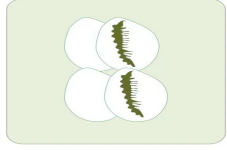


Figure 2: Communication Model of Yorùbá Àrokò Communication System with Feedback Parameter

Table 2: Sample of Àrokò Objects Collected

SN	Àrokò Objects	Description
1.		A weapon wrapped with either palm fronds or grass is used to send a warning of an impending war. The weapon used could either be a cutlass, arrow, or gun.
2.		In Yorùbá numeration and counting, three which stands to <i>ééta</i> . It means something miserable, evil, or bad to befall someone. When three cowrie shells are wrapped in a leaf, it is symbols send to someone owning the sender, that the recipient or receiver of the objects should pay up the debt or face the repercussion.
3.		<i>Òdòyà</i> is the tool used by women in parting the hair, most of the time when they want to plait. It is used to part the hair into the subsection they desire it to be. So, when it is used, it means it is an end to a relationship and that particular relationship cannot be reconciled.
4.		In Àrokò communication, cowries is one of the major materials used in communication. When the cowrie shells have been strung together. It denotes the end of a relationship, particularly between lovers.

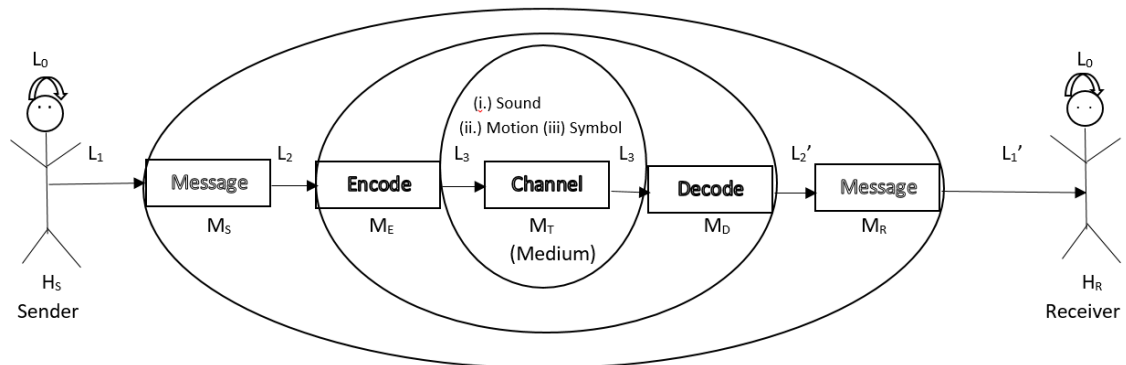
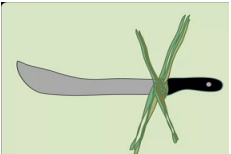
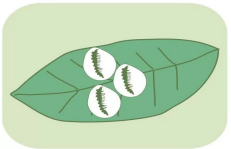
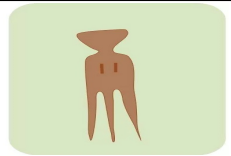


Figure 3: Yorùbá Àrokò Communication System Model

Table 3: Description of *Yorùbá Àrokò* System(YAS) Model

SN	Àrokò Objects	Definition
1.		<p>Cutlass(Àdà)</p> <ol style="list-style-type: none"> 1. Package: Palm fronds or Grass. 2. Media: A courier, servant, or messenger. 3. Language: Standard <i>Yorùbá</i>. 4. Content: Cutlass, arrow, or gun. 5. Caveat: If it is sent from a community or friend already in war with another community. It is a call to help and assist in the war.
2.		<p>Cowries (Owó eyọ)</p> <ol style="list-style-type: none"> 1. Package: Leaf 2. Media: It could be through the sender itself or an individual the receiver can closely associate with the sender. 3. Language: Standard <i>Yorùbá</i>. 4. Content: Three Cowries. 5. Caveat: The number of cowries can bring a change in meaning and interpretation of this particular <i>Àrokò</i>
3.		<p>Comb (òyà)</p> <ol style="list-style-type: none"> 1. Package: Does not necessarily need any package. 2. Media: It can be sent through a courier or the sender deliver the material. 3. Language: Standard <i>Yorùbá</i>. 4. Content: Hairdresser comb (Òyà) 5. Caveat: When a hairdresser comb is sent between friends that have a disagreement, that put an end to their relationship. If the same material is sent between two hairdressers, it means a call to help in plaiting the hair of the sender.

is used to encode or infuse sensually accessible information into the sound.

Speech = Sound + Language

Noise = Speech – Language

Silence = Speech – Sound

Mental = Language - Material

$L_0 = L_1 - \text{Material}$

When language is removed from speech, what remains is noise.

4 System Implementation

In computing, grammar can be defined as the mechanism for formal specification of the elements and structure of a language. Formal means there is a standard and generally accepted standard that everyone adheres to. In modelling the *Àrokò* system of communication, one of the major components involved in the language of the material being sent. To do proper modelling of the language, the grammar of the language is developed. for *Àrokò* system of communication can be derived and there are four major areas on which the production rules will focus on. The first is the one that derives its meaning from the verb of the material being used. The material used is a noun, why the action word, that is the verb gives the meaning of the *Àrokò*. For example, *Òd̀yà* is a noun derived from the verb *yà*, which means to separate. Another example is in the use of *Ab̀èb̀è*, which is also a noun but derives from the verb *b̀èb̀è*, that is to plead. So, when *Ab̀èb̀è* is sent it is used to plead to the receiver of the material. Therefore, when this material is used, the verb of the material is important.

Definition 2 Grammar is four-tuple defined as Grammar $G ::= \langle V_T, V_N, P, S \rangle$

where: V_T is the non-empty finite set of terminal symbols which are also called the alphabet of the language;

V_N is the non-empty set of non-terminal symbols that form the vocabulary of the language. Each string in V_N is composed of V_T ;

P is the finite set of non-empty rules by which each non-terminal is replaced with one or more strings of terminals and non-terminals. They are also called re-written rules; and

S is the start non-terminal symbol. S is an element of V_N . It is a unique member of the non-terminal symbol to indicate the beginning of an expression.

Àrokò communication system starts from the sender as seen in Figure 4, which is the initial state

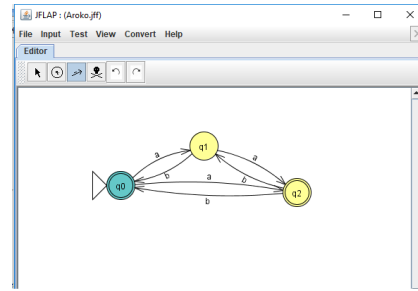


Figure 4: JFLAP Implementation of *Àrokò* Grammar

LHS	RHS
S	→ aA
S	→ aB
S	→ λ
B	→ λ
A	→ bS
B	→ bS
B	→ bA
A	→ aB

Figure 5: Production rules of *Àrokò* Grammar

and can pass through a messenger or directly to the receiver, and the receiver can be the final recipient of the materials, or there could be a response back through the messenger or directly to the sender. The production rules of the grammar of the language are shown in Figure 5.

5 Summary

The study for this work stemmed from a desire to learn more about communication systems before limiting it down to non-verbal communication in humans. The distinctiveness of the non-verbal communication system in the *Yorùbá* community was investigated, and it was discovered to be rich in words and symbol interpretation. This brought about the study centering on *Yorùbá Àrokò* communication system which has existed for years among the *Yorùbá* community. The message system of this particular communication system was looked into and a corresponding model for the *Yorùbá Àrokò* was developed.

The study was able to present the formal representation of the concept and means of communication of the *Yorùbá Àrokò* communication system. Artefacts that can be implemented on a computational instrument were developed and this can be re-used for related tasks.

6 Conclusion

From the various related work examined so far, it is discovered most works of literature talk about the importance and relevance of *Àrokò* in communication (Sidikat, 2015), and few literatures were able to focus on an aspect or two of the areas in which *Àrokò* can be used, the only computational aspect as known in literatures is the cryptography in terms of the secretive part of the *Àrokò* means of communicating. So, it could be seen from existing pieces of literature that none has been able to give a formal and computational model of the communicative context of *Yorùbá Àrokò* System.

When in need of the proper interpretation and understanding of the meaning of *Àrokò*, the study of signs can not be unconnected from the culture of consideration and the background of existing discussions between the sender and the receiver. It is then the meaning of *Àrokò* can be formulated and explained. This work has presented the study and communication concept analysis of *Yorùbá Àrokò*. The work has been able to present a communication model that can serve as a basis for developing communication systems.

References

- Akpabio, E. (2003). *African Communication Systems: An Introductory Text*. B Print Publications, Lagos. ISBN: 978-060-076-0. Retrieved from <http://www.unam.edu.na/staff/eno-akpabio>. Accessed on August 30, 2018.
- Alimi, S. A. (2013). Indigenous communication systems as determinants of cultural resurgence in *Yorùbá* societies of oyo and osun states, nigeria. *Ph. D. Thesis, University of Ibadan.*, pages 1–181.
- Ayantayo, J. K. (2010). The ethical dimension of african indigenous communication systems: An analysis. *LUMINA*, 21(1):1–6.
- Aziken, L. C. and Emeni, F. C. A. (2010). Traditional systems of communication in nigeria: A review for improvement. *Knowledge Review*, 21(4):23–29.
- Bascom, W. (1973). *African Art in Cultural Perspective*. W. Horton, London. Retrieved from <https://www.questia.com/read/102076255/african-art-in-cultural-perspective-an-introduction>. Accessed on 30 August, 2018.
- Dima, I. C., Teodorescu, M., and Gifu, D. (2014). New communication approaches vs. traditional communication. *International Letters of Social and Humanistic Sciences*, 20:46–55.
- Elegbe, O. and Nwachukwu, I. (2017). A cross-cultural analysis of communication patterns between two cultures in southwest nigeria. *Journal of Humanities and Social Sciences*, 2017(9):52–65.
- Elúyemí, O. (1987). African systems of contact and communication. *Nigeria Magazine*, 55(7):1–12.
- Folorunso, O., Akinwale, A. T., Vincent, R. O., and Olabenjo, B. (2010). A mobile-based knowledge management system for ifa: An african traditional oracle. *African Journal of Mathematics and Computer Science Research*, 3(7):114–131.
- Mehrabian, A. (1981). *Silent Messages*. New York: Wadsworth, California. Retrieved from https://www.academia.edu/23744443/Albert_MehrabianSilentMessages on August 30, 2018.
- Metaxas, D. and Zhang, S. (2013). A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31:421–433.
- Mundy, P. and Compton, L. (1991). Indigenous communication and indigenous knowledge. *Development Communication Report*, 4(3):1–3.
- Mundy, P. and Lloyd-Laney, M. (1992). Indigenous communication: appropriate technology. *IT Publications Ltd.*, 19(2):1–10.
- Nabofa, M. Y. (1994). Blood symbolism in african religion. *Religious Studies*, 21(3):389–405. Cambridge University Press. Doi: 10.1017/S0034412500017479.
- Odéjóbí, O. A. (2019). Lecture notes in csc 614. *Department of Computer Science and Engineering, Obáfémí Awólówò University, Ilé-Ife, Nigeria.*, June, 2019.
- Ogundeji, P. A. (1997). The communicative and semi-otic contexts of *àrokò* among the *Yorùbá* symbol-communication systems. *African Languages and Cultures*, 10(2):145–156.
- Olatunji, R. W. (2013). Uses of semiotics in periods of hostilities, armed conflicts and peace building among the *Yorùbá*, south-west nigeria. *International Journal of Arts and Humanities*, 2(4):1.
- Olómọlà, I. (1979). *àrokò*: An indigenous *Yorùbá* semi-otic device. *ODU, New Series*, 19:1–24.
- Olómọlà, I. (2003). *Prominent Traditional Rulers of Yorùbá land*. Ibadan Press Limited.
- Osisanwo, W. (2009). A pragma-semiotic. analysis of *àrokò*: The *Yorùbá* means of symbolic communication. *Contemporary Humanities*, 3:1–16.
- Sidikat, A. A. (2015). The relevance of *àrokò* as a communication device among the *Yorùbá* native speakers of nigeria. *DEGEL: The Journal of the Faculty of Arts and Islamic Studies*, 10:135–146.

Uzochukwu, C. E. and Ekwugha, U. P. (2015). The relevance of the folk media as channels in promoting african cultural values. *Communication Panorama African and Global Perspective*, 1(1):1-11.

Human Evaluation of the Usefulness of Fine-Tuned English Translators for the Guarani Mbya and Nheengatu Indigenous Languages

Claudio Pinhanez, Paulo Cavalin, Julio Nogima

IBM Research, Brazil

{csantosp, pcavalin, jnogima}@br.ibm.com

Abstract

We investigate how useful are machine translators based on the fine-tuning of LLMs with very small amounts of training data, typical of extremely low-resource languages such as Indigenous languages. We started by developing translators for the Guarani Mbya and Nheengatu languages by fine-tuning a WMT-19 German-English translator. We then performed a human evaluation of the usefulness of the results of test sets and compared them to their SacreBLEU scores. We had a level of alignment around 60-70%, although there were about 40% of very wrong translations. The results suggest the need of a filter for bad translations as a way to make the translators useful, possibly only in scenarios of human-AI collaboration such as writing-support assistants.

1 Introduction

In this paper we present a human evaluation of the usefulness of machine translation (MT) models which we trained to translate sentences from two Brazilian Indigenous Languages (BILs), i.e. Guarani Mbya and Nheengatu, to English. The main goal was to evaluate the end-user usefulness of the MT models based on fine-tuning a pre-trained Transformer-based language model, aka Large Language Model (LLM) (Devlin et al., 2019; Raffel et al., 2020), in the case of extremely low-resource languages.

Our method consisted of fine-tuning the WMT19 model (Ng et al., 2019), trained to translate German sentences to English, with both parallel corpora and language resources, to each of the BILs. Since for both Guarani Mbya and Nheengatu data is quite scarce, we relied on resources such as dictionaries (or lexicons) and educational documents to extract as much parallel data as possible in the training set and to compile a set of parallel sentences for testing. We then measured the performance of the models

with SacreBLEU, which is the implementation of the BLEU score (Post, 2018).

It is very difficult to draw conclusions about human usefulness of a translator based only on values from automatic metrics such as SacreBLEU, since they are based on easy-to-perform computations such as word comparison, ignoring often semantic issues. Therefore, to determine the usefulness of the translators, we conducted a human evaluation on the texts generated from the test set inputs. Our analysis consisted of labeling each of the generated outputs in a seven-point scale, ranging to near-perfect quality to very wrong translations.

Results showed that translations were good for only 18% of the Guarani Mbya outputs and 32% for the Nheengatu outputs. Considering content which could be utilized by an user, the results were 35% and 42%, respectively. However, 40% and 42% of the translations were considered very wrong. These results suggest that such translators are more likely to be useful in scenarios of direct human-machine collaboration, such as writing assistants, than of standalone automatic translation. We then compared the human-based usefulness results with the SacreBLEU scores, finding alignments of about 60%.

We believe this work contributes to the understanding of how traditional translation metrics relate to actual end-user usefulness. It also highlights the need of care to use such metrics as system evaluation tools.

2 Datasets

We created two datasets, one for each BIL.

2.1 Guarani Mbya dataset

Sentences from three different sources were used in the construction of the *Guarani Mbya* dataset. The first source was a set of Guarani Mbya short stories with 1,022 sentences, available in both Portuguese

and English (Dooley, 1988a,b). The second comprises 245 texts extracted from PDF files with a pedagogical character (Dooley, 1985). The third source was Robert A. Dooley’s *Lexical Guarani Mbya Dictionary* (Dooley, 2016), a reference work for the language, from which we extracted 2,230 sentence pairs. The last two sources contained sentence pairs in Guarani Mbya and Portuguese only. We converted them to English using a Portuguese-to-English commercial translation service. We have permission from the author to use this data.

After concatenating the data from the three sources, we cleaned it, removing some non-alphanumeric characters (e.g. *, >>, •) and normalizing Unicode values. Then, the dataset was split into training and test sets and finalized by removing repeated sentences and cross-contamination between sets, totaling 3,155 and 300 sentence pairs, respectively.

2.2 Nheengatu dataset

The *Nheengatu* dataset used five different sources containing Nheengatu sentences with Portuguese translations. As with the Guarani Mbya dataset, we converted the Portuguese sentences to English using a Portuguese-to-English commercial translation service¹.

The first source is the *Nheengatu lexicon* (Ávila, 2021) with 6846 sentences extracted from the lexicon examples. For that, we processed the original file made available by the author. The second one is *Corpus Lições* (Ávila, 2021), containing 1,665 samples already available in a spreadsheet format. The other sources, which were directly extracted from PDFs, were: *Texto Anônimo* (Navarro, 2011), with 427 samples; *Brilhos na Floresta* (Ishikawa, 2019), with 590 samples; and *Curso LGA* (Navarro, 2016), with a partial extract of 590 samples.

The Nheengatu dataset contains 7,281 samples, with a random split of 241 samples (10% of the data from all sources except Nheengatu lexicon) for testing and 6,804 samples for training.

3 Machine Translation Models

We trained two models by fine-tuning a pre-trained Transformer-based Language Model to translate from Guarani Mbya and Nheengatu to English. That was done by fine-tuning the parameters of the WMT19 model (Ng et al., 2019), a 315M-parameter German-to-English machine translator pre-trained

¹IBM Watson Language Translation v9.0.0

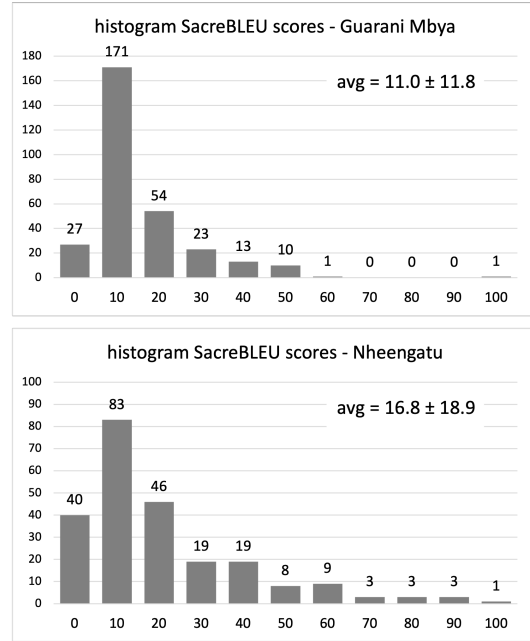


Figure 1: Histograms of the SacreBLEU scores of the Guarani Mbya and Nheengatu translators.

with about 28M pairs of translated sentences and more than 500M back-translated sentences. Both models were fine-tuned for 10 epochs, considering a batch size of 32, learning rate of 2.10^{-5} decaying to 2.10^{-6} according to a cosine function.

4 SacreBLEU Evaluation

To evaluate the results we relied on the the *SacreBLEU* metric which is the BLEU score computed with the SacreBLEU Python package (Post, 2018). We computed sentence-level scores and considered the average of those scores for system-level evaluation.

We observe slightly higher scores from the Nheengatu translator, with a SacreBLEU score of 16.8 against 11.0 from Guarani Mbya. We see, however, that the Nheengatu model resulted in higher standard deviation, with 18.9 against 11.8 of the Guarani Mbya model, which made us question the distribution of the data and compute the histograms of the scores for each test set which are shown in Figure 1. What we observe is a skinnier distribution for Guarani Mbya which may explain the higher standard deviation of Nheengatu.

5 Human Evaluation of Usefulness

In this section we present the results of a human-based evaluation of all the test set outputs of the translators which we conducted to understand the

usefulness of the sentences generated by them. We had two goals in doing a comprehensive manual evaluation of the translations: first, to help us to determine how far the translators are from an actual deployment; and second, to understand how much standard ML metrics can be relied on as a predictor of success in actual human tasks.

The evaluation was performed by one of the authors of the paper by comparing, for all sentences of the test sets, the translation to English from the test set to the generated output. The evaluator did not know both languages but had access to the original text in the Indigenous languages for inspection purposes. Through a process similar to what is used for *thematic networks* (Attride-Stirling, 2001), the categories and their meanings were developed by an iterative process of evaluating sentences, refining the categories, and re-evaluating the sentences until saturation was reached. From that point on, all entries were then evaluated. This process led to the following categories and labels of the usefulness of the translations:

very wrong: the output was completely unrelated to the expected translation or had gross mistakes such as repetitions, words from the source language, or it was empty;

incorrect: no blatant mistakes but there was no relation with the expected text;

mostly incorrect: one or two correct words but mostly of the rest was useless;

usable: the output could be used as a starting point for a translation because it had two or three correct words or it resembled the structure of the expected sentence;

mostly correct: at least two thirds of the generated text were correct but it still had mistakes which needed human correction;

correct: the generated text was an acceptable translation of the original sentence although it could fail to capture completely the meaning of the expected text;

near perfect: the output was almost a literal repetition of the expected text.

The rightmost columns of Table 1 depict the number of sentences evaluated into those different categories for the 300 outputs of the Guarani Mbya test set and the 233 of the Nheengatu test set. For the Guarani Mbya translator, we see about 40% of all outputs are in the *very wrong* category and 26% in the *incorrect* and *mostly incorrect* categories. Of

Guarani Mbya usefulness	SacreBLEU score range					TOTAL		
	0 to 5	5 to 10	10 to 20	20 to 50	50 to 100	#	%	
very wrong	59	46	12	2	0	119	40%	
incorrect	18	21	14	2	0	55	18%	
mostly incorrect	3	8	8	4	0	23	8%	
usable	16	15	8	11	0	50	17%	
mostly correct	2	8	8	16	0	34	11%	
correct	0	1	4	9	0	14	5%	
near perfect	0	1	0	2	2	5	2%	
TOTAL	#	98	100	54	46	2	300	100%
	%	33%	33%	18%	15%	1%	100%	

Nheengatu usefulness	SacreBLEU score range					TOTAL		
	0 to 5	5 to 10	10 to 20	20 to 50	50 to 100	#	%	
very wrong	50	33	13	1	0	97	42%	
incorrect	7	10	8	2	1	28	12%	
mostly incorrect	1	6	0	3	0	10	4%	
usable	2	2	9	10	1	24	10%	
mostly correct	3	3	8	15	4	33	14%	
correct	1	1	3	7	2	14	6%	
near perfect	2	1	5	8	11	27	12%	
TOTAL	#	66	56	46	46	19	233	100%
	%	28%	24%	20%	20%	8%	100%	

Table 1: Results of the human evaluation of usefulness the Guarani Mbya and Nheengatu translators and their relationship with SacreBLEU scores (alignment regions marked with a grey background).

the remaining 34%, about 28% are sentences which need some level of human intervention to be used (categories *usable* and *mostly correct*) and only 7% would be suitable in an automatic translation scenario. The numbers of the Nheengatu translator are better, with 42% in the *very wrong* category but only 16% in the *incorrect* and *mostly incorrect* categories. Of the remaining 42%, 24% would need human correction to be usable and 18% would be suitable for an automatic translation scenario.

Next, we examined how the human evaluation of usefulness of the generated translations related to the SacreBLEU scores. Table 1 also depicts the number of sentences of each category in relation to 5 ranges of SacreBLEU scores, which follow a log-like distribution. If the two methods of evaluation were aligned, we would expect the majority of the sentences to be along the main diagonal of the tables. However, there is a good amount of spread and to quantify it we divided the table in two areas: the cells close to the main diagonal (depicted with a grey background on Table 1) and those in the left-bottom and right-top triangles.

In the results of the Guarani Mbya translator, the main diagonal contains 186 (62%) of all outputs while the non-aligned areas comprise 114 (38%). In the Nheengatu translator, there are 125 (71%) outputs on the main diagonal and 51 (29%) on the non-aligned areas. In general, in about one third of the cases, for both translators, the SacreBLEU score does not seem to be not a good predictor of



Figure 2: Distribution of SacreBLEU scores to each of the qualitative evaluation categories for the Guarani Mbya and Nheengatu translators.

the usefulness of a translation.

Figure 2 provides a visual rendition of the data on Table 1 which shows more clearly that the Nheengatu SacreBLEU scores seem to be better correlated with the usefulness evaluation than the scores of the Guarani Mbya translator.

6 Final Discussion

In this paper we explored two forms of evaluation of two translators from Guarani Mbya and Nheengatu languages to English. The first evaluation method, totally automatic, used the traditional SacreBLEU metric for translators, resulting in sentence averages of 11.0 ± 11.8 and 16.8 ± 18.9 respectively. The second form of evaluation was based on a human-created scale of usefulness established through an iterative process based on thematic networks. Results indicated that, for the Guarani Mbya translator, 40% of all generated sentences were totally useless, 26% had too many mistakes to be usable, about 28% could sometimes help knowledgeable end-users, and 7% were ready to be used. The Nheengatu translator had a better performance with 42% useless, 16% almost useless, 24% usable, and 18% with no errors. The two metrics had about 62% and 71% of alignment,

respectively. These results seem to indicate that the translators, at this stage, can be only used in demos and initial prototypes.

It is very rare to find any kind of human-based usefulness testing of ML language systems, and even more in ML translator systems. We believe this kind of evaluation is particularly important in contexts of extremely low-resource languages such as the ones studied in this paper, since small amounts of data may impact the quality of traditional human-free ML metrics. Moreover, in this work we developed an evaluation which was specific to the task and based on the characteristics of the actual data, making it more ecologically valid. Unlike other works, we did not use human beings to validate a ML metric but instead we developed a more comprehensive metric which is directly related to the intended use. We see this as a major contribution of this work.

The work has important limitations which should be highlighted. First and foremost, only one human evaluator was used, a non-speaker of both languages. We plan to do, in future works, studies with multiple and language-knowledgeable evaluators to further validate our results. Another issue is that the Nheengatu translator was built with more than twice the number of training samples of the Guarani Mbya, what may explain its superior results.

Beyond those issues, the results of the human evaluation suggest more focused ways to improve the end-user performance which go beyond the traditional focus on simply increasing overall accuracy. In particular, around 40% of the outputs were very wrong, in a way that possibly they can be filtered out by a simple ML detector built directly with the data. Notice that, as shown in Table 1, only half of those outputs are easily detectable by the SacreBLEU score (0 to 5), and therefore a simplistic focus on improving the scores may not be enough to fix the problem.

Finally, we want to underscore the importance of ML developers to explore and have direct contact with the output data. During the evaluation process we could notice some other issues and errors which suggested readily available opportunities for improvement. This is an important benefit of human-based evaluations, which are often shun by developers as wasteful and time-consuming. Manually exploring and evaluating the output should be, in our opinion, a fundamental process in the construction of machine translation systems.

References

- Jennifer Attride-Stirling. 2001. Thematic networks: an analytic tool for qualitative research. *Qualitative research*, 1(3):385–405.
- J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proc. of ACL'19*.
- Robert Dooley. 1985. Nhanhembo'e aguã nhandeayvupy [1-5].
- Robert A. Dooley. 1988a. Arquivo de textos indígenas – guaraní (dialeto mbyá) [1].
- Robert A. Dooley. 1988b. Arquivo de textos indígenas – guaraní (dialeto mbyá) [2].
- Robert A. Dooley. 2016. [Léxico guarani, dialeto mbyá: Guarani-português](#).
- Noemia Kazue Ishikawa. 2019. *Brilhos na Floresta*. Editora Valer; Editora Inpa, Manaus.
- Eduardo de Almeida Navarro. 2011. [Um texto anônimo, em língua geral amazônica, do século xviii](#). *Revista USP*, (90):181–192.
- Eduardo de Almeida Navarro. 2016. *Curso de língua geral (nheengatu ou tupi moderno): a língua das origens da civilização amazônica*, 2 edition. Edição do Autor, São Paulo.
- Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. [Facebook FAIR's WMT19 news translation task submission](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- M. T. Ávila. 2021. [Proposta de dicionário nheengatu-português](#). Tese de doutorado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. Recuperado em 2023-12-27, de www.teses.usp.br.

A Universal Dependencies Treebank for Nheengatu

Leonel Figueiredo de Alencar

Universidade Federal do Ceará

Av. da Universidade, 2683 - 60020-180 Fortaleza, Brazil

leonel.de.alencar@ufc.br

Abstract


We present UD_Nheengatu-CompLin, the inaugural treebank for Nheengatu, an endangered Indigenous language of Brazil with limited digital resources. This treebank stands as the largest among Indigenous American languages in version 2.13 of the Universal Dependencies collection. The developmental version comprises 1,336 trees, encompassing 13,246 tokens and 13,374 words. In a 10-fold cross-validation experiment using UDPipe 1.2, parsing with gold tokenization and gold tags achieved a labeled attachment score (LAS) of 81.17 ± 1.02 , outperforming Yauti, the rule-based analyzer employed for sentence annotation.

1 Introduction

Universal Dependencies (henceforth UD) provides a framework for consistent morphosyntactic annotation across languages of different families, aiming at both linguistic typology and natural language processing (Nivre et al., 2016; de Marneffe et al., 2021). The UD collection has grown from 10 treebanks of 10 European languages in version 1.0 of January 2015 to 259 treebanks of 148 languages from all continents in version 2.13 of November 2023 (Zeman et al., 2023). However, the enormous diversity of Indigenous languages in the Americas is still underrepresented despite the efforts in the last five years.¹

This paper introduces UD_Nheengatu-CompLin, the first UD treebank for Nheengatu (ISO 639-3: yr1), an endangered Indigenous language of Brazil, also known as Modern Tupi and *Língua Geral Amazônica* (hereafter LGA). Although it made its debut in UD v2.11 on November 15, 2022, with only 196 trees, it has since expanded significantly. With 1,239 trees totaling 12,621 tokens, it stands

¹Following the recommendations in *The Chicago Manual of Style Online* (2024) and elsewhere, we capitalize *Indigenous* in the sense used in this paper.

as the largest treebank for an Indigenous American language in UD v2.13. To our knowledge, no analogous resource for Nheengatu exists. It is made available under a  license.

2 Related work

Wagner et al. (2016) adapted the UD annotation guidelines to Arapaho, an Algonquian language spoken in Wyoming, USA. Shipibo-Konibo, however, seems to have been the first Indigenous American language with a treebank under the UD framework (Vasquez et al., 2018).² There followed Mbya Guaraní (Thomas, 2019), Yupik (Park et al., 2021), K'iche' (Tyers and Henderson, 2021), Apurinã (Rueter et al., 2021), Nahuatl (Pugh et al., 2022), Tupinamba, and ten other languages, mostly Tupian of Brazil (Martín Rodríguez et al., 2022; Santos et al., 2024). As expected of treebanks for low-resource languages, they are “opportunistic corpora” (McEnery and Hardie, 2012, p. 11) with no reported inter-annotator agreement.

Parsing experiments with these treebanks showed that performance is heavily dependent on factors like gold part-of-speech (POS) tags and training data size. For instance, parsing Shipibo-Konibo with gold POS tags yielded a labeled attachment score (LAS) of 81.25 ± 3.45 , while parsing raw text resulted in a score of 30.39 ± 1.34 , indicating a significant drop in performance (Vasquez et al., 2018). This is not surprising given the small size of the treebank with only 407 trees and 2,706 tokens. Similarly, for the Nahuatl treebank, which had a larger size of 10,356 tokens and 939 trees, UDPipe 1 (Straka et al., 2016; Straka and Straková, 2017) was used to obtain a LAS score of 68.1 ± 2.0 with normalized text (Pugh et al., 2022).

Nheengatu, with a Digital Language Support Level of only 0.07 (Simons et al., 2022; Eberhard

²This treebank has never been part of any release of the UD collection. Instead, the UD homepage lists it among the “possible future extensions”.

et al., 2023), is among many minority languages impacted by the digital divide, despite recent initiatives. For example, da Rocha D'Angelis et al. (2021) discusses the localization of a smartphone operating system for Nheengatu. However, this system does not provide any text input enhancement technologies, e.g., word completion, spelling correction, etc. After summarizing previous directly related work, de Alencar (2023) proposes a tool called Yauti for the UD annotation of Nheengatu. Cavalin et al. (2023) included Nheengatu in a study of language identification.

3 Nheengatu, the “good language”

Nheengatu originated in the 17th century in Maranhão from Tupinamba, one of the many varieties of Tupi, which was dominant along the Brazilian coast in the 16th century (Edelweiss, 1969; Borges, 1996; Rodrigues, 1996; Freire, 2011; Rodrigues and Cabral, 2011; Navarro, 2012; Finbow, 2023). The Portuguese colonizers adopted Tupi as *língua geral*, i.e. lingua franca, of which other varieties besides the LGA developed (de Lurdes Zanolli, 2022; Leite, 2013). Description and teaching of Tupi, e.g., Anchieta (1595); Figueira (1621), were incumbent on Jesuits (Edelweiss, 1969; de Almeida Navarro, 2009; Altman, 2022). Not Portuguese, but Tupi was Brazil's de facto first national language (Drumond, 1964). It was widespread among black Africans as well as Europeans and their descendants of Indigenous women, some of these mixed families attaining high economic status and social prestige (Moore, 2014). Seixas (1853) is the earliest known usage of the term *Nheengatu* ‘good language’ to designate the LGA.

A Royal Charter of 1689 made Tupi the official language of the State of Maranhão and Grão-Pará until an analogous document in 1727 prohibited it in favor of the Portuguese language (Moore, 2014). However, as D'Angelis (2023) points out, the mere existence of a document stating a preference for a particular language does not necessarily guarantee its widespread adoption. In fact, by 1750, except for some colonial administrators who came from Portugal, the LGA was still the predominant language spoken throughout the colony (Moore, 2014). It continuously spread along the Amazon River and its tributaries, like the Rio Negro, eventually reaching Colombia and Venezuela. In the middle of the 19th century, the LGA was the most widely spoken language in the Brazilian Amazon, including larger

cities such as Belém. Documentation of Nheengatu boomed from this time until the early 20th century (Altman, 2022). On the one hand, emperor Pedro II promoted field research on Nheengatu, which resulted in the publication of oral Nheengatu literature and grammars, e.g., de Magalhães (1876). On the other, Nheengatu was part of the curriculum of the Seminary of Belém, and Church representatives produced teaching materials (Seixas, 1853; Aguiar, 1898; Costa, 1909).

The *Cabanagem* revolt (1835-1845) and mass immigration from the Northeast starting in 1877, among other factors, triggered Nheengatu's continual decline (Navarro et al., 2017). Today, as a first language, it is limited to São Gabriel da Cachoeira in the Upper Rio Negro, where it is co-official, having replaced the original non-Tupi Arawak languages of the Bare, Baniwa, and Warekena, whose languages are extinct or moribund (Eberhard et al., 2023). Nheengatu itself, with reportedly 6000 speakers in Brazil and 8000 in Colombia, where it ranks 6b and 7 on the EGIDS scale, respectively, is severely endangered, being “nearly extinct” in Venezuela, with 8b status and “[v]ery few, if any, speakers left” (Eberhard et al., 2023). Nheengatu as a contact language has also dramatically diminished (Finbow, 2020). Fortunately, diverse revitalization initiatives, e.g., in the Middle Amazon River (Lima Schwade, 2021) and the Lower Tapajós River, have targeted Nheengatu (Silva Meirelles, 2020). Besides, Indigenous people whose original languages have long gone extinct, from places as far away from the Amazon region as the Ceará State, are learning Nheengatu to affirm their ethnic identity (Filho, 2010). In 2021, the Monsenhor Tabosa municipality in Ceará adopted “Tupinheengatu” as a co-official language (Government, 2021).

All this background places Nheengatu in a unique position among the approximately 150 Indigenous languages that are still alive in Brazil, according to Storto (2019). Unlike any other, Nheengatu is supra-ethnic and has never been a tribal language (Borges, 1996; Navarro, 2012). Its influence on Brazilian Portuguese is unparalleled (de Souza Martins, 2012, 2014). Not only that, but Nheengatu has also had a significant impact on intellectuals of the stature of Mario de Andrade, Villa-Lobos, and Guimarães Rosa (Avila and Trevisan, 2015; Campoi, 2015; Pucci, 2017; Toni and Fresca, 2022). Moore (2014, p. 108) states: “Nheengatu has a notable charm. People

delight in learning it and regard it with affection.” Indeed, since the last decade, non-Indigenous learners have contributed significantly to the stock of texts in Nheengatu, e.g., by translating literary classics such as Graciliano Ramos, Saint-Exupéry, and Tolstoy (Avila, 2016; Trevisan, 2017; Costa, 2019). August 2023 marked a significant milestone: the Federal Supreme Court and the National Council of Justice published a translation of the Brazilian Constitution into Nheengatu (Lucchesi et al., 2023), making it the first Indigenous language to receive such an honor.

4 Overview of the treebank

Sentence lengths in the UD_Nheengatu-CompLin treebank range from 2 to approximately 50 words (Figure 1), with a mean and median of 10.01 and 8.0 words, respectively, and a standard deviation of 6.72, reflecting the richness found in Nheengatu texts, as represented in Table 1.

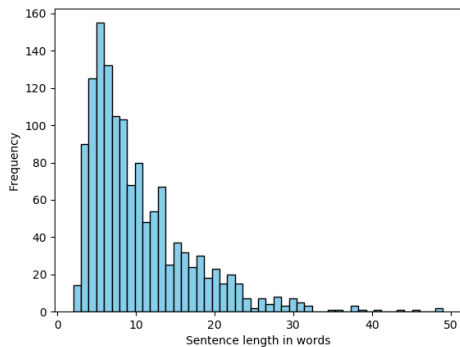


Figure 1: Frequency distribution of sentence length in the treebank.

Table 2 presents statistics for the current development version, calculated using UD’s `conllu-stats.pl` script. It includes corresponding data from UD_Guajajara-TuDeT (Gerardi et al., 2022; Martín Rodríguez et al., 2022; Santos et al., 2024), the second-largest treebank for a Brazilian Indigenous language in the UD collection. Guajajara also pertains to the Tupian family. The Nheengatu treebank surpasses the Guajajara in most dimensions (Section 4.4).

The second column of Table 1 contains the first field of the sentence identifier (Section 4.1).³ About half of the sentences stem from Avila (2021),

³NTLN2019 = do Brasil (2019), MooreFP1994 = Moore et al. (1994), TerraPreta2013 = Bird et al. (2013), Stradelli2014 = Stradelli (1929, 2014), DLGA2019 = Müller et al. (2019).

Freq.	Source	Rel. Freq.
705	Avila2021	0.5273
216	Navarro2016	0.1617
86	Magalhaes1876	0.0644
61	Cruz2011	0.0457
56	Alencar2021	0.0420
48	NTLN2019	0.0360
46	Rodrigues1890	0.0345
38	MooreFP1994	0.0285
23	Casasnovas2006	0.0173
23	Amorim1928	0.0173
16	Sympson1877	0.0120
7	TerraPreta2013	0.0052
3	Aguiar1909	0.0022
2	Stradelli2014	0.0015
2	Melgueiro2022	0.0015
2	DLGA2019	0.0015
1	Seixas1853	0.0007
1	Hartt1938	0.0007
1336	Total	1.0000

Table 1: Frequency of treebank examples per bibliographical source.

on which we have mostly based the selection of sources. With circa 8,000 lemmas and 4,000 unique examples, this is certainly the most comprehensive dictionary of a Brazilian Indigenous language, perhaps only rivaled by Navarro’s (2015) dictionary of Ancient Tupi. The entries have a rich microstructure covering semantic, grammatical, and etymological aspects, anchored in a wide-coverage research of practically all known sources of Nheengatu from the 18th to the 21st century.

Making up 16% of the treebank, the second largest group of sentences derives from de Almeida Navarro (2016). This is a self-contained coursebook with 13 lessons containing both constructed and authentic contemporary as well as historical texts, accompanied by didactic translations into Portuguese. The lessons follow a grammatical progression that facilitates the annotation. The treebank presently covers almost all examples up to the 4th lesson. Sympson (1877); Casasnovas (2006) are two other important coursebooks (Table 1).

The 3rd portion of the treebank derives from de Magalhães (1876), perhaps the most influential oeuvre of 19th-century Nheengatu literature. Rodrigues (1890); de Amorim (1928) contain analogous collections of fables and myths. da Cruz

Treebank	Sentences	Words	Lemmas	Forms	Fusions	Features	Dependency Relations
Nheengatu	1336	13374	1244	1707	89	71	36
Guajajara	1182	9160	593	1314	138	72	29

Table 2: Comparison of statistics between UD_Nheengatu-CompLin and UD_Guajajara-TuDeT.

(2011) makes up the 4th portion. This is the most comprehensive description of the phonology and grammar of 21st-century Nheengatu as spoken by the Bare, Baniwa, and Warekena in the Upper Rio Negro. The 5th treebank portion consists of a sample from the test set of constructed sentences expressing a qualifying predication, as described in de Alencar (2021). Diverse studies have shown the importance of biblical texts for NLP (McCarthy et al., 2020; Liu et al., 2021). An indispensable textual resource documenting late 20th-century Nheengatu is the New Testament translation (do Brasil, 2019), of which the treebank features 92 sentences. 44 stem from Avila (2021). We manually extracted and adapted the other 48 sentences (Table 1), such a limited number being due to annotation difficulties. The treebank contains all 38 sentences from Moore et al. (1994), a concise but fairly complete description of Nheengatu phonology and grammar based on the transcribed speech of two native speakers from the Upper Rio Negro. The examples show to what extent Nheengatu changed structurally towards Portuguese and to what extent it remained true to Tubinamba.

The treebank only contains a few examples from textual materials by Indigenous writers, e.g., Bird et al. (2013); Filho and Neto (2016); da Silva et al. (2021); Yamã et al. (2021); Melgueiro (2022) (Table 1). Incorporating more extensive passages beyond what would be considered fair use requires permission from authors. We are already contacting some of them about this.

Our ultimate goal with the treebank is to acknowledge the linguistic significance, cultural richness, and social relevance of Nheengatu, encompassing all the texts from the 19th and early 20th centuries that are in the public domain, e.g., Seixas (1853); Hartt (1872); de Magalhães (1876); Symson (1877); Rodrigues (1890); Aguiar (1898); Costa (1909); de Amorim (1928); Stradelli (1929); Hartt (1938). Apart from copyright restrictions, contemporary texts pose greater challenges to morphosyntactic annotation in the context of UD due to a lack of interlinear glossing or suitable transla-

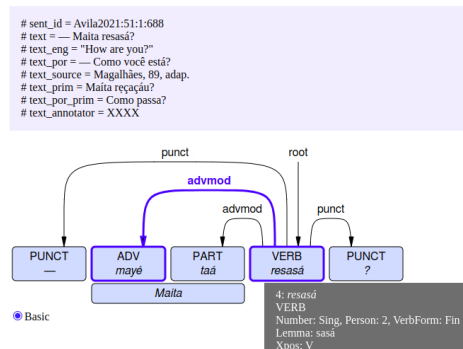


Figure 2: Dependency tree of (1) highlighting the features of the verb form *resasá*.

tions. They also exhibit strong orthographic variation and grammatical or lexical idiosyncrasies. We aim to overcome these challenges by involving Indigenous speakers with a background in linguistics in the sentence annotation workflow.

4.1 Metadata

UD does not specify a rigid scheme for metadata. The official `validate.py` script only requires the CoNLL-U files of a treebank to have two attributes: `text` and `sent_id` (Conllu). Therefore, one encounters great variability in the types and names of metadata attributes in the validated treebanks of the UD collection. In our treebank, sentences additionally have the obligatory attributes `text_eng`, `text_por`, `text_source` and `text_annotator`, which encode the English and Portuguese translations, the source of the sentence, and the annotator (Figure 2).⁴ Unless otherwise noted, we use, if available, the translation provided in the same publication we extracted the Nheengatu example from, translating it to English or Portuguese as appropriate.

`sent_id` is a unique sentence identifier, consisting of four colon-separated pieces of information, namely, (i) an abbreviation keying to the publication the sentence stems from, (ii) an integer identifying a complete text or a continuous text fragment

⁴The graph of Figure 2 was produced with <https://urd2.let.rug.nl/~kleiweg/conllu/>.

within the publication, (iii) a sequencing index for the sentence in this text, and (iv) a count number for the sentences from the same source. Examples (1)–(3) help clarify this. In (1) (respectively Figure 2) and (2), Avila2021 refers to Avila (2021).⁵ The third and second field in (1) and (2) identify the first two sentences of the 51st text fragment of the treebank stemming from Avila (2021), which are the 688th and 689th sentence from this source. In (3), 1:2 designates the second sentence of the first myth of de Magalhães (1876).

- (1) — *Maita resasá?* ‘How are you?’ (Avila2021:51:1:688) (de Magalhães, 1876, p. 89)
- (2) — *Se katuntu.* ‘I’m just fine.’ (Avila2021:51:2:689) (de Magalhães, 1876, p. 89)
- (3) *Pituna ukiri uikú í ripí-pe.* ‘The night was sleeping at the bottom of the water.’ (Magalhães1876:1:2:2)

In case of isolated sentences, the second and third fields are set to 0, see (4)–(6). Example (4) actually stems from a story but is cited without any additional context.

- (4) *Yepé paá uwapika igara gantime, amú uwapika yakumame.* ‘It is said that one was sitting in the bow of the canoe, the other was sitting in the stern.’ (Avila2021:0:0:342) (Casasnovas, 2006, p. 75)
- (5) *Setimã pinima pá.* ‘Her leg is all painted up.’ (Cruz2011:0:0:41)
- (6) *Aikú suakí.* ‘I’m close to her.’ (Navarro2016:0:0:203)

The `text_source` attribute includes various types of information that help to locate the example within the original publication. The treebank sentences from Avila (2021) simply reproduce the string in the form of a bibliographic key and a page number that accompanies the dictionary examples. For instance, the primary source of the sentence in Figure 2 is de Magalhães (1876, p. 89).

To facilitate treebank usage for a wide range of purposes, we provide additional metadata. We limit our discussion here to `text_orig` and `text_prim`. Both convey the verbatim text of an example when

⁵Boldface indicates the morphemes the tokenizer splits off, as explained in Section 4.3.

it differs from the value of `text`. The `text_orig` attribute applies to an example extracted from the source identified in the `sent_id` attribute (Figure 3), while `text_prim` indicates that the source in the `sent_id` attribute is not primary (Figure 2). A total of 37.43% of the treebank sentences have one or both of these attributes, which can be relevant for training or evaluating a language detector or a spelling converter.

4.2 Annotation methodology

The construction of a treebank for one of the Indigenous languages of Brazil is particularly challenging. A total of approximately 150 languages compete for human resources to perform this task. An annotator must be familiar not only with the lexicon and grammar structure of the particular language but also with the annotation framework. It seems that the challenge has not been so appealing to the Brazilian NLP and computational linguistics communities. The treebanks referred to in Section 2 owe their existence to the participation of foreign researchers or institutions.

At first sight, it looks like UD theory only requires high school-level knowledge of traditional concepts such as parts of speech and syntactic relations, e.g., subject, object, and indirect object. Such a simplistic view will soon vanish once one starts annotating complex sentences from authentic texts and delves into the UD documentation, where one comes across non-trivial concepts such as “open clausal complement” (`xcomp`), “depictive predicate”, etc. Familiar-sounding concepts such as “indirect object”, “apposition”, or “adverbial clause” are employed in UD in a technical sense whose understanding demands a background in syntactic theory. UD’s inventory of 17 parts of speech includes categories such as particles that are not part of the traditional descriptions of Portuguese, which are generally limited to up to ten categories (Cunha and Cintra, 1985; Macambira, 1999).

The Nheengatu treebank has been annotated by a team of three non-Indigenous annotators, consisting of a senior linguist (SLIN) and two undergrad students — of whom one (EULIN) is much more experienced in the annotation task than the other (UULIN). All three are foreign-language learners of Nheengatu. SLIN and EULIN roughly possess the grammatical and lexical knowledge of de Almeida Navarro’s (2016) coursebook. UULIN is less familiar with the language but has some knowledge of Ancient Tupi. SLIN is acquainted

```

>>> import Yauti
>>> s = '''Tapiira unhehê: – Aramé/advj aikú asú. (p. 182) A anta falou: – Então estou-me indo. -
Tapiira onhehê: – Aramé a ikô xa çô.'''
>>> Yauti.parseExample(s, 'Magalhaes1876', 2, 40, 83, annotator='XXXX')
# sent_id = Magalhaes1876:2:40:83
# text = Tapiira unhehê: – Aramé aikú asú.
# text_eng = The tapir said: – Then I'm leaving.
# text_por = A anta falou: – Então estou-me indo.
# text_source = p. 182
# text_orig = Tapiira onhehê: – Aramé a ikô xa çô.
# text_annotator = XXXX
1 Tapiira tapiira NOUN N Number=Sing 6 nsubj TokenRange=0:7
2 unhehê unhehê - - - 6 - SpaceAfter=No|TokenRange
=8:14
3 : : PUNCT PUNCT - 6 punct TokenRange=14:15
4 - - PUNCT PUNCT - 6 punct TokenRange=15:16
5 Aramé aramé ADV ADVJ AdvType=Cau 6 advmod TokenRange=17:22
6 aikú ikú VERB V Number=Sing|Person=1|VerbForm=Fin 0 root
TokenRange=23:27
7 asú sú VERB V Number=Sing|Person=1|VerbForm=Fin 6 parataxi
8 - - SpaceAfter=No|TokenRange=28:31
9 . . PUNCT PUNCT - 6 punct SpaceAfter=No|TokenRange
=31:32

```

Figure 3: Analysis of a novel example with Yauti.

with most lexical and grammatical descriptions of Nheengatu, e.g., de Magalhães (1876); Sympson (1877); Stradelli (1929); Moore et al. (1994); Casanovas (2006); da Cruz (2011); Moore (2014); de Almeida Navarro (2016); Avila (2021).

We adopted the following labor division: EULIN and UULIN annotated, respectively, 165 and 45 sentences from de Almeida Navarro (2016), all of which SLIN revised, totaling 15.7% of the treebank. EULIN and UULIN also revised 46 and 29, respectively, of each other’s sentences. SLIN annotated all 1,126 remaining sentences, i.e., 84.3% of the treebank.

All sentences were first annotated with Yauti (de Alencar, 2023) and, in case of errors, manually corrected. Typically, the annotation workflow roughly consisted of the following steps: (i) select an example for annotation; (ii) format the example; (iii) apply Yauti to the formatted example; (iv) check the resulting CoNLL-U output for any remaining ambiguities and unknown words; (v) if necessary, update Yauti’s glossary and manually annotate the example with XPOS tags or token creation functions, as described in de Alencar (2023); (vi) reapply Yauti on the example; (vii) manually correct any errors; (viii) insert the serialized CoNLL-U data in the treebank file; (ix) run `validate.py` on the file and correct any detected errors. Figure 3 exemplifies the annotation of an example. The advj XPOS tag enables disambiguation. Yauti fails to recognize the second word due to a spelling mistake, the correct form being *unheẽ* ‘it says’. Yauti also renders the Portuguese translation into English by means of Google Translate using the `deep_translator` library.⁶

⁶<https://pypi.org/project/deep-translator/>

4.3 Spelling normalization, tokenization, and lemmatization

One of the factors that hinder the development of computational tools and resources for minority languages is the lack of orthography standardization (Mager et al., 2018; Ebrahimi et al., 2023). This problem especially affects both historical and contemporary Nheengatu, an exclusively oral language until very recently.⁷ On the one hand, each of the researchers that have collected oral stories, recorded dialogues, or produced vocabularies and grammar descriptions since the 19th century coined their own spelling system, e.g., Seixas (1853); de Magalhães (1876); Sympson (1877); Rodrigues (1890); Aguiar (1898); Costa (1909); de Amorim (1928); Stradelli (1929). On the other hand, ethnic, cultural, and religious heterogeneity and geographical dispersion of the speaker communities have prevented agreement on a common system or at least a reduced number of standards. As Avila (2021) observes, not only does each publication use its own orthography, but there is often variation within a single publication. Contemporary Nheengatu has far more than the four spelling systems identified by D’Angelis (2023). We looked up seven common words, e.g., pronouns and forms of *munhã* ‘to make’, across 20 publications, about half of which were by Indigenous writers, and found out that none coincides in all spellings. For example, *yam*, *yã*, and *nyã* are variants of demonstrative *nhaã* ‘that’ in some recent publications.

Orthographic variation in Nheengatu texts results from differences not only in the mapping of phonemes onto graphemes, possibly related to di-

⁷Avila’s (2021) bibliography only includes Indigenous writers from the early 2000s onward.

alectal pronunciations, but also in word segmentation. Person and number are marked by prefixes, of which there are two series, namely, the *active, dynamic* or *verbal* (IP_A) and the *inactive, stative* or *nominal* (IP_E) (Moore et al., 1994; da Cruz, 2011; Moore, 2014; Finbow, 2020). In both historical and contemporary texts, these prefixes are sometimes spelled together, sometimes separately from their heads.⁸ For example, *semayã* ‘my mother’ in one text corresponds to *se mãya, çe mãya* and *sé manha* in other texts. This sort of variation impacts many other morphemes, with the additional complication of the use of a hyphen as a separator in some texts.

To make the construction of the treebank manageable, we decided to adopt Avila’s (2021) orthographic system (henceforth AVO) due to its practical advantages. First, it provides the most comprehensive description of the language, particularly regarding the lexicon, facilitating the lexical lookup of words in the treebank. Second, Yauti heavily relies on Avila’s (2021) lexical and grammatical information. Third, AVO closely aligns with de Almeida Navarro’s (2016) orthography, allowing those teaching or learning the language with this coursebook to easily consult the treebank. The treebank already includes 216 examples directly extracted from de Almeida Navarro (2016). Finally, AVO shares many commonalities with orthographies in use by speaker communities.

Following de Almeida Navarro (2016), Avila (2021) treats the syllabic IP_E prefixes, e.g., 1st and 3rd person singular *se* and *i* in (2) and (7), respectively, as *second class pronouns*, separating them from their heads, an approach also adopted by speakers of so-called Traditional Nheengatu (Yamã et al., 2021). In (2), the IP_E functions as an agreement marker of the stative verb *katú* ‘to be fine’, while it is a pronoun realizing the internal argument of the noun *resá* ‘eyes’ in (8) and of the postposition *irumu* ‘with’ in (7) and (9). It seems that, to properly reflect the role of the IP_E as an inflectional morpheme, *se katú* ‘I’m fine’ in (2) should be treated as a single syntactic word. While syntactic words with an internal white space are, in principle, permitted, they are discouraged by the UD guidelines (Universal Dependencies). This has led us to uniformly adopt de Almeida Navarro’s (2016) and Avila’s (2021) approach, tokenizing syllabic IP_E prefixes as separate syntactic words in all situations. By contrast, both authors treat the

relational non-contiguity prefix R² and its head as a single syntactic word, e.g., *setimã* ‘her leg’ and *suakí* ‘near her’ in (5) and (6), which we also adhere to, despite the functional parallelism with the *i* IP_E, e.g., (7).

- (7) *Makití i manha usú, usú i irumu.* ‘Where his mother went, he went with her.’ (Avila2021:14:2:158) (Rodrigues, 1890, p. 233)
- (8) *Kunhã uyumuseẽ-kwáu ixé arama, aé umurí-kwáu tẽ ixé, se resá ti amuyeréu aintá i xupé, amukití aintá uikú.* ‘A woman can sweeten herself for me, she can even please me, my eyes don’t turn to her, they are turned to the other side.’ (Avila2021:0:0:87) (de Amorim, 1928, p. 366)
- (9) — *Resú-putari se irumu?* ‘“Do you want to go with me?”’ (Avila2021:53:1:696)

In a few cases, Yauti automatically splits tokens into distinct syntactic words. In (1), the content question particle *taá* fuses with the interrogative adverb *mayé* ‘how’. In (2), we have an enclitic adverb (de Almeida Navarro, 2016). Sentences (3) and (4) exemplify the clitic allomorphs of postposition *upé* ‘in’. In (8) and (9), the capability and volition auxiliaries *kwáu* ‘can’ and *putari* ‘to want’ incorporate into the main verb (da Cruz, 2011).

Avila (2021, p. 145) goes beyond a mere spelling adaptation of usage examples from the literature. He often normalizes historical variants to align with the contemporary form in Upper Rio Negro Nheengatu. For instance, in (1), the original form *reçaçáu* transforms into *resasá*, although his dictionary also registers historical variant *sasáu*. Additionally, he adjusts original punctuation to adhere to current Portuguese conventions and undertakes various interventions to enhance readability for contemporary speakers.

In the general case, Yauti automatically carries out lemmatization. It strips off the plural suffix from nouns and pronouns and the person-number prefixes from conjugated active verbs, filling in the 3rd CoNLL-U column with the appropriate lemma and encoding the morphosyntactic properties of the affix as features in the sixth column (Figures 2 and 3). Yauti’s capabilities in this domain, however, are still restricted to inflectional morphology. To parse derivational morphology, e.g., evaluative, collective, privative, and aspectual suffixes, it is

⁸This variation affects IP_E prefixes more often.

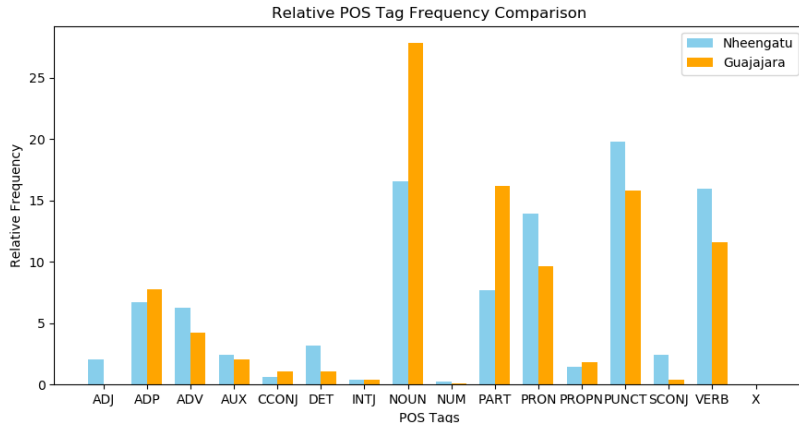


Figure 4: Relative frequency of parts of speech in UD_Nheengatu-CompLin and UD_Guajajara-TuDeT.

necessary to annotate the examples with special tags (de Alencar, 2023).

4.4 Aspects of the annotation

Of the 17 universal parts-of-speech tags (UPOS), only SYM and X have not been used because no sentence with such words has yet occurred. Usually, the tag assigned to a particular word in the treebank corresponds to Avila’s (2021) taxonomy, which mostly matches the one of da Cruz (2011). Discussing the various word classification proposals for Nheengatu is beyond the scope of this paper. We focus on property concept words (Peng, 2016), which da Cruz (2011) classifies as stative verbs. Following Moore (2014); Avila (2021), we treat these words as adjectives when they do not inflect for person and number.

Figure 4 compares the relative frequency of UPOS in UD’s Nheengatu and Guajajara treebanks. Except for adjectives, absent in the Guajajara treebank, the two tagsets coincide. The Nheengatu treebank has one feature less but much more dependency relations than the Guajajara (Table 2), which lacks, e.g., *acl:relcl*, *amod*, *cop*, *csubj*, *nmod:poss*, and *nsubj*. The two treebanks share 17 feature names, e.g., *Rel*, *Red*, and *Person[psor]* for relational prefixes, reduplication, and possessor’s person. *Clusivity* is one of the 15 feature names of Guajajara that are missing in Nheengatu, which failed to inherit this property from Tupinamba (Rodrigues, 1990, 2013). On the other hand, *Clitic*, *Compound*, *Definite*, *Deixis*, *Derivation*, *Number[psor]* and *PartType* are some of the 14 feature names of Nheengatu that are absent in Guajajara. In the UD collection, only Nheengatu pos-

sesses *Number[grnd]=Sing* and *Person[grnd]=3*, which encode the corresponding features of the internal argument of a postposition, i.e., the *landmark* or *ground* (Tosco, 2006), when it is expressed by the relational prefix R^2 , as in (6). We speculate that some of the discrepancies between the Nheengatu and Guajajara treebanks might be due to the changes the former underwent as lingua franca.

5 Parsing experiment

In this section, we report on a 10-fold cross-validation experiment with UDPipe (Straka et al., 2016; Straka and Straková, 2017). Our purpose was to assess the usefulness of the treebank for parsing, to bootstrap sentence annotation.

```

1 udpipe --train model training_file
2 udpipe --tokenize --tokenizer=ranges ↔
   ↔--accuracy --tag --parse model ↔
   ↔test_file
3 udpipe --accuracy --parse model ↔
   ↔test_file

```

Listing 1: Commands for training and testing the models.

Although UDPipe 2.0 attains better parsing results (Straka, 2018; Straka et al., 2019), due to time constraints, we limited ourselves in the experiment to the light-weight UDPipe 1.2 (Straka et al., 2016; Straka and Straková, 2017). Using the *KFold* function from the *scikit-learn* library (Pedregosa et al., 2011) with *shuffle=True* and *random_state=42* for reproducibility, we divided the treebank sentences into ten equal-sized folds, training and testing ten times, each time with a different fold as the test set and the remaining nine folds as the training set. We used the commands in Listing 1 for training and evaluating the models, which pretty much

correspond to the default settings (Straka, 2023).

While Listing 1:2 treats the test data as raw text, also performing tokenization and tagging, Listing 1:3 takes into account the gold tokenization with the gold tags. In each of the 10 executions of these commands, UDPipe 1.2. aggregates the performance results into reports like the ones in Appendix A. With Listing 1:2, accuracy in tokenization, tagging, and parsing is computed using the F1 score, i.e. the harmonic mean of precision and recall (Straka et al., 2016; Zeman et al., 2017). Tables 3 and 4 exhibit the averages of the F1 scores and standard deviations for these three dimensions computed with the NumPy library’s mean and std functions from the values of the reports generated by the ten runs of Listing 1:2. Tokenization encompasses not only splitting up text into sentences and these, in turn, into surface tokens but also two other tasks, namely, the identification of multiword tokens and syntactic words. Besides UPOS, tagging involves correctly assigning language-specific part-of-speech tags (XPOS) (Appendix B), morphological features (FEATS), and lemmas. Parsing is assessed in terms of the unlabeled attachment (UAS) and labeled attachment (LAS) scores. Table 5 presents the average UAS and LAS scores with standard deviations for parsing from gold tokenization with gold tags, computed over ten executions of Listing 1:3 as previously described. UDPipe 1.2 outperforms the rule-based Yauti morphosyntactic analyzer, which attained 80.0 and 73.2, respectively, in an analogous setting (de Alencar, 2023).

Tokenization Metric	F1 Score (%)
Tokenizer tokens	94.376 ± 1.19
Tokenizer multiword tokens	86.187 ± 10.28
Tokenizer words	94.279 ± 1.20
Tokenizer sentences	66.102 ± 4.53

Table 3: Tokenization results.

Tagging/Parsing Metric	F1 Score (%)
Tagging - UPOS	89.039 ± 1.11
Tagging - XPOS	88.16 ± 1.17
Tagging - FEATS	87.289 ± 1.17
Tagging - Lemmas	91.598 ± 1.42
Parsing - UAS	70.466 ± 1.77
Parsing - LAS	64.506 ± 1.85

Table 4: Tagging, UAS, and LAS F1 scores for parsing raw text.

	UAS (%)	LAS (%)
Average ± SD	86.30 ± 0.96	81.17 ± 1.02

Table 5: Parsing from gold tokenization with gold tags. SD = standard deviation.

6 Final remarks

We are continually revising the annotated sentences and adding new ones. We will train a model with UDPipe 2.0 to assess its impact on accelerating annotation. Given the growth rate of the UD_Nheengatu-CompLin treebank, we anticipate reaching 1800 sentences by the next UD release on May 15, 2024. A further interesting question to pursue is understanding whether the discrepancies from the other Tupian treebanks stem from Nheengatu history or theoretical preferences.

Acknowledgements

Diverse people and institutions have contributed to the version of the UD_Nheengatu-CompLin treebank presented in this paper. The Ceará State Foundation to Support Scientific and Technological Development (*Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico* — Funcap) provided scientific initiation scholarships for Vivianne Anselmo Nascimento and Dominick Maia Alexandre, who annotated and revised numerous treebank sentences. Thanks are especially due to Dominick Maia Alexandre for her diligent effort in annotating or revising more than two hundred sentences. We would like to express our immense gratitude to Eduardo de Almeida Navarro for allowing us to include the text materials from his coursebook (de Almeida Navarro, 2016) in our treebank. We are also indebted to the Federal University of Amazonas Press (*Editora da Universidade Federal do Amazonas* — UFAM), particularly to its director, Sérgio Freire, for granting permission to incorporate texts from Casasnovas (2006) into the treebank. We owe Marcel Twardowsky Avila a huge debt of gratitude. He kindly made an XML version of his Nheengatu dictionary (Avila, 2021) available to us. He generously shared his philological expertise on Nheengatu with us, clarifying many lexical and grammatical questions about the language and providing an adaptation of the myth “How the night appeared” (de Magalhães, 1876). We also thank the ongoing contributions that will mostly be reflected in the next release of the UD collection. Do-

minick Maia Alexandre and Juliana Lopes Gurgel have been carefully reviewing the annotation of a random sample of 200 sentences. Since January 2024, Juliana Lopes Gurgel has been annotating a great number of texts as part of her engagement as a scholarship holder with the DACILAT project, funded by the São Paulo State Research Support Foundation (*Fundação de Amparo à Pesquisa do Estado de São Paulo* – FAPESP) under Process No. 22/09158-5.

The final version of the paper has immensely benefited from the comments and suggestions of the two anonymous reviewers, to whom we are deeply thankful. We also acknowledge the use of AI-based tools Grammarly, Quillbot, and ChatGPT3.5 for the improvement of spelling, grammar, vocabulary, and style. We also utilized ChatGPT3.5 to speed up the writing of Python, Bash, and LaTeX code. AI suggestions were carefully examined, tested, and often corrected by us, whereby we take full responsibility for the form and content of the paper.

References

- Costa Aguiar. 1898. *Doutrina cristã destinada aos nativos do Amazonas em nhíngatu' com tradução portuguesa em face*. Pap. e Tip. Pacheco, Silva & C., Petrópolis.
- Maria Cristina Fernandes Salles Altman. 2022. *As partes da oração na tradição gramatical do Tupinambá / Nheengatu*. *Limite. Revista de Estudos Portugueses y de la Lusofonia*, 6:11–51.
- José de Anchieta. 1595. *Arte de Grammatica da Lingoa mais usada na costa do Brasil*. Antonio de Mariz, Coimbra.
- Marcel Twardowsky Avila. 2016. *Estudo e prática da tradução da obra infantil “A terra dos meninos pelados”, de Graciliano Ramos, do português para o Nheengatu*. Dissertação de mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. Acesso em: 2023-12-18.
- Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Ph.D. thesis, Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.
- Marcel Twardowsky Avila and Rodrigo Godinho Trevisan. 2015. *Jaguanhém: um estudo sobre a linguagem do Iauaretê*. *Magma*, 22(12):297–335.
- Steven Bird, Katie Gelbart, and Isaac McAlister, editors. 2013. *Fábulas de Terra Preta: Uma coletânea bilíngue*. sine nomine, Manaus.
- Luiz Carlos Borges. 1996. O nheengatú: uma língua amazônica. *Papia*, 4(2):44–55.
- Juliana Flávia de Assis Lorenção Campoi. 2015. *A literatura brasileira em nheengatu: uma construção de narrativas no século XIX*. Mestrado em literatura brasileira, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. Accessed: 2023-12-19.
- Afonso Casanovas. 2006. *Noções de língua geral ou nheengatú: gramática, lendas e vocabulário*, 2 edition. Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, Manaus.
- Paulo Cavalin, Pedro Domingues, Julio Nogima, and Claudio Pinhanez. 2023. *Understanding native language identification for Brazilian indigenous languages*. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 12–18, Toronto, Canada. Association for Computational Linguistics.
- Conllu. 2022. Conll-u format. <https://universaldependencies.org/u/overview/tokenization.html>. Accessed: 2024-01-09.
- Adriano Luis Costa. 2019. *Do português ao nheengatu: tradução da obra “De quanta terra precisa o homem?”*, de Leon Tolstoi. Dissertação de mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.
- D. Frederico Costa. 1909. *Carta pastoral de D. Frederico Costa bispo do Amazonas a seus amados diocesanos*. Typ. Minerva, Fortaleza.
- Celso Cunha and Lindley Cintra. 1985. *Nova Gramática do Português Contemporâneo*, 2 edition. Nova Fronteira, Rio de Janeiro.
- Alina da Cruz. 2011. *Fonologia e gramática do nheengatú: A língua falada pelos povos Baré, Warekena e Baniwa*. LOT, Utrecht.
- Wilmar da Rocha D’Angelis, Mateus Coimbra de Oliveira, and Michéli Carolíni de Deus Lima Schwade. 2021. Acesso ao mundo digital ou acesso digital ao mundo? *Revista Digital de Políticas Linguísticas*, 15:134–158.
- Florêncio Cordeiro da Silva, Aline da Cruz, and Ademar dos Santos Lima, editors. 2021. *Mayé yamyã bûgu: Uma abordagem sociolinguística sobre a origem do bongo*. Dom Modesto, Blumenau.
- Leonel Figueiredo de Alencar. 2021. *Uma gramática computacional de um fragmento do nheengatu / A computational grammar for a fragment of nheengatu*. *Revista de Estudos da Linguagem*, 29(3):1717–1777.
- Leonel Figueiredo de Alencar. 2023. *Yauti: A tool for morphosyntactic analysis of Nheengatu within the Universal Dependencies framework*. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 135–145, Porto Alegre, RS, Brasil. SBC.

- Eduardo de Almeida Navarro. 2009. *Anchieta, literato y humanista*. *Língua e Literatura*, 29:177–191.
- Eduardo de Almeida Navarro. 2016. *Curso de Língua Geral (nheengatu ou tupi moderno): A língua das origens da civilização amazônica*, second edition. Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, São Paulo.
- Antonio Brandão de Amorim. 1928. Lendas em Nheengatu e em Português. *Revista do Instituto Histórico e Geográfico Brasileiro*, 154(100):9–475. Tomo 100, vol. 154 (2º de 1926).
- Maria de Lurdes Zanoli. 2022. *O nheengatu de São Paulo (língua geral ou língua brasílica): para uma reconstrução da área linguística das capitânias de São Vicente e de São Paulo*. Ph.D. thesis, Universidade de São Paulo, São Paulo. Tese (Doutorado).
- José Vieira Couto de Magalhães. 1876. *O selvagem*. Typographia da Reforma, Rio de Janeiro.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.
- José de Souza Martins. 2012. Você fala nheengatu? O Estado de S. Paulo. Page C6.
- José de Souza Martins. 2014. Book flap. In Ermanno Stradelli, editor, *Vocabulário português-nheengatu, nheengatu-português*. Ateliê Editorial, Cotia.
- Missão Novas Tribos do Brasil, editor. 2019. *Novo Testamento na língua Nyengatu*, 2nd edition. Sociedade Bíblica do Brasil, Barueri, SP. Original work published in 1973.
- Carlos Drumond. 1964. Das tupi, die erste National-sprache Brasiliens. *Staden-Jahrbuch*, XI/XII:19–29.
- Wilmar da Rocha D'Angelis. 2023. *A língua Nheengatu e suas ortografias: questões técnicas e de política linguística*. *LIAMES: Línguas Indígenas Americanas*, 23(00):e023004.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, twenty-sixth edition. SIL International, Dallas.
- Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaña, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. *Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages*. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.
- Frederico G. Edelweiss. 1969. *Estudos Tupis e Tupi-Guaranis: confrontos e revisões*. Livraria Brasileira Editora, Rio de Janeiro.
- Luis Figueira. 1621. *Arte da lingva brasilica*. Manoel da Silva.
- Florêncio Almeida Baz Filho and Antônio Fernandes Góes Neto, editors. 2016. *Nheengatu Tapajowara: Livro do Projeto de Extensão Curso de Nheengatu UFOPA/GCI*, 2 edition. SELO Gráfica Editora, Santarém, PA.
- Florêncio Almeida Vaz Filho. 2010. *A Emergência étnica dos povos indígenas do baixo Rio Tapajós, Amazônia*. Ph.D. thesis, Universidade Federal da Bahia, Salvador. Tese (Doutorado). Programa de Pós-Graduação em Ciências Sociais, Faculdade de Filosofia e Ciências Humanas. Área de concentração em Antropologia.
- Thomas Finbow. 2020. *Nheengatu Dâw: A preliminary study of the phonetic, phonological and morpho-syntactic aspects of a case of Tupi-Guarani and Nadahup Contact in the Upper Rio Negro*. *Cadernos De Linguística*, 1(3):01–21.
- Thomas Finbow. 2023. *The nature and emergence of the Língua Geral Amazônica according to Mufwene's Language Ecology Model*. *Revista do GEL*, 19(2):75–112.
- José Ribamar Bessa Freire. 2011. *Rio Babel: A história das línguas na Amazônia*, second edition. EdUERJ, Rio de Janeiro.
- Fabício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2022. *TuDeT: Tupian Dependency Treebank (v0.4)*.
- Ceará Government. 2021. Língua nativa de povos indígenas é adotada como cooficial de Monsenhor Tabosa. <https://bit.ly/3RPPqgc>. Accessed: 2023-12-19.
- Charles Frederick Hartt. 1872. Notes on the Lingoa Geral or Modern Tupi of the Amazonas. *Transactions of the American Philological Association*, 3:58–76.
- Charles Frederick Hartt. 1938. Notas sobre a língua geral, ou tupí moderno do Amazonas. *Anais da Biblioteca Nacional do Rio de Janeiro*, LI:305–390. [1929].
- Fabiana Raquel Leite. 2013. *A Língua Geral Paulista e o "Vocabulário Elementar da Língua Geral Brasílica"*. Master's thesis, Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas, SP. Dissertação (mestrado).
- Michéli Carolíni de Deus Lima Schwade. 2021. *"Tupi" do Rio Andirá: o Nheengatu no Médio Rio Amazonas*. Tese (doutorado), Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas, SP. Accessed: 17 dez. 2023.

- Ling Liu, Zach Ryan, and Mans Hulden. 2021. [The usefulness of Bibles in low-resource machine translation](#). In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. Association for Computational Linguistics.
- Marco Lucchesi, José Ribamar Bessa Freira, Luis Geraldo Sant’Ana Lanfredi, Andréa Jane Silva de Medeiros, and Luanna Marley, editors. 2023. *Mundu Sa Turusu Waá : Ubêuwa Mayé Míra Itá Uikú Arãma Purãga Iké Brazíu Upé*. Supremo Tribunal Federal, Conselho Nacional de Justiça, Brasília.
- José Rebouças Macambira. 1999. *Estrutura Morfosintática do Português*, 9 edition. Pioneira, São Paulo.
- Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. [Challenges of language technologies for the indigenous languages of the Americas](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F. Gerardi. 2022. [Tupían language resources: Data, tools, analyses](#). In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.
- Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. [The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.
- Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.
- Edilson Martins Melgueiro. 2022. *O Nheengatu de Stradelli aos dias atuais: uma contribuição aos estudos lexicais de línguas Tupi-Guaraní em perspectiva diacrônica*. Ph.D. thesis, Universidade de Brasília.
- Denny Moore. 2014. Historical development of Nheengatu (Língua Geral Amazônica). In Salikoko S. Mufwene, editor, *Iberian Imperialism and Language Evolution in Latin America*, pages 108–142. University of Chicago Press, Chicago.
- Denny Moore, Sidney Facundes, and Nádia Pires. 1994. [Nheengatu \(Língua Geral Amazônica\), its history, and the effects of language contact](#). In *Proceedings of the Meeting of the Society for the Study of the Indigenous Languages of the Americas, July 2-4, 1993 and the Hokan-Penutian workshop, July 3, 1993*, Report / Survey of California and other Indian Languages ; 8, pages 93–118, Berkeley, CA. [University of California].
- Jean-Claude Muller, Wolf Dietrich, Ruth Monserrat, Cândida Barros, Karl-Heinz Arenz, and Gabriel Prudente, editors. 2019. *Dicionário de Língua Geral Amazônica*. Universitätsverlag Potsdam – Museu Paraense Emílio Goeldi, Potsdam – Belém/Pará. Primeira transcrição por Gabriel Prudente. Edição diplomática, revisada e ampliada com comentários e anexos por Wolf Dietrich, Ruth Monserrat e Jean-Claude Muller.
- Eduardo Navarro, Marcel Ávila, and Rodrigo Trevisan. 2017. [O Nheengatu, entre a vida e a morte: A tradução literária como possível instrumento de sua revitalização lexical](#). *Revista Letras Raras*, 6(2):9–29.
- Eduardo de Almeida Navarro. 2012. O último refúgio da língua geral no Brasil. *Estudos Avançados*, 26(76):245–254.
- Eduardo de Almeida Navarro. 2015. *Dicionário tupi antigo, a língua indígena clássica do Brasil: vocabulário português-tupi e dicionário tupi-português, tupinismos no português do Brasil, etimologias de topônimos e antropônimos de origem tupi*. Global.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. [Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik](#). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Siyao Peng. 2016. [Property concept words in six Amazonian languages](#). Master’s thesis, Faculty of Humanities, Leiden University, Linguistics (MA).
- M. D. Pucci. 2017. [Influência da voz indígena na música brasileira](#). *Música Popular em Revista*, 4(2):5–30. Accessed: 19 dez. 2023.

- Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. [Universal Dependencies for western sierra Puebla Nahuatl](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.
- Aryon D. Rodrigues. 1990. You and I = neither you nor I: The personal system of Tupinambá (Tupi-Guaraní). In Doris L. Payne, editor, *Amazonian Linguistics: Studies in Lowland South American Languages*, pages 393–405. University of Texas Press, Austin.
- Aryon Dall’Igna Rodrigues. 1996. As línguas gerais sul-americanas. *Papia*, 4(2):6–18.
- Aryon Dall’Igna Rodrigues. 2013. [Some cases of regrammaticalization in Tupí-Guaraní languages](#). *Revista Brasileira de Linguística Antropológica*, 2(2):231–240.
- Aryon Dall’Igna Rodrigues and Ana Suely Arruda Câmara Cabral. 2011. [A contribution to the linguistic history of the língua geral amazônica](#). *ALFA: Revista de Linguística*, 55(2).
- João Barbosa Rodrigues. 1890. *Poranduba amazonense ou kochiyma-uara porandub, 1872-1887*. Typ. de G. Leuzinger & Filhos, Rio de Janeiro.
- Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. [Apurinã Universal Dependencies treebank](#). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 28–33, Online. Association for Computational Linguistics.
- Luana Luiza Santos, Carolina Coelho Aragon, and Fabrício Gerardi. 2024. [Línguas minoritárias e anotações sintáticas de corpora: experiências de pesquisa na iniciação científica](#). *Letras de hoje*, 59(1):1–9. Published: 2024-01-10.
- Manoel Justiniano de Seixas. 1853. *Vocabulario da lingua indigena geral para o uso do Seminario Episcopal do Pará*. Typ. de Mattos e Comp^a., Pará.
- Sâmela Ramos da Silva Silva Meirelles. 2020. [A reinserção de uma língua destituída: o Nheengatu no Baixo Tapajós](#). Ph.D. thesis, Universidade Estadual de Campinas, Instituto de Estudos da Linguagem. Accessed: 17 dez. 2023.
- Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. [Assessing digital language support on a global scale](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Luciana Raccanello Storto. 2019. *Línguas indígenas: tradição, universais e diversidade*. Mercado de Letras, Campinas, SP.
- Ermanno Stradelli. 2014. *Vocabulário português-nheengatu, nheengatu-português*. Ateliê Editorial, Cotia, SP. Original work published in 1929.
- Ermanno Stradelli. 1929. Vocabulários da lingua geral portuguez-nheengatú e nheengatú-portuguez, precedidos de um esboço de Grammatica nheenga-umbuê-sáua mirí e seguidos de contos em lingua geral nheengatú poranduaa. *Revista do Instituto Historico e Geographico Brasileiro*, 158(104):9–768.
- Milan Straka. 2018. [UDPipe 2.0 prototype at CoNLL 2018 UD shared task](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.
- Milan Straka. 2023. *UDPipe 1 User’s Manual*.
- Milan Straka, Jan Hajič, and Jana Straková. 2016. [UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).
- Milan Straka and Jana Straková. 2017. [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*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Milan Straka, Jana Straková, and Jan Hajic. 2019. [UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging](#). In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.
- Pedro Luiz Sympson. 1877. *Grammatica da lingua brazílica geral, fallada pelos aborigines das provincias do Pará e Amazonas*. Typographia do Commercio do Amazonas, Manaus.
- The Chicago Manual of Style Online. 2024. Capitalization. <https://www.chicagomanualofstyle.org/qanda/data/faq/topics/Capitalization/faq0106.html>. Accessed: February 6, 2024.
- Guillaume Thomas. 2019. [Universal Dependencies for Mbyá Guaraní](#). In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France. Association for Computational Linguistics.
- Flávia Camargo Toni and Camila Fresca. 2022. [Natureza e modernismo: Mário de Andrade e Villa-Lobos antes da Semana](#). *Estudos Avançados*, 36(104):143–183.

- Mauro Tosco. 2006. Towards a geometry of adpositional systems: A preliminary investigation of Gawwada. In Pier Giorgio Borbone, Alessandro Mengozzi, and Mauro Tosco, editors, *Loquentes Linguis: Studi Linguistici e Orientali in Onore di Fabrizio Pennacchietti*, pages 695–702. Harrassowitz, Wiesbaden.
- Rodrigo Godinho Trevisan. 2017. *Tradução comentada da obra “Le Petit Prince”, de Antoine de Saint-Exupéry, do francês ao nheengatu*. Dissertação de mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. Acesso em: 2023-12-18.
- Francis M. Tyers and Robert Henderson. 2021. A corpus of k’iche’ annotated for morphosyntactic structure. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.
- Universal Dependencies. 2023. Tokenization and word segmentation. <https://universaldependencies.org/u/overview/tokenization.html>. Accessed: 2023-12-25.
- Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. *Toward Universal Dependencies for Shipibo-konibo*. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium. Association for Computational Linguistics.
- Irina Wagner, Andrew Cowell, and Jena D. Hwang. 2016. *Applying Universal Dependency to the Arahapaho language*. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 171–179, Berlin, Germany. Association for Computational Linguistics.
- Yaguarê Yamã, Elias Yaguakãng, Egídia Reis, and Mario José. 2021. *Dicionário e estudo de nheengatu tradicional*, 2 edition. Cintra, São Paulo.
- Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielè Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arican, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Ben-goetxea, İbrahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnè Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilaraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Drostanova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaz Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Franssen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Gričiūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guillier, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mý, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ȑlájídé Ishola, Artan Islamaj, Kaoru Ito, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşikara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóga, Andre Kåsen, Tolga Kayadelen,

Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdulatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyên Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayo Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Pacosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Pheilan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker,

Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Riebler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roşca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoal Sadde, Pegah Safari, Aleksí Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinhór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umot Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Hórdarson, Vilhjálmur Hósteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilian Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2023. [Universal dependencies 2.13](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Uřešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. [CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Example parsing reports

Listings 2 and 3 display the commands of the first run of the ten-fold cross-validation, which generate the reports reproduced further below.

```
udpipe --tokenize --tokenizer=ranges ←
      ←--accuracy --tag --parse model1. ←
      ←output test1.conllu
```

Listing 2: First run of the 10-fold cross-validation: parsing from raw text.

Number of SpaceAfter=No features in gold data: 373

Tokenizer tokens - system: 1277, gold: 1313, precision: 96.01%, recall: 93.37%, f1: 94.67%

Tokenizer multiword tokens - system: 10, gold: 11, precision: 90.00%, recall: 81.82%, f1: 85.71%

Tokenizer words - system: 1287, gold: 1324, precision: 95.96%, recall: 93.28%, f1: 94.60%

Tokenizer sentences - system: 104, gold: 134, precision: 77.88%, recall: 60.45%, f1: 68.07%

Tagging from plain text (CoNLL17 F1 score) - gold forms: 1324, upostag: 90.00%, xpostag: 89.70%, feats: 88.70%, alltags: 87.09%, lemmas: 92.38%

Parsing from plain text with computed tags (CoNLL17 F1 score) - gold forms: 1324, UAS: 73.54%, LAS: 67.79%

```
udpipe --accuracy --parse model1. ←
      ←output test1.conllu
```

Listing 3: First run of the 10-fold cross-validation: Parsing with gold tokenization and gold tags.

Parsing from gold tokenization with gold tags - forms: 1324, UAS: 87.39%, LAS: 83.23%

B Language-specific tagset (XPOS)

XPOS	Abbreviation	Abbreviation expansion
A	adj.	first class adjective
A2	adj. 2 ^a cl.	second class adjective
ADP	postp.	postposition
ADV	adv.	adverb
ADVA	adv. manner	adverb of manner
ADVC	adv. loc.	locative adverb
ADVD	adv. dem.	demonstrative adverb
ADVDI	adv. dem. dist.	distal demonstrative adverb
ADVDX	adv. dem. prox.	proximal demonstrative adverb
ADVG	adv. gr.	degree adverb
ADVJ	adv. conj.	causal conjunctive adverb
ADV L	adv. rel.	relative adverb
ADVLA	adv. rel. man.	manner relative adverb
ADVLC	adv. rel. loc.	locative relative adverb
ADVLT	adv. rel. temp.	temporal relative adverb

Table 6: XPOS tags (part 1).

Tables 6, 7 and 8 explain UD_Nheengatu-CompLin’s language-specific part-of-speech tags (XPOS) as employed in Yauti’s full-form [lexicon](#). The second column reproduces the Portuguese abbreviations for word classes of Yauti’s [glossary](#), which are fully translated into English in the third column.

XPOS	Abbreviation	Abbreviation expansion
ADVM	adv. mod.	modal adverb
ADVNC	adv. ind. loc.	indefinite locative adverb
ADVNT	adv. ind. temp.	indefinite temporal adverb
ADVO	adv. ord.	ordinal adverb
ADVP	adv. conj. opos.	concessive conjunctive adverb
ADVR	adv. interr.	interrogative adverb
ADVRA	adv. interr. man.	manner interrogative adverb
ADVRC	adv. interr. loc.	locative interrogative adverb
ADVRT	adv. interr. temp.	temporal interrogative adverb
ADVRU	adv. interr. caus.	causal interrogative adverb
ADVS	adv. intens.	intensity adverb
ADVT	adv. temp.	temporal adverb
AFF	part. affirm.	affirmation particle
ART	art. indef.	indefinite article
ASSUM	part. assum.	assumption particle
AUXFR	aux. flex. pre.	preverbal inflected auxiliary
AUXFS	aux. flex. post.	postverbal inflected auxiliary
AUXN	aux. non-flex.	noninflected auxiliary
CARD	num. card.	cardinal numeral
CCONJ	cconj.	coordinating conjunction
CERT	part. cert.	certainty particle
CLADP	postp. encl.	enclitic postposition
CLADV	adv. encl.	enclitic adverb
COND	part. cond.	conditional particle
CONJ	conj.	conjunction
CONS	part. cons.	consent particle
COP	cop.	copula verb
CQ	part. interr. cont.	content question particle
DEM	pron. dem.	demonstrative pronoun
DEMS	pron. dem. dist.	distal demonstrative pronoun
DEMSN	pron. dem. dist. non-flex.	noninflected distal demonstrative pronoun
DEMX	pron. dem. prox.	proximal demonstrative pronoun
EMP	pron. enf.	emphasis pronoun
EXST	part. exist.	existential particle
FOC	part. focus	focus particle
FRUST	part. frust.	frustrative particle
FUT	part. fut.	future particle
IMPF	part. imperf.	imperfective particle
IND	pron. indef.	indefinite pronoun
INDQ	pron. quant.	indefinite quantifier pronoun
INT	pron. interr.	interrogative pronoun
INTJ	interj.	interjection
MOD	part. mod.	modal particle
N	s.	common noun
NEC	part. neces.	necessity deontic particle

Table 7: XPOS tags (part 2).

Tag	Abbreviation	Abbreviation expansion
NEG	part. neg.	negation particle
NEGI	part. neg. imp.	negative imperative particle
ORD	num. ord.	ordinal numeral
PART	part.	particle
PFV	part. perf.	perfective particle
PQ	part. interr. pol.	polar question particle
PREC	part. prec.	precatative particle
PREF	pref.	prefix
PREP	prep.	preposition
PRET	part. pret.	past tense particle
PRON	pron.	first class pronoun
PRON2	pron. 2 ^a cl.	second class pronoun
PROPN	s. próprio	proper noun
PROTST	part. prot.	protestative particle
PRSV	part. pres.	presentative particle
REL	pron. rel.	relative pronoun
RELF	pron. rel. livre	free relative pronoun
RPRT	part. report.	reportative particle
SCONJ	sconj.	postverbal subordinating conjunction
SCONJR	sconj. pre.	preverbal subordinating conjunction
SUFF	suf.	suffix
TOT	pron. quant. univ.	universal quantifier pronoun
TOTAL	part. tot.	totalitive particle
V	v.	first class verb
V2	v. 2 ^a cl.	second class verb
V3	v. 3 ^a cl.	third class verb
VSUFF	v. suff.	noninflectionable suffixal verb

Table 8: XPOS tags (part 3).

Network-based Approach for Stopwords Detection

Felermينو D. M. A. Ali^{1,3}, Gabriel de Jesus²
Henrique Lopes Cardoso², Sérgio Nunes², Rui Sousa-Silva³

¹Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)

²Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC)

³Centro de Linguística da Universidade do Porto (CLUP)

{up202100778, hlc, sergio.nunes}@fe.up.pt

gabriel.jesus@inesctec.pt, rssilva@letras.up.pt

Abstract

Stopword lists, an essential resource for natural language processing and information retrieval, are often unavailable for low-resource languages. Creating these lists is time-consuming and expensive, making automated stopwords detection a viable alternative. This paper introduces a novel stopwords detection approach that exploits the topological properties of co-occurrence networks to identify function words. By leveraging the connectivity patterns of function words in these networks, the proposed approach aims to achieve higher precision compared to traditional frequency-based methods. To assess the effectiveness of the network-based approach, we constructed co-occurrence networks for Tetun and Emakhuwa (low-resourced languages), as well as English and Portuguese. We then compared the performance of this approach with traditional frequency-based methods. The results indicate that the network-based approach consistently outperforms traditional methods, with in-degree emerging as the most reliable indicator of function words. This finding suggests promising prospects for automatically generating stopwords lists in other low-resource languages, paving the way for developing natural language processing tools for these linguistic contexts.

Keywords: Stopwords detection, Low-resource languages, Tetun, Emakhuwa.

1 Introduction

In natural language processing (NLP) and information retrieval (IR), stopwords are function words, such as prepositions, pronouns, and conjunctions, that are frequently removed due to their high frequency and minimal information content. A common practice in various NLP and IR tasks is to remove stopwords during the preprocessing stage to reduce execution time, enhance overall performance, and improve the effectiveness of retrieval

systems (Croft et al., 2009).

Many existing methods often either rely on a predefined list of stopwords or are computed using traditional unsupervised methods, such as term frequency (TF) (Baeza-Yates and Ribeiro-Neto, 2011; Croft et al., 2009), normalized inverse document frequency (NIDF) (Lo et al., 2005), inverse document frequency (IDF), term frequency-inverse document frequency (TF-IDF), and term and document frequency (TDF) (Ferilli, 2021). Considering the observed high topological connectivity of function words in network properties (Chen et al., 2018; Gao et al., 2014; Liang et al., 2009), our objective is to investigate the application of co-occurrence networks' properties for automated stopwords detection. This investigation is grounded on the assumption that the attributes of a co-occurrence network may prove more effective than traditional unsupervised frequency-based approaches in stopwords detection tasks.

We employed co-occurrence network methods to validate the assumption, assuming two sequential terms in the text corpus form pairs of linking nodes and the graph is directed (see Figure 1). The datasets used for the experiment were collected from four languages: English, Portuguese, Tetun, and Emakhuwa. Tetun is one of Timor-Leste's official languages alongside Portuguese, spoken by 79.04% of a 1.17 million population (de Jesus, 2023), while Emakhuwa, also known as Makua, Macua, or Makhwa, is a Bantu language primarily spoken in the northern and central areas of Mozambique with an estimated 7 million speakers (Ali et al., 2021).

Subsequently, these datasets underwent preprocessing to align with the requirements of our task. Following that, we constructed the directed co-occurrence network and evaluated the effectiveness of network attributes against traditional unsupervised frequency-based methods using precision at different levels. The results demonstrated that

our proposed approach outperformed all frequency-based methods for stopword detection in both high- and low-resource languages. This outcome implies the adaptability and applicability of our solution to automate the stopword list construction process in other low-resource languages, providing a valuable and efficient tool for language processing tasks in diverse linguistic contexts.

The remaining sections of this paper are organized as follows. Section 5 describes related works. The approach is outlined in Section 2. Then, Section 3 presents the experiment and evaluation. Section 4 presents the results obtained and their discussion. Finally, Section 6 summarizes our conclusion and possible future work.

2 Approach

We experimented with English, Portuguese, and two low-resourced languages (Emakhuwa and Tetun). We pre-processed the raw text data in both cases, then constructed a co-occurrence network by generating a directed graph $G = (V, A)$ from pre-processed text data. V denotes the nodes (i.e., word types) and A the edges. Each node $v \in V$ corresponds to a word from the vocabulary of the pre-processed text, whereas $a \in A$ is an adjacency arc, which represents the connection between a pair of consecutive words, from the first to the second.

Since we are interested in analyzing the network properties and unsupervised approaches concerning stopword detection, we compute the following node (i.e., word type) attributes:

1. **Network properties:** degree, indegree, outdegree, weighted indegree, weighted outdegree, closeness centrality, harmonic closeness centrality, eccentricity, and betweenness centrality.
2. **Traditional unsupervised methods:** term frequency (TF), normalized term frequency (NTF), inversed document frequency (IDF), document frequency (DF), normalized inverse document frequency (NIDF), term frequency - normalized inverse document frequency (TF-IDF), and term-document frequency (TDF).

For convention purposes, we use the following notation: C corresponds to the corpus of each language, and $n = |C|$ the number of sentences of corpus C , while $V = t_1, \dots, t_m$ is the vocabulary of C , i.e., its word types (unique words) in C . For

each term $t_i \in V$, o_i corresponds to the number of occurrences in C , n_i to the number of sentences in which it appears, and o_i^c to the number of occurrences in sentence $c \in C$. Thus, the total number of tokens in C is given by $o = \sum_i o_i$. In addition, from the network perspective, we denote M as an adjacency matrix, which contains boolean values indicating if there is a direct link between node i and j , for $m_{ij} \in M$.

The detail of network properties is explained in Section 2.1. Section 2.2 presents the traditional metrics. Finally, Sections 2.3 and 2.4 describe the data collection and preparation processes.

2.1 Network properties

The details of network properties are the following:

In-Degree The in-degree of node t_i is the total number of connections onto node i , and is the sum of the i th row of the adjacency matrix M :

$$t_i^{in} = \sum_j m_{ij} \quad (1)$$

Out-Degree The out-degree of node t_i is the total number of connections coming from node i , and is the sum of the i th column of the adjacency matrix M :

$$t_i^{out} = \sum_j m_{ij} \quad (2)$$

Degree The sum of all connections to the nodes t_i .

$$t_i^{degree} = t_i^{in} + t_i^{out} \quad (3)$$

Average Degree is simply the mean of all the node degrees in a network.

$$\frac{1}{n} \sum_{i=1}^n t_i^{degree} \quad (4)$$

Average Weighted Degree sum of the weights of all links attached to node i .

$$\frac{\sum_{i=1}^n \sum_{j=1}^n w(i, j)}{n} \quad (5)$$

Average Path Length the average length of shortest paths between any two nodes.

$$\frac{\sum_{i=1}^n \sum_{j=1}^n d(i, j)}{n(n-1)} \quad (6)$$

Where $d(i, j)$ is the shortest path length between nodes i and j , and n is the number of nodes in the network.

Original	Eusébio (25 january 1942 – 5 january 2014) was a Portuguese football player. He was born in Mozambique
Pre-processed	eusébio january january he was a portuguese football player he born in mozambique
Vocabulary	eusébio, january, he, was, a, portuguese, football, player, born, in, mozambique

Table 1: Text pre-processing example.

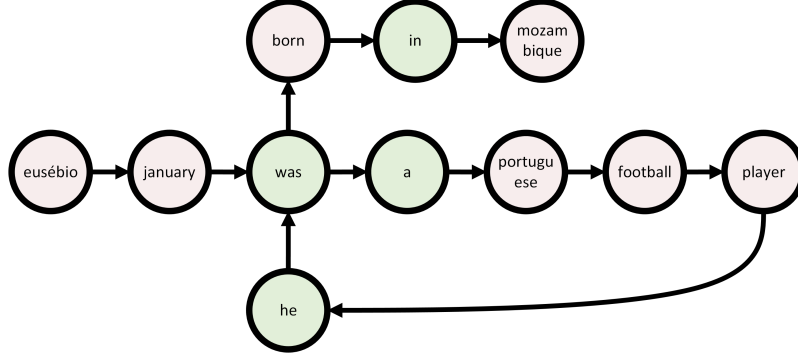


Figure 1: Co-occurrence network generated from the pre-processed text in Table 1.

Betweenness Centrality Measure the amount of influence that a node t_i has over the flow of information in the network:

$$C_B(i) = \sum_{j < k} g_{jk}(j)g_{jk} \quad (7)$$

When Normalized:

$$C_B(i) = \frac{\sum_{j < k} \frac{g_{jk}(j)}{g_{jk}}}{(N-1)(N-2)} \quad (8)$$

Where g_{jk} denotes the number of shortest paths connecting nodes j and k , $g_{jk}(i)$ is the number of those paths that pass through node i , and N is the number of nodes in the giant component.

2.2 Traditional unsupervised methods

The conventional frequency-based unsupervised approaches commonly used in stopword detection tasks are the following:

Term Frequency (TF) The amount of times a term appears in the corpus:

$$tf(t_i) = o_i \quad (9)$$

Document Frequency (DF) The number of sentences in which a term occurs:

$$df(t_i) = n_i \quad (10)$$

Normalized Term Frequency (NTF) TF normalized in accordance with the number of tokens in the corpus as a whole:

$$ntf(t_i) = -\log\left(\frac{o_i}{O}\right) \quad (11)$$

Inverse Document Frequency (IDF) (Church and Gale, 1995) Based on how frequently the term occurs in the corpus of sentences, the more it appears in sentences, the less information there is:

$$idf(t_i) = \log\left(\frac{n}{n_i}\right) \quad (12)$$

Normalized IDF (NIDF) (Robertson and Jones, 1976) IDF adjusted by 0.5 to reduce extreme values in relation to the number of sentences that do not contain the term ($n - n_i$):

$$nidf(t_i) = \log\left(\frac{(n - n_i) + 0.5}{n_i + 0.5}\right) \quad (13)$$

Term frequency-inverse document frequency (TF-IDF) (Sparck Jones, 1972) The product of NTF and NIDF:

$$TF * IDF(t_i) = ntf(t_i) \times nidf(t_i) \quad (14)$$

Term Document Frequency (TDF) (Ferilli, 2021) Amount of times a term appears in the corpus times the number of sentences in which it appears:

$$tdf(t_i) = o_i \cdot n_i \quad (15)$$

2.3 Data Collection

We collected the datasets for Portuguese, English, Tetun, and Emakhuwa from different data sources. For Portuguese and English, we used WikiCLIR (Sasaki et al., 2018), a dataset containing Wikipedia articles from 25 different languages. It is one of the large-scale datasets constructed from Wikipedia’s articles, making it ideal for our experiment as Wikipedia articles have wide coverage in terms of topics.

On the other hand, unlike English and Portuguese, datasets for Tetun and Emakhuwa are not easily accessible. That is why Tetun’s dataset is entirely composed of news articles. These articles were extracted from two news online portals in Timor-Leste, Timor News¹; and Tatoli². We scraped the content using BeautifulSoup³ and then built the corpus.

The Emakhuwa dataset is composed of a collection of the following data: Emakhuwa side of the parallel corpora from Ali et al. (2021) as well as Wikipedia articles⁴, radio transcripts, and Emakhuwa’s translation of the Constitution of the Republic of Mozambique. The details of the datasets are shown in Table 2.

2.4 Data Preparation

Since the datasets are collected from various data sources, we pre-processed them to exclude unnecessary characters and reduce the data size to be more efficient. The pre-processing task comprises lowercasing, tokenization, and removing punctuation, special characters, and numbers.

For Portuguese, English, and Emakhuwa, we used the general white space and punctuation-based tokenizer from NLTK⁵ and a Tetun tokenizer⁶ was used for Tetun.

Extra cleaning was done to reduce the vocabulary in Portuguese and English, as we noticed that Wikipedia articles typically contain a mixture of words from different languages. This happens because Wikipedia documents are usually translations of articles originally written in a different language, so some words, such as names and nouns,

¹<https://www.timornews.tl>

²<https://tatoli.tl>

³<https://www.crummy.com/software/BeautifulSoup/>

⁴<https://incubator.wikimedia.org/wiki/Category:Wp/vmw>

⁵<https://www.nltk.org/>

⁶<https://pypi.org/project/tetun-tokenizer/>

are kept as they are in the source languages. Thus, we removed all unknown terms from the language’s word list to reduce the vocabulary. For Portuguese, we removed terms that do not appear in the Natura (University of Minho, 2021) dictionary, whereas for English, we used the english-words (Wiens, 2021) dictionary. The statistics for each network are displayed in Table 3.

3 Experiment and Evaluation

Our experiments consist of, first, conducting feature selection to reduce our analysis to the most relevant network attributes. After that, we subdivided the experiments into high-resource languages (English and Portuguese) and low-resource languages (Tetun and Emakhuwa). The experiments and evaluation are described in the following subsections.

3.1 Feature Selection

For each term in English and Portuguese vocabulary (i.e., nodes), we provided a Boolean value as true if they correspond to an actual stopword and false if they do not. For that, we used the stopword list available in the NLTK toolkit. Then, we load nodes and edges (i.e., co-occurrence) on Gephi⁷ using the edge sum strategy and then computed the network’s properties mentioned in Section 2.

To select the most relevant network features, we evaluated each network attribute with respect to information gain related to the target class (i.e., stopword or not stopword). Here, we use the Weka data mining toolkit (Frank et al., 2016) aiming to reduce our analysis to the top four relevant features. Weka calculates the information gain with respect to the target class by using the following formula:

$$InfoGain(c, a) = H(c) - H(c|a) \quad (16)$$

Where $H(c, a)$ is the information for the dataset, c is the target class, a is the attribute, $H(c)$ is the entropy for the dataset before any change, and $H(c|a)$ is the conditional entropy for the dataset given the attribute a .

3.2 Portuguese and English

We run stopwords filtering based on the values computed from attributes in Section 2, following two strategies:

⁷<https://gephi.org/>

Language	#Documents	Data source(s)
English	50,000	WikiCLIR
Portuguese	50,000	WikiCLIR
Tetun	32,666	Timor News and Tatoli portals
Emakhuwa	52,238	Wikipedia, parallel corpora (Ali et al., 2021), radio transcripts, Mozambican’s Constitution

Table 2: Dataset details. Documents are each line or paragraph of the corpus.

Measure	English	Portuguese	Tetun	Emakhuwa
# Nodes	18,369	57,960	22,111	53,880
# Edges	482,828	1,041,846	229,675	292,578
Diameter	7	11	21	17
Avg. Degree	26.285	17.976	10.387	5.43
Avg. Weighted Degree	150.934	84.277	49.534	14.215
Avg. Path Length	2.57	4.159	3.451	4.858

Table 3: Network statistics.

- Descending order (high to low): TF, NTF, DF, betweennesscentrality, indegree, outdegree, and degree.
- Ascending order (low to high): IDF, NIDF, and TF-IDF.

For evaluation, we adopt Precision N top position ($P@N$), which is given as the fraction of words that are stopwords:

$$P@N = \frac{\text{stopwords}}{N} \quad (17)$$

According to NLTK toolkit’s stopword list, English has 126 stopwords, whereas Portuguese has 176. So, to evaluate precision, we used cutoff values from 25 ($P@25$) to 200 ($P@200$) with an interval of 25.

3.3 Tetun and Emakhuwa

Tetun and Emakhuwa do not have a ground truth list of stopwords, leading us to adopt the approach outlined in Section 3.2 for the stopwords filtering process. To assess precision, we strategically chose to translate the top 50 words into English, taking into account the widespread understanding and use of the English language. Following translation, each word was compared to entries in the English stopword list; if there was a match, it was classified as a stopword; otherwise, we considered it not to be a stopword.

4 Result and Discussion

We summarize our experimental results in Portuguese and English and discuss them in Sec-

tion 4.1. In Section 4.2 is our observation of the approaches applied to Tetun and Emakhuwa.

Rank	Portuguese	Score
1	Betweenness Centrality	0.01223
2	Indegree	0.01162
3	Degree	0.01137
4	Outdegree	0.01123
5	TF-IDF	0.00988
6	Weighted Outdegree	0.00985
7	TF	0.00983
8	NTF	0.00983
9	IDF	0.00982
10	NIDF	0.00982
11	DF	0.00982
12	Weighted Degree	0.00979
13	Weighted Indegree	0.00973
14	TDF	0.00969
15	Harmonic Closeness Centrality	0.00753
16	Closeness Centrality	0.00730
17	Eccentricity	0.00128

Table 4: Feature ranking with information gain.

4.1 Portuguese and English

Table 5 and Table 4 provides the results of the importance of network features for stopword detection. For English, degree was the most relevant attribute, followed by in-degree. On the other hand, in Portuguese, betweenness centrality was at the top of the ranking. In-degree followed next, making a more stable feature as it ranked second in English and Portuguese. Then, degree and out-degree followed the list.

Based on these results, the betweenness centrality, indegree, outdegree, and degree are selected for the remaining experiments.

Table 6 shows results with the English network, where in-degree obtained the highest scores for all

Rank	English	Score
1	Degree	0.03525
2	Indegree	0.03515
3	DF	0.03342
4	NIDF	0.03330
5	IDF	0.03330
6	Outdegree	0.03298
7	TF	0.03272
8	Weighted Outdegree	0.03269
9	TF-IDF	0.03252
10	Weighted Degree	0.03238
11	NTF	0.03210
12	TDF	0.03170
13	Betweenness Centrality	0.03150
14	Weighted Indegree	0.03038
15	Harmonic Closeness Centrality	0.02912
16	Closeness Centrality	0.02795
17	Eccentricity	0.00632

Table 5: Feature ranking with information gain.

approaches. Similar results can be found on degree, the second best performing approach, outperforming all frequency-based approaches except $P@175$ and $P@200$. Betweenness centrality, on the other hand, has provided results very close to frequency-based approaches; however, it has small margin advantages in $P@25$. Finally, out-degree was shown to be unworthy of English as it scored slightly below the frequency-based approaches.

Table 7 shows the results in Portuguese. Here, the network-based approaches have a clear advantage over frequency-based approaches. Betweenness centrality attained the highest scores, except for precision at the top 75 words. However, regarding complexity, betweenness centrality was the least efficient approach as it takes approximately $O(N^3)$ (Barthelemy, 2004) to compute the values. In-degree and degree, on the other hand, obtained similar results to betweenness centrality, where the degree was better than in-degree by small margins. This can be explained by the fact that degree is the summation of in-degree and out-degree, so the degree is always proportional to both in- and out-degree. However, like English, out-degree performed worse than all other network-based approaches.

Overall, our results support the claims by Gao et al. (2014); Chen et al. (2018); Liang et al. (2009) that stopwords are highly connected nodes in a co-occurrence network. However, we further suggest that degree and in-degree are more effective approaches for stopwords detection than unsupervised frequency-based ones.

4.2 Tetun and Emakhuwa

To investigate the precision of network-based properties for stopwords detection further, we also experimented on Tetun and Emakhuwa. The results are presented in Table 8. Due to extra manual effort to project stopwords from the English language, we only focused on 50 top-ranking words. Here, the results show a clear advantage of network-based approaches over frequency-based ones. Also, in-degree outperforms the other approaches in Tetun and Emakhuwa. We visualized the top-25 stopwords for Tetun in Figure 2 and Emakhuwa in Figure 3. The network was filtered by in-degree and partitioned into stopwords (green) and not stopwords (red). Also, each node has a label that shows the original word and its translation, separated by a colon. The precision drop for the Emakhuwa language can be better visualized in Figure 3, which shows five misses. We believe that this drop in precision is because the Emakhuwa corpus is predominantly made up of religious texts, which contain a high number of words from religious themes, such as "Yesu" (meaning Jesus), "Muluku" / "Yehova" (meaning God), and others.

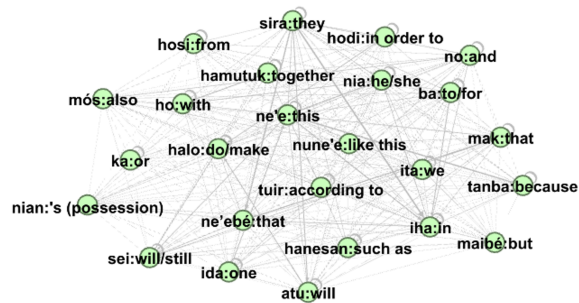


Figure 2: Top-25 of the Tetun stopwords identified using in-degree.

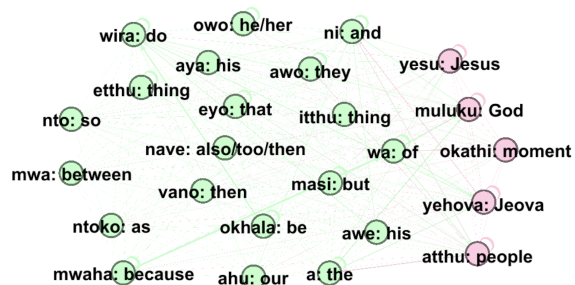


Figure 3: Top-25 stopwords of Emakhuwa identified using in-degree.

Approach	P@25	P@50	P@75	P@100	P@125	P@150	P@175	P@200
tf	0.960	0.800	0.706	0.600	0.528	0.480	0.451	0.420
ntf	0.960	0.800	0.706	0.600	0.528	0.480	0.451	0.420
df	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
idf	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
nidf	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
tf-idf	0.920	0.800	0.693	0.610	0.560	0.500	0.457	0.415
tdf	0.960	0.800	0.693	0.620	0.552	0.500	0.457	0.415
betweennesscentrality	1.000	0.800	0.693	0.620	0.552	0.487	0.429	0.390
indegree	1.000	0.880	0.800	0.690	0.608	0.54	0.469	0.415
outdegree	0.920	0.780	0.667	0.570	0.504	0.467	0.423	0.380
degree	1.000	0.840	0.800	0.700	0.608	0.520	0.460	0.411

Table 6: Stopword detection Precision for English.

Approach	P@25	P@50	P@75	P@100	P@125	P@150	P@175	P@200
tf	1.000	0.68	0.533	0.49	0.432	0.373	0.331	0.310
ntf	1.000	0.68	0.533	0.49	0.432	0.373	0.331	0.310
df	0.920	0.68	0.56	0.520	0.448	0.386	0.337	0.310
idf	0.920	0.68	0.56	0.520	0.448	0.386	0.337	0.310
nidf	0.920	0.68	0.56	0.520	0.448	0.386	0.337	0.310
tf-idf	0.920	0.68	0.546	0.520	0.448	0.373	0.337	0.315
tdf	0.960	0.680	0.546	0.510	0.440	0.373	0.331	0.310
betweennesscentrality	1.000	0.920	0.747	0.620	0.552	0.480	0.440	0.400
indegree	1.000	0.860	0.733	0.600	0.528	0.460	0.417	0.375
outdegree	0.960	0.800	0.707	0.590	0.512	0.447	0.411	0.370
degree	1.000	0.900	0.773	0.620	0.528	0.467	0.406	0.380

Table 7: Stopword detection Precision for Portuguese.

Approach	Tetun		Emakhuwa	
	P@25	P@50	P@25	P@50
tf	0.840	0.720	0.760	0.740
ntf	0.840	0.720	0.760	0.740
idf	0.880	0.760	0.800	0.760
nidf	0.640	0.500	0.800	0.760
tf-idf	0.880	0.760	0.800	0.760
tdf	0.880	0.740	0.760	0.760
betweennesscentrality	0.920	0.820	0.720	0.780
indegree	1.000	0.900	0.760	0.820
outdegree	0.960	0.840	0.760	0.800
degree	1.000	0.860	0.760	0.800

Table 8: Stopword detection Precision for Tetun and Emakhuwa

5 Related Works

The concept of stopwords was initially introduced by Luhn (1957) and as their application in the domains of information retrieval and natural language processing became evident, lists of stopwords were compiled for various languages. Several automated approaches for stopword detection have been proposed to streamline the process and eliminate manual effort. The conventional method for identifying stopwords uses term frequency (TF) (Manning et al., 2009; Croft et al., 2009). Lo et al. (2005)

introduced the normalized inverse document frequency (NIDF) in their experiment with TREC⁸. These two techniques, along with other unsupervised frequency-based approaches, such as inverse document frequency (IDF) and term frequency-inverse document frequency (TF-IDF), among others, have served as the standard mechanism for stopword detection. More recently, Ferilli (2021) proposed another approach called term-document frequency (TDF), which proves particularly effective when dealing with small-sized corpora.

From a linguistic perspective, stopwords are function words, which serve grammatical or structural roles in sentences. These function words are typically high-frequency terms such as articles, prepositions, conjunctions, and pronouns, frequently occurring in conjunction with other words. This high-frequency terms has been validated across different languages through various studies (Chen et al., 2018; Gao et al., 2014; Liang et al., 2009), employing co-occurrence networks constructed based on the linear relation of words. Gao et al. (2014) in their analysis of six languages (Arabic, Chinese, English, French, Russian, and Spanish), observed that function words, including

⁸<https://trec.nist.gov/>

“y,” “en,” and “a,” in the Spanish network, consistently ranked highest in degrees across all languages. Focusing on Chinese and English, Liang et al. (2009) reported that words with the highest connections (degree) were functional words, such as “a,” “the,” and “of,” in English networks. Similarly, Chen et al. (2018) identified stopwords with the highest degree in Chinese co-occurrence networks, suggesting their role as hubs and indicating high betweenness centrality.

While the aforementioned studies offer evidence of a potential correlation between network properties and “stopwordness”, to our knowledge, there has been no systematic evaluation of the advantages of complex network features in stopword detection. This is why this study investigates complex network properties for stopword detection.

6 Conclusion and Future Work

This paper presents an automated approach for stopword detection, leveraging network properties derived from connecting pairs of sequential terms within text corpora. The datasets underwent pre-processing before constructing pairs of sequential terms for selecting network features. Selecting these features involved evaluating the information gained from each one. The chosen network features were then employed in experimentation with the aforementioned four languages. Overall results indicate that network features are more effective than existing frequency-based approaches in stopword detection, with in-degree as the most reliable feature.

In future work, we aim to apply the proposed approach to construct stopword lists for Tetun and Emakhuwa. Furthermore, we will develop and use ground truth stopwords for both languages to conduct a more comprehensive evaluation of the effectiveness of the proposed approach.

Acknowledgements

This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC). Felermimo Ali is supported by a PhD studentship (with reference SFRH/BD/151435/2021), funded by Fundação para a Ciência e a Tecnologia (FCT). Similarly, Gabriel de Jesus also benefits from a scholarship

funded by FCT, identified by reference number SFRH/BD/151437/2021.

References

- Felermimo D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. [Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique](#). In *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*.
- Ricardo Baeza-Yates and Berthier A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.
- Marc Barthelemy. 2004. [Betweenness centrality in large complex networks](#). *The European Physical Journal B - Condensed Matter*, 38(2):163–168.
- Heng Chen, Xinying Chen, and Haitao Liu. 2018. [How does language change as a lexical network? an investigation based on written chinese word co-occurrence networks](#). *PLOS ONE*, 13(2):1–22.
- Kenneth Church and William Gale. 1995. [Inverse document frequency \(IDF\): A measure of deviations from Poisson](#). In *Third Workshop on Very Large Corpora*.
- W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines - Information Retrieval in Practice*. Pearson Education.
- Gabriel de Jesus. 2023. [Text information retrieval in Tetun](#). In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 429–435. Springer.
- Stefano Ferilli. 2021. [Automatic multilingual stopwords identification from very small corpora](#). *Electronics*, 10(17).
- Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The weka workbench. online appendix for "data mining: Practical machine learning tools and techniques". Fourth Edition.
- Yuyang Gao, Wei Liang, Yuming Shi, and Qiuling Huang. 2014. [Comparison of directed and weighted co-occurrence networks of six languages](#). *Physica A: Statistical Mechanics and its Applications*, 393:579–589.
- Wei Liang, Yuming Shi, Chi K. Tse, Jing Liu, Yanli Wang, and Xunqiang Cui. 2009. [Comparison of co-occurrence networks of the chinese and english languages](#). *Physica A: Statistical Mechanics and its Applications*, 388(23):4901–4909.
- Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. 2005. [Automatically building a stopword list for an information retrieval system](#). *J. Digit. Inf. Manag.*, 3(1):3–8.

- Hans Peter Luhn. 1957. [A statistical approach to mechanized encoding and searching of literary information](#). *IBM J. Res. Dev.*, 1(4):309–317.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Online edition, Cambridge UP.
- Stephen E. Robertson and Karen Spärck Jones. 1976. [Relevance weighting of search terms](#). *J. Am. Soc. Inf. Sci.*, 27(3):129–146.
- Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. [Cross-lingual learning-to-rank with shared representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463, New Orleans, Louisiana. Association for Computational Linguistics.
- Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval”.
- UMINHO University of Minho. 2021. [words-pt: Dicionário natura](#). Accessed: 2024-01-05.
- Matt Wiens. 2021. [english-words-py](#). Accessed: 2024-01-05.

Grammar Induction for Brazilian Indigenous Languages

Diego Pedro Gonçalves da Silva
Núcleo Interinstitucional
de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas
e de Computação
diegopedro@usp.br

Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional
de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas
e de Computação
taspardo@icmc.usp.br

Abstract

This paper investigates the issue of grammar induction for Brazilian indigenous languages, mainly focusing on unsupervised methods, but also testing a large language model for the task. Grammar induction poses several challenges, particularly when applied to low-resource languages, a characteristic commonly associated with indigenous languages. The primary objective of this paper is to discover syntactically related words in sentences. In addition to the contributions to linguistic studies, as in language description and structural analysis, grammar induction may help in varied Natural Language Processing tasks, as it could help detect parsing errors, enhance parsing results, and reveal pertinent relations for open information extraction purposes. The findings reveal that, even with a limited corpus, it is feasible to identify syntactically related words, especially for some relations. To the best of our knowledge, this represents a pioneering attempt to undertake grammar induction for Brazilian indigenous languages.

1 Introduction

In the year 2001, there were 6,981 languages spoken globally, some of which linguists predict will confront the threat of extinction by the year 2100 (Harrison, 2008). One of the reasons for this decline may be associated with political and social discrimination directed toward its speakers, thereby exerting an influence on subsequent generations. This influence may manifest as parents refraining from transmitting their native languages to their offspring, driven by concerns regarding perceived limitations in future opportunities (Harrison, 2008; Cruz, 2011). The consequences of a language extinction across social, political, and cultural spheres are profound and incalculable. The cumulative wisdom amassed across generations, transmitted exclusively through oral communication, irreversibly dissipates (Harrison, 2008).

In Brazil, according to data provided by *Instituto Brasileiro de Geografia e Estatística* (IBGE), there were 244 indigenous languages documented in the country in 2010 (Morello, 2016). Predominantly, these languages belong to the Tupi family, which comprises more than 40 distinct languages (Ferraz Gerardi et al., 2023). The expansive influence exerted by the Tupi language family constitutes the most extensive diffusion globally. This facilitates mutual comprehension among languages within this linguistic group, many of which share cognates (Ferraz Gerardi et al., 2023). Among the indigenous languages prevalent in Brazil, Ticuna, spoken by 46 thousand individuals, Guarani-Caiuí, with 43 thousand speakers, and Caingangue, with 37 thousand speakers, emerge as the most widely spoken ones according to IBGE (Morello, 2016). A considerable number of Brazilian indigenous languages are spoken by fewer than 100 individuals (Cruz, 2011).

Promoting literacy among indigenous children in their native language and attempting to digitalize their language constitutes strategic initiatives to mitigate language decline (Taylor, 1985; Azevedo, 2016). However, the rise of the internet may have hastened the extinction of indigenous languages, given that the prevalence of dominant languages significantly contributes to the functional loss of indigenous languages (Kornai, 2013). The content deficit of the indigenous languages adversely affects the development of technological tools for these languages, such as translation systems. These tools would be useful for disseminating information and facilitating learning, consequently, contributing to preserving the language.

Artificial Intelligence systems emerge as a significant initiative to contribute to the advance of language technologies (Pinhanez et al., 2023; de Lima et al., 2021). Addressing this challenge involves considering alternatives, such as the use of com-

parable texts to build parallel corpora¹, and the use of grammar induction for learning syntactical structuring patterns and lexical clustering for detecting semantically-related terms for a (probably low-resource) language of interest. Grammar induction is the focus of this paper.

In Natural Language Processing (NLP) applications, Grammar Induction (GI) proves useful for various tasks, including grammar checking, information extraction, and text simplification, to name a few. Grammar induction can be approached in an Unsupervised way (UGI), in a Semi-Supervised way (SSGI), or in a Supervised way (SGI). SGI methods demonstrated remarkable efficacy in many works, achieving accuracy rates exceeding 95% (Lin et al., 2022) for the English language, while their unsupervised counterparts present a considerable challenge, often falling short of this benchmark.

This study focuses on unsupervised approaches to induce grammar within the context of dependency paradigm, which seeks to model the dependency relations among syntactic elements. Illustrative instances are provided in the form of a Nheengatu sentence presented in Figure 1, along with its Portuguese translation portrayed in Figure 2. These sentences were extracted from the Nheengatu CompLin treebank (Avila, 2021) identified with ID *Avila2021:0:0:647*. The arrows delineate the relationships between two tokens, wherein the arrow originates from the head term and is directed toward the dependent term.

Good methods for grammar induction include Large Language Models (LLM) (Shen et al., 2021) and neural networks (He et al., 2018) and both methods need a huge amount of data for training. Due to the limited amount of available digital data in indigenous languages, we test two different approaches to discover related words in an unsupervised way: Dependency Model with Valence (DMV) (Klein and Manning., 2004), the most influential model in grammar induction tasks; and Mutual Information (MI), a measure that has demonstrated efficacy to retrieve syntactic structures (Futrell et al., 2019; Hoover et al., 2021). Furthermore, we also evaluate an LLM for the tasks.

The investigation specifically centers on twelve indigenous languages spoken in Brazil, most of which were annotated as a part of the TuLaR

¹It is not rare to use the Bible for such end, as it is published in many languages.

(Tupían Language Resources) project within the “Universal Dependencies” (UD) framework (Nivre et al., 2020). Notably, seven of these languages are affiliated with the Tupi family. To the best of our knowledge, this is the first unsupervised grammar induction study within the domain of Brazilian indigenous languages. We provide the code from this project at Github².

The next section brings a brief literature review on the topic of grammar induction. Section 3 presents the methods that we test, while Section 4 shows and discusses the achieved results. Discussion and final remarks are presented in Sections 5 and 6.

2 Related Work

In recent decades, Grammar Induction has been applied in different contexts and diverse applications. Varied methodologies have been employed, with the DMV (Klein and Manning., 2004) emerging as the most prevalent and widely recognized approach. This approach was the first to surpass the right-branching baseline, wherein the rightmost word functions as the head of the immediately adjacent left word, for grammatical structure induction.

Contemporary methods involve the utilization of neural networks (He et al., 2018) and LLM (Shen et al., 2021). Nevertheless, these innovative models may exhibit limitations when applied to languages with limited resources, particularly indigenous languages, and notably in the context of dependency grammar.

A noteworthy approach is the application of the MI measure, which has been harnessed to induce constituent grammar (Solan et al., 2005), and dependency relations for languages like Japanese (de Paiva Alves, 1996) and Portuguese (da Silva and Pardo, 2023).

Several initiatives have advanced in the domain of grammatical induction for languages with limited linguistic resources. Dahl et al. (2023) introduced a method employing Womb Grammars, a technique designed for the translational mapping of languages, in which grammar has been described to languages with no grammar description, to facilitate the induction of the Ch’ol language³.

²<https://github.com/diegodpgs/PROPORInd>

³Ch’ol is an indigenous language of Mexico that lacks a formally documented grammar. However, it is noteworthy that the grammatical induction methodology articulated in this study relies on the use of syntactic relations, by definition, using supervised training.

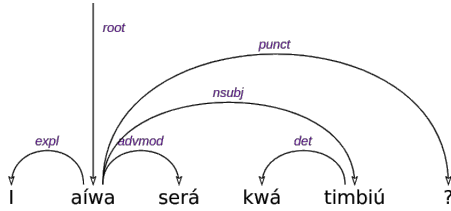


Figure 1: An example sentence for Nheengatu Language (Avila, 2021)

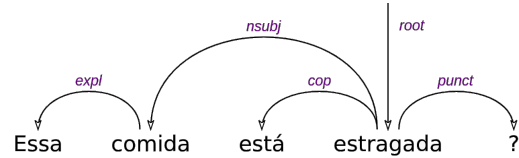


Figure 2: The translation to Portuguese of the sentence presented in Figure 1 (Avila, 2021)

In what follows we present the data and the methods that we explore in this paper.

3 Methodology

We use data from UD version 2.13⁴. This contains 245 treebanks (i.e., corpora with sentences and their corresponding syntactical dependency analyses) for 141 languages. Almost 50 languages are indigenous or ethnic representative. Of these, twelve are spoken in Brazil and nine in Russia. The twelve languages used in this work are: Akuntsu, Guajajara, Kaapor, Karo, Makurap, Munduruku, Tupinamba, Nheengatu, Apurina, Bororo, Xavante, and Madi.

All these languages include 36,322 tokens, 8,632 types, and 5,000 sentences. A detailed description of these languages is presented in Table 1. The first column describes the language used in the experiment, the second shows the linguistic family, and the third column describes the number of different Syntactic Relations (SR) used in the annotations. Subsequent columns detail the number of tokens, vocabulary size, and complexity (computed as the type-token ratio). Higher complexity indicates greater sparsity. The final three columns present the number of sentences, the average number of tokens per sentence, and the standard deviation for token counting.

Nheengatu may stand out as the most extensively documented Brazilian indigenous language, dating back to its description in the first Brazilian indigenous language dictionary in 1756 (Avila, 2021). Moreover, numerous texts in Nheengatu were authored during the eighteenth century, further contributing to its rich documentation. The Nheengatu treebank is the largest one: 12,743 tokens (35% of all treebanks) and 1,913 types (22.1% of all treebanks). About 99.8% of all sentences have a length of up to 40 tokens (including punctuation), which is compared to almost all European languages avail-

able in UD initiative. For instance, in German, Czech and Russian, which are the biggest treebanks in UD, about 93% of sentences have a length of fewer than 40 tokens.

Since the UD repository only provides test sets, we perform cross-validation such that the test set is split into five folds: one for test and four for training. Three different grammar induction methods are used: MI, DMV, and LLM. In the present study, it is pertinent to emphasize that our approach is entirely unsupervised. Therefore, our training data solely comprises raw text, with the exception of the DMV method which incorporates gold Part of Speech (POS) tags.

The first works on grammar induction applied a dynamic programming algorithm on $O(n^3)$ for constituency grammar (Sankaran, 2010; Cohen et al., 2008), which is computationally expensive for longer sentences. For this reason, most works on grammar induction were trained on sentences up to 40 tokens (Kim et al., 2019). In this paper, we tested the models on sentences of lengths up to 10 and up to 40 tokens, to evaluate the impact of different sentence size. The tree models used in this work are described in subsections 3.1, 3.2, and 3.3. These models are unsupervised, except for the LLM that, besides the zero-shot approach, we also used one and two-shot learning.

3.1 DMV Model

The DMV stands as a prevalent model for grammar induction, serving as a baseline in several works on unsupervised grammar induction (Shen et al., 2021; Yang et al., 2020). This model operates by generating syntactic trees in a top-down fashion using generative unsupervised training. The idea behind the DMV model is to estimate the syntactic tree by using the Expectation-Maximization (EM) algorithm. For each branch to be generated, it uses probability distributions to make decisions on when and which branch to generate.

We experimented with DMV using the same set-

⁴<http://hdl.handle.net/11234/1-5287>

Table 1: Indigenous languages in Brazil used in this study

<i>Language</i>	<i>Family</i>	<i>SR</i>	<i>Tokens</i>	<i>Types</i>	<i>Complexity</i>	<i>Sentences</i>	μ	σ
Xavante	Macro-Je	22	1,597	385	0.241	148	10.791	6.423
Tupinambá	Tupian	26	4,508	1,970	0.437	581	7.759	5.946
Nheengatu	Tupian	32	12,743	1,913	0.150	1,239	10.285	6.736
Munduruku	Tupian	26	1,022	399	0.390	158	6.468	5.977
Makurap	Tupian	15	178	95	0.533	37	4.811	1.998
Madi	Arawan	17	115	68	0.591	20	5.750	3.048
Karo	Tupian	25	2,319	773	0.333	674	3.441	1.523
Kaapor	Tupian	22	366	221	0.603	83	4.410	2.024
Guajajara	Tupian	27	9,160	1,515	0.165	1,182	7.750	4.041
Bororo	Bororoan	29	1,905	762	0.400	371	5.135	5.512
Apurina	Arawakan	26	941	373	0.396	152	6.191	3.258
Akuntsu	Tupian	21	1,468	506	0.344	343	4.280	2.556
All	-	35	36,322	8,632	4.208	5,000	7.264	5.450

- 1 Na sentença "Aiwana, paá, aintá uyaxiú", as relações de dependência sintática são mostradas abaixo no formato (token dependente -> token cabeça)
- 2 (Aiwana -> uyaxiú)
- 3 (, -> paá)
- 4 (, -> paá)
- 5 (aintá -> uyaxiú)
- 6 (. -> -> uyaxiú)
- 7 Liste as relações de dependência sintática na sentença "Yané tuixawa umaná ana mira amusuaxarawara usikié tenhê waá.", usando o formato (token dependente -> token cabeça).

Figure 3: An example of prompt for the Nheengatu language in one shot learning

ting provided by He et al. (2018). It is pertinent to note that this model exhibits limitations in training with longer sentences, attributed to the $O(n^3)$ time complexity of the EM algorithm (Cohen et al., 2008; Spitkovsky et al., 2010). However, given the relatively small treebanks employed in this investigation, the DMV is executed with 10 epochs on each fold using cross-validation assessments.

3.2 MI-based Model

Generally defining, the MI measure indicates the dependency among elements of interest. In our case, it is used to determine words that are more probable to be syntactically related. Equation (1) shows how it is computed for head (h) words and their dependents (d).

$$MI(D, H) = \sum_{d \in D} \sum_{h \in H} P(d, h) \log_2 \frac{P(d, h)}{P(d)P(h)} \quad (1)$$

To compute it, we performed word pair permutations within each sentence, considering every possible configuration. The total number of permuted pairs is described by $\sum_{d=1}^{DW} n - d$,

where n is the number of tokens in the sentence, including punctuation, and DW is the distance between the words in the sentence. For instance, for the sentence "I love the sun", the word pairs for $DW=1$ is <I, love>, <love, the>, <the sun>. Using $DW=n$, the number of pairs is described by binomial coefficients $\binom{n}{k}$, with k representing two (tokens per pair). This setting produces the pairs <I,love>, <I,the>, <I,sun>, <love,the>, <love,sun> and <the,sun>. We train all models using different DW values and choose $DW=2$ as the best performance.

That permutation process resulted in the creation of the final set of Sentence Permutations (SP), comprising pairs of tokens where the first token precedes the second in the sentence sequence. Following this, MI was computed for each word pair within the SP. Finally, we take the n pairs with the highest MI and compare them to manually annotated sentences.

Since corpora used in this work are very small, we perform an edit distance smoothing. For each token in the test that was not in the training set, we searched for the most similar morphological token in the training set using edit distance. For instance,

if the token “*uyapí*” does not appear in the training set, the edit distance is applied to find the most lexically related word in the training set, such as “*uyari*”. Then the frequency of the token “*uyari*” is assigned to the token “*uyapí*”. Since there will always be a lexically related token, all tokens in the test set will have a frequency. For bigrams found in the test and not in the training set, we apply a derived simple Laplace smoothing by attributing frequency equal to $1/\text{size of the vocabulary}$.

3.3 Large Language Model

LLM are models that are trained with a massive amount of data and require a huge computational structure. They can be used in a wide number of tasks such as information extraction, summarization, and question answering, to name a few (Wei et al., 2022). We did not build the LLM using native languages, instead, since we do not have enough data, we used LLM trained in Portuguese. Since the native languages used in this work are spoken in Brazil, and their vocabularies eventually incorporate some Portuguese words, we believe that is possible to find some syntactic relations using LLM even if that language has never been used for training.

We aim to demonstrate the limits and potentialities of LLMs to learn syntactic information in languages with lower resources. We use the chatGPT 3.5 API provided by OpenAI. Differently from the experiments on MI and DMV, we select only three languages to conduct experiments with the LLM. As we wanted to analyze the influence of a larger treebank, we tested with Nheengatu. Average sentence length can also play a role in dependency grammar induction and, therefore, we chose the Karo language, whose sentences are shorter. Finally, we wanted to study the influence of the language family, and language Bororo was chosen for having the largest treebank among those languages not belonging to the Tupian family.

We performed zero, one shot, and two shots learning. In +1 shot learning, we use two different prompts: using a fixed sentence and a random sentence for composing the prompt. For the fixed sentence, we chose a sentence of length seven, which is approximately the average of all languages used in this study. The chosen sentence is the one with the most frequent tokens in the treebank. For the prompt that applies a random sentence, we have random sentences with lengths up to 40 tokens in

the training set to be included in the prompt. Since the answers provided by the model are not always the same, we tested the prompts on 30 sentences for each of the five folds of cross-validation. This experiment resulted in 2,250 requests to OpenAI API. We also tested different prompts in Portuguese language and chose the best one. An example of a prompt for one shot is shown in Figure 3.

4 Results

In this study, we adopt the 37 syntactic relations of the UD initiative⁵, yet not all languages that we examined utilize all of these relations. As demonstrated in Table 1, Makurap employs only 15 syntactic relations, while Nheengatu utilizes 32. It is noteworthy that Guajajara does not include any occurrence of the subject relation. This study concentrates exclusively on syntactic relations that constitute a minimum of 10% of the respective treebank annotations. Due to limited data, we did not consider the subtypes of some syntactic relations.

We present results for the standard evaluation metrics: Undirected Dependency Accuracy (UDA) and Directed Dependency Accuracy (DDA). Comparing with the reference annotations, these metrics compute how many relations (for word pairs) were correctly predicted, considering or not the relation direction, respectively.

Overall, it is interesting that, despite the limited size of the treebanks, the induction methods for these languages achieved good results, even better than some reported results for non-indigenous languages, such as German, English, and Chinese, using DMV (Klein and Manning., 2004).

In general, Akuntsu and Karo emerged as languages exhibiting the best outcomes, whereas Guajajara and Xavante posed notable challenges. These results are not related to the family origin or annotation. Akuntsu, Karo, and Guajajara were annotated using the same annotation protocol within the same project (Gerardi et al., 2021). However, Akuntsu and Karo are two languages spoken in the state of Rondônia, but Guajajara and Kaapor, which are also spoken in the same state (Maranhão) and come from the same family, Tupian, present different outcomes.

No discernible correlation is observed between vocabulary size and treebank size; however, a subtle correlation is discerned between sentence length

⁵<https://universaldependencies.org/u/dep/index.html>

and associated scores. Across all settings, the “object” dependency relation was the most correctly detected one, yet substantial variation exists among languages.

MI presented the best results on UDA; on the other hand, DMV was better on DDA. As may be expected, LLM presented the worst results.

The syntactic relations that were more correctly induced (with the highest scores) with DMV are *punct* (punctuation) with 20.8%, *obj* (object) with 18.7%, and *nsubj* (subject) with 16.7%. However, MI presents the highest incidence of *obj* with 26% and *nsubj* with 18%, followed by *advmod* (adverbial modifier) with 8%. The selection of these syntactic relations is based on their prevalence within the treebank. Nonetheless, our code is accessible for retrieving data related to other syntactic relations as well.

The detailed results are presented in Subsections 4.1, 4.2, and 4.3. The summarized results are presented in Table 2. The last three lines present the most correctly induced syntactic relations (1 SR), the second most correctly induced syntactic relations (2 SR), and the third most correctly induced syntactic relations (3 SR), respectively. Due to space limitation, we presented only the results for DMV using the DDA metric⁶.

Table 2: Summarized results

	DMV	MI	LLM
UDA 10	0.5135	0.5692	0.4165
UDA 40	0.4654	0.5089	0.4212
DDA 10	0.3201	0.3122	0.2779
DDA 40	0.2808	0.1687	0.2720
1 SR	obj	obj	obj
2 SR	nsubj	nsubj	case
3 SR	punct	advmod	advmod

4.1 DMV

The results for DDA are presented in Table 3. DMV can induce correctly 89% of all object relations on Akuntsu, but only 11% on Kaapor. Despite presenting good results on small corpora such as those of Makurap and Madi, DMV struggles to induce some important syntactic relations. This pattern is similar when evaluated using UDA metrics.

⁶Detailed results may be found at <https://github.com/diegodpgs/PROPORInd>

4.2 MI

The use of edit distance yielded notable improvements, showcasing a 29.5% enhancement in MI for UDA and a 13.6% boost for DDA. While the results based on MI lag behind DMV in terms of DDA metrics, it is crucial to highlight the superiority of MI in UDA metrics. Moreover, it manifests superior outcomes in the context of induced object and subject relations. Notably, in the Makurap language, all object relations were accurately induced, and, in the Madi language, every subject was correctly induced.

4.3 LLM

Differently from experiments with DMV and MI, we did not use weighted average for LLM because the Nheengatu language presents 75% of the available corpora. The results presented in Table 2 refer to the average of all settings. As we expected, the zero-shot for all languages and all settings yielded the least favorable results on average, with 0.290 for UDA and 0.142 for DDA; transitioning to one-shot learning, UDA improved to 0.413, and DDA to 0.264; in two-shot learning, the model achieved 0.427 for UDA and 0.285 for DDA. When sentences were not fixed, the model exhibited competence with scores of 0.431 for UDA and 0.286 for DDA. However, when fixed sentences were employed in the prompt for one and two-shot learning, the overall performance deteriorated, resulting in an average of 0.406 for UDA and 0.263 for DDA. This result may be due to the distribution of the sentences, since that, with no fixed sentence, almost 150 different sentences were tested in the prompt. However, to induce object relations, using a fixed sentence in the prompt presented better results.

Different from MI and DMV, LLM may be influenced by the size of the treebank. When using one and two-shot learning, Nheengatu presents 0.440 DDA, against 0.406 in Karo and 0.410 in Bororo. This result is different from the DMV and MI approaches, in which Nheengatu presents the poorest scores. Nonetheless, the induction of particular dependency relations may not necessarily exhibit a correlation with treebank size. In the cases of Karo and Bororo languages, accurate induction of object relations is achieved with notable proficiency. In contrast, the Nheengatu language demonstrates a lower level of accuracy in this regard. These outcomes align with the findings obtained through both DMV and MI approaches.

Table 3: Results for DMV with DDA metric

DDA for sentences ≤ 10 tokens							
Language		1 SR		2 SR		3 SR	
Akuntsu	0.5661	0.8957	obj	0.5783	nsubj	0.5551	punct
Apurina	0.4248	0.7460	obj	0.7227	nsubj	0.2321	punct
Bororo	0.3832	0.8696	case	0.6992	obl	0.5489	nsubj
Guajajara	0.1669	0.4690	obl	0.2142	discourse	0.0730	punct
Kaapor	0.2500	0.7843	obj	0.4921	nsubj	0.1765	advmod
Karo	0.3803	0.5882	nsubj	0.4595	advmod		
Madi	0.4186	0.4545	punct	0.2500	obj		
Makurap	0.4696	0.6667	advmod	0.3750	discourse		
Munduruku	0.4074	0.8077	case	0.6846	obl	0.5000	punct
Nheengatu	0.3671	0.5756	advmod	0.5579	nsubj	0.2271	punct
Tupinamba	0.3138	0.5111	punct	0.4100	obl		
Xavante	0.3264	0.7500	dep	0.3099	punct	0.1176	nsubj
μ	0.3729	0.6765		0.4795		0.3038	
μ weighted	0.3201	0.5907		0.4492		0.2193	
DDA for sentences ≤ 40 tokens							
Akuntsu	0.5641	0.8800	obj	0.6077	nsubj	0.5879	punct
Apurina	0.3907	0.8488	obj	0.7211	nsubj	0.2153	punct
Bororo	0.3579	0.6647	punct	0.6497	obl	0.5020	nsubj
Guajajara	0.1704	0.4223	obl	0.2135	discourse	0.0900	punct
Kaapor	0.2287	0.8302	obj	0.4242	nsubj	0.2432	advmod
Karo	0.3301	0.5882	nsubj	0.4757	advmod		
Madi	0.3585	0.4167	punct				
Makurap	0.4348	0.6250	advmod	0.4375	discourse		
Munduruku	0.3784	0.9029	case	0.6506	nsubj	0.5909	obl
Nheengatu	0.2943	0.5376	nsubj	0.4918	advmod	0.1613	punct
Tupinamba	0.2572	0.4835	punct	0.3921	obl		
Xavante	0.3110	0.6348	dep	0.2800	punct		
μ	0.3397	0.6529		0.4858		0.3415	
μ weighted	0.2808	0.5512		0.4198		0.1916	

5 Discussion

Despite the effectiveness of modern approaches such as neural networks and LLM, simple methods such as MI can perform better when applied to low language resources. For some sentences, we identified that the LLM likely employed the straight-forward right-branching algorithm. It is necessary to note that an explicit evaluation of the comparative efficacy of these methodologies against the right-branching baseline, established at 0.38 for the English language (Klein and Manning, 2004), was not conducted and remains for future work.

The MI models present good results, but the induced syntactic tree could have missing elements, as presented in Appendix A. It can be solved by optimization, which could also be a matter of future work.

It is essential to highlight that the indigenous languages utilized in this study exhibit distinct syntactic characteristics, including the absence of certain crucial syntactic relations (such as nsubj in Guajajara, for example), as well as unique sentence structures. These nuances may influence the obtained outcomes. In-depth linguistic inquiries or even anthropological investigations may be necessary to elucidate the variations in results across different languages.

6 Final Remarks

We presented a study on grammar induction for different Brazilian indigenous languages. We demonstrate the efficacy of inducing syntactically related words for low-resource languages using some well-known approaches and a current LLM-based strat-

egy, mainly in inducing specific relations, such as object and subject relations. Such methods may be very useful to uncover syntactic structures for languages for which the grammar was not yet described or to refine NLP parsing methods.

Future work includes the investigation of other induction methods and the exploitation of language-specific features that may improve the results.

The interested reader may find other details about this and other related work at the web portal of the POeTiSA project (*P*ortuguese processing - *T*owards *S*yntactic *A*nalysis and *P*arsing)⁷.

Acknowledgments

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. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

References

- Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Ph.D. thesis, Universidade de São Paulo.
- Marta Maria Azevedo. 2016. *Urbanização e migração na cidade de São Gabriel da Cachoeira, Amazonas*. In *Anais do XV Encontro Nacional de Estudos Populacionais*, pages 1–14.
- Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. 2008. *Logistic normal priors for unsupervised probabilistic grammar induction*. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pages 321–328.
- Aline Da Cruz. 2011. *Fonologia e gramática do nheengatú: A língua geral falada pelos povos baré, warekena e baniwa*. Ph.D. Thesis, Vrije Universiteit Amsterdam.
- Diego Pedro Gonçalves da Silva and Thiago Alexandre Salgueiro Pardo. 2023. *Indução gramatical para o português: a contribuição da informação mútua para descoberta de relações de dependência*. In *Proceedings of the 14th Brazilian Symposium on Information Technology and Human Language*, pages 298–307.
- Veronica Dahl, Gemma Bel-Enguix, Velina Tirado, and Emilio Miralles. 2023. *Grammar induction for under-resourced languages: the case of ch’ol*. In *Proceedings of the Analysis, Verification and Transformation for Declarative Programming and Intelligent Systems: Essays Dedicated to Manuel Hermenegildo on the Occasion of His 60th Birthday*, pages 113–132.
- Tiago Barbosa de Lima, André C. A. Nascimento, Pericles Miranda, and Rafael Ferreira Mello. 2021. *Analysis of a Brazilian indigenous corpus using machine learning methods*. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 118–129.
- Eduardo de Paiva Alves. 1996. *The selection of the most probable dependency structure in Japanese using mutual information*. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 372–374.
- Fabício Ferraz Gerardi, Tiago Tresoldi, Carolina Coelho Aragon, Stanislav Reichert, Jonas Gregorio de Souza, and Francisco Silva Noelli. 2023. *Lexical phylogenetics of the Tupí-Guaraní family: Language, archaeology, and the problem of chronology*. *Plos one*, 18(6):1–25.
- Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. *Syntactic dependencies correspond to word pairs with high mutual information*. In *Proceedings of the 5th international conference on dependency linguistics*, pages 3–13.
- Fabício Ferraz Gerardi, Stanislav Reichert, and Carolina Coelho Aragon. 2021. *Tuled (Tupían lexical database): introducing a database of a South American language family*. *Language Resources and Evaluation*, 55(4):997–1015.
- K. David Harrison. 2008. *When languages die: The extinction of the world’s languages and the erosion of human knowledge*. *Oxford University Press*.
- Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. *Unsupervised learning of syntactic structure with invertible neural projections*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.
- Jacob Louis Hoover, Wenyu Du, Alessandro Sordani, and Timothy J. O’Donnell. 2021. *Linguistic dependencies and statistical dependence*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963.
- Yoon Kim, Chris Dyer, and Alexander M. Rush. 2019. *Compound probabilistic context-free grammars for grammar induction*. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2369–2385. Association for Computational Linguistics.
- Dan Klein and Christopher D. Manning. 2004. *Corpus-based induction of syntactic structure: Models of dependency and constituency*. In *Proceedings of the*

⁷<https://sites.google.com/icmc.usp.br/poetisa>

42nd Annual Meeting of the Association for Computational Linguistics, page 478–485.

András Kornai. 2013. [Digital language death](#). *PloS one*, 8(10):1–11.

Boda Lin, Zijun Yao, Jiaxin Shi, Shulin Cao, Binghao Tang, Si Li, Yong Luo, Juanzi Li, and Lei Hou. 2022. [Dependency parsing via sequence generation](#). In *Findings of the Association for Computational Linguistics*, pages 7339–7353.

Sidney Facundes Moore, Denny and Nádia Pires. 1994. [Nheengatu \(língua geral amazônica\), its history, and the effects of language contact](#). In *Proceedings of the Meeting of SSILA and the Hoka-Penutian Workshop*, pages 93–118.

Rosângela Morello. 2016. [Censos nacionais e perspectivas políticas para as línguas brasileiras](#). *Revista Brasileira de Estudos de População*, 33(2):431–439.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4033.

Claudio S Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. [Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6174–6182.

Baskaran Sankaran. 2010. [A survey of unsupervised grammar induction](#). *Manuscript, Simon Fraser University 47*, pages 1–63.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron C. Courville. 2021. [Structformer: Joint unsupervised induction of dependency and constituency structure from masked language modeling](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 7196–7209.

Zach Solan, David Horn, Eytan Ruppín, and Shimon Edelman. 2005. [Unsupervised learning of natural languages](#). In *Proceedings of the National Academy of Sciences*, pages 11629–11634.

Valentin I. Spitzkovsky, Hiyán Alshawi, and Daniel Jurafsky. 2010. [From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing](#). In *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 751–759.

Gerald Taylor. 1985. [Apontamentos sobre o nheengatu falado no rio negro, brasil](#). *Amérindia: revue d'ethnolinguistique amérindienne*, pages 5–23.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*, pages 1–30.

Songlin Yang, Yong Jiang, Wenjuan Han, and Kewei Tu. 2020. [Second-order unsupervised neural dependency parsing](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3911–3924.

A Illustration of grammar induction for Nheengatu

We present a sample of the induced relations for the sentence *Aikwé awá ururi indé u reyuri putari tē ne rupí?*, which corresponds to *Was there anybody to bring you or did you yourself want to come?* in English, using DMV, MI, and LLM methods. The cited sentence represents a transcription of speech delivered by an indigenous Nheengatu speaker (Moore and Pires, 1994). It is important to note that the orthography utilized is not the original form, but has been adjusted to adhere to the UD framework.

In Figures 4, 5, and 6, the color orange means that the model correctly predicted the relation according the UDA measure (which does not evaluate the direction of the arrow), and green means that the model correctly predicted the direction too, as informed by the reference annotation (in Figure 7).



Figure 4: Induced relations using DMV

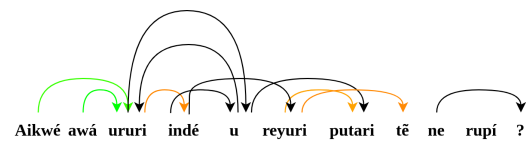


Figure 5: Induced relations using MI



Figure 6: Induced relations using LLM



Figure 7: Reference annotation in the treebank

NLP Tools for African Languages: Overview

Joaquim Mussandi

Lisbon University
Instituto Superior Técnico
Luanda University

Instituto de Tecnologias de Informação e Comunicação
joaquim.mussandi@tecnico.ulisboa.pt

Andreas Wichert

Lisbon University
Instituto Superior Técnico
andreas.wichert@tecnico.ulisboa.pt

Abstract

In Natural Language Processing (NLP), languages are classified into low-resource languages and high-resource languages, therefore, rich. African languages are grouped among those with few resources, due to little investment and consequently little interest from researchers. To understand this phenomenon, and without wanting to enter into a discussion in the historical-anthropological, sociological forum, or other areas, we carried out a study to identify the languages that have PLN resources, report the points of greatest consensus among experts, identify researchers' motivations, possible opportunities, challenges, possible linguistic patterns, institutions, events and projects that stimulate them, the most used techniques for constructing datasets in vernaculars and the different forms of corpus training. From what we have seen, it seems to us to be a promising field of study.

1 Introduction

Languages are the main communication mechanisms. In this era of digitalization, there is a need for African languages to keep up with this dynamic. Technological linguistic resources are essential for the development of the economic, political, financial, educational, medical and tourist sectors, etc. as they constitute the basis for the development of more advanced research in Artificial Intelligence. Automatic text translation, voice search, sentiment analysis, data analysis and event prediction (Siminyu et al., 2021, 2020) are some of the domains that require contextualized linguistic corpus. However, there is a shortage of funding, documentation and human resources to overcome the challenges in building technological resources in Natural Language Processing (NLP) in languages of African origin (Siminyu et al., 2021; Adda et al., 2016; Ayogu and Abu, 2021), as well as mitigating the possibility of extinction of some languages that

are under threat (Sands, 2018). At the same time, there is a need to reduce the difference between the richest languages and those with few resources (Adda et al., 2016).

The proliferation of the Internet in urban and suburban areas of Africa could mean an increase in data in digital format in these languages, which consequently increases their visibility within and outside the continent with the possibility of collecting data through Web tracking (Kandybowicz and Torrence, 2017). The corpus annotated at the token level, identification of orthographic patterns, grammatical classification of texts at the morphological and morphosyntactic level in vernaculars are differentials for the development of other research in this domain (du Toit and Puttkammer, 2021). Like the challenges of languages with greater resources, for corpora with text-to-speech, actors must be found to record such extracts through simulated scenes (Siminyu et al., 2021) or social communication professionals, teachers, judges, etc. as long as they speak the respective languages.

In this overview we surveyed the languages of African origin that have PLN tools, the institutions that motivate the holding of events and projects to build technological tools in languages of African origin, opportunities and challenges, particularities in corpus construction in these languages. This field of study is promising.

2 Initial considerations: definition of research questions

There are, in the literature, several studies carried out on African languages. Some study them in the linguistic aspect (Matsinhe and Fernando, 2008), others study them in the political aspect (Kanana Erastus and Erastus, 2013) combining linguistic history to identify common ancestors through language (phylogenetic) (Schryver et al., 2015), territories, ethnicities, culture of the peo-

ple (Pinto and Silva, 2022; Ki-Zerbo, 2010), in an attempt to revitalize languages threatened with extinction (Sands, 2018). NLP specialists motivated to provide linguistic resources for the technology industry are also invited (Kanana Erastus and Erastus, 2013; Loubser and Puttkammer, 2020a; Niek-erk et al., 2017; Siminyu et al., 2020).

1. Technologically, is there research carried out on African languages?
2. But, why study African languages from a technological perspective?
3. What NLP tools exist in African languages?

These questions are answered in this article as we read it further.

2.1 African languages as a factor in economic and socio-cultural development

The economic sector is one of the most affected by the lack of technological resources in African languages. In Africa, generally, the languages of the former colonial powers are spoken in large urban centers and local languages are spoken mainly by young people and adults in the interior regions. Classes in national systems are mostly taught in the languages of colonial powers, sometimes providing children with illustrations and examples outside of everyday games, increasing the degree of difficulty in learning. Especially because knowledge of linguistic history helps in the search for the necessary bases for inferences about the cultural history of its speakers (Ki-Zerbo, 2010).

Siminyu et al. state that with an investment in NLP, governments would be the biggest beneficiaries, as they would have factual data for making decisions about investment within the scope of public policies, the analysis of feelings about the needs of local customers, to direct private investors, for example, example (Siminyu et al., 2021). Unfortunately, some governments restrict national languages for use in certain domains considered informal, such as intra-community communication, interpretive roles in local courts, use by politicians at rallies, etc. For Kanana, Tanzania, Ethiopia and Egypt, with Swahili, Amharic and Arabic languages, respectively, and most Arabic-speaking countries, are references in the development of native languages that today serve as national languages used for education, business and commerce (Kanana Erastus and Erastus, 2013).

3 Background: work carried out

NLP resources, in particular, can be a means of accelerating the process of digital inclusion. This vision is also shared by Pauw and Schryver who direct their research in seven African languages of Bantu origin, namely Ciluba (Republic of Congo), Gikuyu, Kikamba (Kenya), Sotho and/or Soto and Venda (South Africa), the Nilotic Maa (Kenya) and the Defoid Yoruba (Nigeria) whose spelling is similar (diacritics), having developed applications and technological components under the approach of data-driven methods (Pauw and Gilles-Maurice de Schryver, 2009). Jakobus and Puttkammer have developed some language technologies for four official South African languages, namely Ndebele, Swati, Xhosa and Zulu. These technologies are summarized as lemmatizer, part-of-speech tagger, morphological analyzer for each of the languages. This was possible after building a corpus in each of these languages (du Toit and Puttkammer, 2021). In another research, an implementation of artificial neural networks was used in word embedding to build language technologies in which the data was modeled sequentially to perform all part-of-speech tagging tasks (POS-tagging, grammatical marking of the text), lemmatization, Named Entity Recognition¹ (NER), compound analysis. It is a language model that, for translation, given a sequence of input words, predicts the output, with different lexicons and lengths. This approach was based on ten South African languages², having presented good results, the best in the state of art at the height (Loubser and Puttkammer, 2020a).

Two other studies carried out mainly, but not exclusively, in the former territory of the Kongo kingdom, on the Kikongo with a focus on phylogenetics up to half a century. The first done by Gilles-Maurice et. al. directed from two perspectives: the first to present the character-based phylogenetic classification applied to lexical data, which became known as Kikongo Language Clouster (KLC) and, second, to present an exhaustive overview of the field of lexico-statistics of Bantu languages about KLC (Schryver et al., 2015). The second study carried out by Bostoen to examine variation in the expression of tense and aspect (TA) in a universe of 23 varieties of modern and two historical Bantu

¹NER is a technique used in NLP to categorize and identify key information in a text.

²The languages studied are: Ndebele, Afrikaans, Xhosa, Zulu, Swati, Sepedi, Sotho, Swana, Tsonga and Venda.

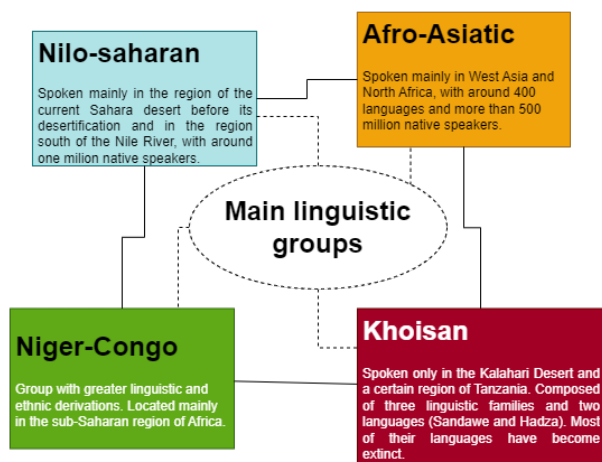


Figure 1: Main African linguistic groups

languages, including Kikongo (Dom and Bostoen, 2015).

Yamaguchi and Tanaka-Ishii include two national languages of Angola in their research. These are Umbundu and Kimbundu (Hiroshi Yamaguchi and Tanaka-Ishii, 2012). Outside the domain of NLP, Kikongo is also a regional language (spoken in Angola, DRC, Congo and Gabon) and has been studied by academics for its valorization, for example, in understanding the complexity of its morphology for preliminary exploration of the ordering of affixes verbal (Pinto and Silva, 2022).

Voice searches, automatic text translation, speech-to-text conversion are tasks that receive greater attention from researchers, however, for languages with greater economic power. For African languages, Siminyu et al. performed data collection for open source corpus annotations for varieties of NLP tasks and creating baselines for machine learning tasks. And they state that NLP contributes to the promotion, appreciation and dissemination of linguistic diversity (multilingual), as well as inclusion (Siminyu et al., 2021).

3.1 Understanding African languages: origins

There are over 7 thousand languages spoken in the world. The African continent, like the others, is multilingual. UNESCO recognizes around 2,092 languages and dialects spoken in Africa, in addition to creoles, a mixture of the languages of former settlers and local native languages (Ki-Zerbo, 2010). These languages have various origins, briefly stratified into four main groups, illustrated in Figure 1.

Sub-Saharan Africa, as a result of what hap-

pened at the Berlin conference in 1985, became the most plurilingual, pluricultural and pluriethnic part of the continent (Ki-Zerbo, 2010; Pinto and Silva, 2022). Cameroon alone shares around 70 languages with neighboring countries. One of these countries is Nigeria, with which it shares around 45 languages (Kanana Erastus and Erastus, 2013). Zimbabwe and South Africa are the countries in Africa with the most official languages, 16 in the first and 11 in the second with legal support in their constitution, in addition to another 24 languages (Loubser and Puttkammer, 2020b).

3.2 Main Linguistic Events, Projects and Institutions

Since 1953, UNESCO has debated the use of local languages for African education and to rescue sociocultural values prior to the colonial period, the specific educational need. This strategy would be characterized as effective, firstly, by the use of an appropriate teaching medium, contextualized teaching techniques, culturally appropriate curricular content and sufficient financial and material resources. In addition to supporting various initiatives such as *AI4D*, *International Conference Language Technologies for All (LT4All)* and others (Siminyu et al., 2021). UNESCO celebrates the day of reflection on the languages of minority groups on February 21st. And, from 2022 to 2032, it coordinates the implementation of the international decade of indigenous languages declared by the United Nations in 2019, known as *Los Pinos Declaration on the Decade of Indigenous Languages* (Siminyu et al., 2021).

In 2020, it held the competition called *AI4D-Africa Language Dataset Challenge*³ with the objective of creating and curating a dataset on African languages spoken in South Africa, Ghana and Uganda, with quality data to be used in future studies for language models (Siminyu et al., 2021). *AI4D* has already held seven events, including six competitions and a *hackathon* in African languages, namely:

- In November 2020, *GIZ NLP Agricultural Keyword Spotter for Luganda*;
- In February 2021, *AI4D Swahili News Classification Challenge*, participants were allowed from Tanzania, Kenya, Malawi, Uganda and Rwanda;

³Details here: <https://zindi.africa/competitions/ai4d-african-language-dataset-challenge>.

Language	Mains tools	Sources
Afrikaans	Corpus and Morphological analyser	(Eiselen and Puttkammer, 2014)
Ciluba	Corpus and Morphological analyser	(Pauw and Gilles-Maurice de Schryver, 2009)
Chichewa	Morphological analyser	(Keet, 2016)
Swahili	Machine Translation, Morphological analyser	(Keet, 2016)
Zulu	Machine Translation	(Eiselen and Puttkammer, 2014; Keet, 2016)
Shona (Xhosa)	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Setswana	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Ndebele	Morphological analyser, Language Model	(Eiselen and Puttkammer, 2014; Keet, 2016)
Sepedi	Morphological analyser	(Eiselen and Puttkammer, 2014)
Yuruba	Machine Translation	(Siminyu et al., 2021; Ayogu and Abu, 2021)
Swati	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Venda	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Sotho	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Tsonga	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Luganda	Agricultural Keyword Spotter	(Siminyu et al., 2021)
Chichewa	News Classification	(Siminyu et al., 2021)
Fongbe	Machine Translation	(Siminyu et al., 2021)
Ewe	Machine Translation	(Siminyu et al., 2021)
Tunisian Arabizi	Social Media Sentiment Analysis	(Siminyu et al., 2021)
Wolof	Automatic Speech Recognition	(Siminyu et al., 2021)
Kinyrwanza	CHAT-GPT for Covid-19,	(Forum, 2022)
Kikongo	Clouster	(Schryver et al., 2015)
Umdundu	Corpus	(Hiroshi Yamaguchi and Tanaka-Ishii, 2012)
Kimbundu	Corpus	(Hiroshi Yamaguchi and Tanaka-Ishii, 2012)
Basaa	Machine translation	(Adda et al., 2016)
Myene	Machine translation	(Adda et al., 2016)
Embosi	Machine translation	(Adda et al., 2016)
Emakhuwa	Corpus	(Ali et al., 2021)
Lingala	Corpus	(Sene-Mongaba, 2015)

Table 1: Main NLP toolkits for African languages

- As of March 2021, *AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi*;
- In May 2021, *AI4D Yoruba Machine Translation Challenge*;
- In May 2021, *AI4D Takwimu Lab - Machine Translation Challenge: French to Fongbe and Ewe*;
- In May 2021, *AI4D Chichewa News Classification Challenge*;
- Coming May 2021, *AI4D Baamtu Datamation - Automatic Speech Recognition in WOLOF*.

AI4D - Artificial Intelligence for Development is an initiative that involves entities from the academic, business and governmental community whose vision is to leverage AI through high-quality research,

responsible innovation and strengthening talent, as well as enabling better political decisions by governments to its people, integrating African languages into digital platforms.

In 2016, Adda et al. carried out a project that became known as **BULB**⁴ (*Breaking the Unwritten Language Barrier*), which brought together linguists and computer scientists, with the aim of supporting the documentation of languages with few resources. To achieve this, tools adapted to the documentary needs of linguists were developed, taking advantage of NLP technology and knowledge, especially in *automatic speech recognition* and *machine translation*. Using three African languages from the Bantu family, from the **Nigerian-Congolese** linguistic group, namely: Basaa, Myene and Embosi. The project was divided into three

⁴Available at:<https://www.bulb-project.org/>

main stages: in the first stage, data was collected from a large audio corpus (100h per language) at a reasonable cost. To achieve this, standard mobile devices and dedicated software were used – Lig-Aikuma. The recorded data was repeated by a reference speaker to improve signal quality and translated orally into French. In the second stage, automatic transcription of the Bantu languages was carried out at the phoneme level and the French translation at the word level. Recognized Bantu phonemes and French words were automatically aligned. In the third and final stage, tools were developed in close cooperation and discussion between linguists, speech and language technologists, to support linguists in their work, taking into account their needs and technological capabilities (Adda et al., 2016).

Totoeba is a project that involves around 500 languages, with emphasis on those with few resources, whose objective is to create automatic translation tools and models covering the widest languages in the world, in an open way. It has a comprehensive collection of diverse datasets in hundreds of languages with systematic language and *script* annotation. Provides a growing number of pre-trained baseline models for individual language pairs and selected language groups. In line with the main objective is the purpose of encouraging people to develop machine translation in real-world cases into multiple languages (Tiedemann, 2020). With around 500 GB of compressed data for 2,961 language pairs, covering 555 languages. Unlike the annotation made by Agić and Vulić (Agić and Vulić, 2019), Tiedemann labeled the stored languages with IDs that associate them with the name of the source *corpus* for the training datasets (Tiedemann, 2020).

The *Department of Arts and Culture, and the Department of Science and Innovation* of the South African Republic, more than two decades later, continue to fund several NLP projects related to African languages, such as HLTs, SADiLar⁵ (*South African Center for Digital Language Resources*), SARIR⁶ (*South African Research Infrastructure*

⁵**SADiLar:** is a language training center focusing on all the official languages of South Africa, in the humanities and social sciences, in the linguistic-technological domain. The center is supported by the Department of Science and Innovation for the creation, management and distribution of digital language resources. Available at: <https://sadilar.org/index.php/en/>

⁶**SARIR:** SARIR is an intervention high-level strategic and systemic approach to providing research infrastructure

Roadmap).

JW300 is a project that involves 343 languages and a total of 1,335,376 articles, around 109 million sentences, 1.48 billion tokens. This data is collected by tracking publications made on the jw.org portal. and cover topics from different areas. The primary language of published information is English. Each published article has an identifier to identify it in any language. These articles were converted to clean text in HTML format with one sentence per line, having aligned more than 50 thousand language pairs with more than 90 thousand parallel sentences per language pair on average. The high-resource languages (English, French, German, Portuguese and Italian) stood out in performance (Agić and Vulić, 2019).

FAIR Forward – Artificial Intelligence for all⁷ is a project that is part of the “*Digital Transformation for Sustainable Development* from German Federal Ministry for Economic Cooperation and Development (BMZ) and implemented by GIZ. FAIR Forward collaborates closely with the other flagship projects of this initiative, *the global e-learning platform Atingi, the Centers for Digital Transformation in Africa and Asia, the BMZ Digilab, the Data Economy and Data4policy* as well as business initiative projects *Make-IT*. At the African level, Ghana, Rwanda, Kenya, South Africa and Uganda collaborate. While in Asia it has Indonesia and India. With a view to achieving three objectives: *Access to Training Data and AI Technologies for Local Innovation, strengthening local technical know-how on AI and Developing Policy Frameworks for Ethical AI, Data Protection and Privacy*. FAIR Forward, the Mozilla Foundation and local partners from Rwanda, Uganda and Kenya, contribute to the development of *open AI training datasets* in the languages Luganda and Kiswahili⁸ and Kinyarwanda, which under funding from its local government and partners, developed a national *chatbot* that reports on the status of the Covid-19 pandemic in the local language, Kinyarwanda (Forum, 2022). South Africa has produced documents to standardize its local languages⁹, including for the education sector¹⁰. The measures contained in these documents

across the public research system, building on existing capabilities and strengths and taking advantage of future needs.

⁷Details at <https://www.bmz-digital.global/en/overview-of-initiatives/fair-forward/>

⁸It is spoken by over 150 million people.

⁹<https://www.gov.za/documents/>

¹⁰https://www.gov.za/sites/default/files/gcis_document/201409/langframe0.pdf

help preserve the cultural significance of each language, consequently, they contribute to the tools and resources that serve its community of language researchers (du Toit and Puttkammer, 2021).

3.3 Corpus in African languages

Artificial Intelligence applications use, for their experiments, preferably huge data sets to improve their performances. In this sector there is a deficit for African languages, in some cases compromising the results of these experiences and in other cases hindering their implementation. Corpus are an important component of the NLP tools necessary for developing your multilingual solutions, such as in the development of programs to perform tasks such as automatic text translation, information extraction, text classification, sentiment analysis, text summarization, among other tasks. (du Toit and Puttkammer, 2021). For its construction, specialists from related areas are needed, such as linguists, translators, computer engineers, as well as native speakers of the language (Siminyu et al., 2020). African languages classified among low-resource languages lack a varied corpus. The few found in the literature are from institutional and individual or independent initiatives with a little more than a dozen African languages, with emphasis on:

- SADiLAR, South African Government repository for datasets in local languages: <https://repo.sadilar.org/handle/20.500.12185/1> (du Toit and Puttkammer, 2021);
- Vector words in 157 languages, available at: <https://fasttext.cc/docs/en/crawl-vectors.html>
- Kinyarwanda dataset <https://digitalumuganda.com/dataset/> corpus that serves as the basis for the Government of Rwanda's chatbot functionalities.
- The Alliance of Digital Humanities Organizations (ADHO) held its annual conferences from 2009 to 2019. The data available at <https://aflat.org/> serves to catalog its results and make them available to the community of researchers.
- Universal Dependency: coprora developed by a group of independent researchers, available at: <https://universaldependencies.org/format.html>

- Tatoeba is a project composed of more than 300 languages spoken around the world, including audio of the words, available at: https://tatoeba.org/pt-br/stats/sentences_by_language (Agić and Vulić, 2019).

The collection of data from native speakers of African languages, as well as on other continents, has followed some legal assumptions, regardless of whether the data is in textual format, images, audio or video, whether via web tracking or other alternatives (Siminyu et al., 2020).

3.3.1 Corpus construction techniques

The process of building a corpus follows at least five steps, namely: identification of a primary source of data (data source), definition of the protocol, pre-processing of texts, annotations of the corpus data, part-of- speech tagging¹¹ (level of morphological decomposition) and lemmatization (du Toit and Puttkammer, 2021; Loubser and Puttkammer, 2020b). However, the form of implementation according to the specificity of each language or vernacular may imply some changes from one researcher to another, as well as depending on the methodology. It has already been used in bag of word, Markov model (Hidden Markov Models) (Loubser and Puttkammer, 2020b), text alignment under heuristics (Tiedemann, 2014) and currently it is recurrently used in artificial neural networks with various configurations (Loubser and Puttkammer, 2020b). However, pre-processing texts in African languages is a relatively difficult task due to the fact that many of these languages use lexicons borrowed from the official languages spoken in the respective countries and in some cases, the lack of orthographic uniformity due to the lack of a writing standard (Siminyu et al., 2021). For these cases, manual text cleaning is used first, after which regular expressions are used to check terms in official languages using linguistic detectors (Siminyu et al., 2021; Agić and Vulić, 2019). However, for Tiedemann, the use of this tool is preceded by cleaning characters and strings that violate Unicode encoding principles using the re-coding forced encoding mode (Tiedemann, 2020).

3.3.2 Main sources of data collection

A data source represents the source that feeds the rest of the process of creating NLP technologi-

¹¹Part-Of Speech Tagging is preferable to be done by linguists

cal tools. This is where the main problem lies in building technological tools for solving NLP problems in African languages (du Toit and Puttkammer, 2021; Niekerk et al., 2017). While there tends to be little writing in African languages, there is some media content online in local languages. Some international media powerhouses, such as the BBC and Voice Of America, have versions of their websites aimed at African audiences with content exclusively in African languages (Siminyu et al., 2021), as do a few religious institutions (Tiedemann, 2020). Siminyu et al. present three ways of collecting data, namely: data scraping from online sources, which, through the provision of online data in African languages, performs automatic or manual collection to verify accuracy, in addition to transcriptions of TED talks/films, transcriptions of radio and texts. The translators, with sentences created from scratch based on certain themes, with highly experienced local translators. And later online collection by tracking; and finally audio recordings, with the conversion of texts into audio using recordings of actors (Siminyu et al., 2021), as well as crowdsourcing (Nzeyimana, 2020). In other experiments, a set of recordings was used in a large corpus of speech with around 100 hours per language. To achieve this, standard mobile devices and dedicated software were used – Lig-Aikuma. The recorded data was repeated by reference speakers to improve signal quality and translated orally into French (Adda et al., 2016). Having the internet as the basis for obtaining texts in main and non-main languages, another research was carried out with more than 200 languages from four continents, with greater emphasis on some spoken in Africa (between main and non-main languages), in which, given a document written in several languages, an attempt was made to identify certain parts of the same document in other languages (Hiroshi Yamaguchi and Tanaka-Ishii, 2012). The language identification process is preceded by the segmentation of documents into each of the languages in the document, duly labeled, in a small amount of data, on the one hand, due to the scarcity of data in non-main languages, on the other hand, due to the difficulty of using certain machine learning methods with reduced learning corporuses.

With around 500 GB of compressed data for 2,961 language pairs, covering 555 languages. Unlike the annotation made by Agić and Vulić (Agić and Vulić, 2019), Tiedemann labeled the stored languages with IDs that associate them with the

name of the source corpus for the training datasets (Tiedemann, 2020). Briefly, the data is collected on the Web in video, text, audio format (Adda et al., 2016; Agić and Vulić, 2019; Tiedemann, 2020), in languages with greater resources and through experienced speakers they are re-recorded and translated into African languages in text format (Adda et al., 2016). In some cases, actors speaking local languages perform with free themes in environments (Siminyu et al., 2021, 2020).

3.3.3 Machine Learning

To build language models, machine learning has been widely used; Successful examples for language models have generally used supervised learning combined with reinforcement learning. For Agić and Vulić to build linguistic tools for natural language processing without supervision, it is not possible to achieve the minimum quality required (Agić and Vulić, 2019). In either form, artificial neural networks demonstrate excellent results (Loubser and Puttkammer, 2020b; Agić and Vulić, 2019; Tiedemann, 2014). BERT with simple two-layer variations was used for language model training of low-resource languages (Nzeyimana, 2020).

3.4 Analysis and Discussion

NLP resources are used to perform various tasks, generally related to texts and voice. For texts, are used in sentiment analysis, text summarization, automatic translation, document classification, information extraction, document similarity search, keyword extraction, etc (Sefara et al., 2022). As for voice, the tasks are text-to-speech conversion, audio analysis, audio transcription or automatic speech recognition, music information retrieval, audio classification, real-life applications, etc. (Li and Màrquez, 2010; Sefara et al., 2022). NLP provides technology for, for example, people with visual impairments, to obtain information, such as books, in audio form, allowing access to information, which is one of several forms of social inclusion.

The Figure 2 presents the types and quantification (percentage) of NLP tools for African languages, based on Table 1. It is worth adding that all the tools mentioned here are based on a generally annotated linguistic corpus.

Languages of African origin spoken exclusively in Portuguese-speaking African Countries (PALOP) are among those most in need of NLP resources. Angola, for example, has more than

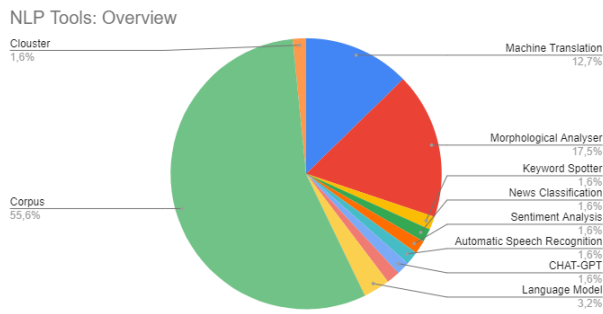


Figure 2: NLP tools overview

eleven languages of African origin, only three (Kimbundu e Umbundu) have unannotated linguistic corpus, in inaccessible repositories (Hiroshi Yamaguchi and Tanaka-Ishii, 2012), and Kikongo, which was included in a phylogenetic study carried out in the former territory belonging to the Kongo kingdom (Schryver et al., 2015). The same applies to Mozambique, which has few resources for the Emakhuwa language (Ali et al., 2021).

Artificial Neural Networks (ANN) are widely used as the main tool in building language models for machine translation (Wang et al., 2019), and researchers in African languages have adopted them (Loubser and Puttkammer, 2020b).

On the other that, (Nzeyimana, 2020) argues that language models pre-trained on high-quality monolingual corpora generally present the best performances, especially for morphologically rich languages. This is the case for most African languages.

However, to create the corpus, common principles in NLP are observed, which include cleaning the texts (as plain text), detecting words in the official Western languages (main languages) using the monolingual and multilingual gold standard, annotating the corpus, which are generally done in four different ways: Part-Of-Speech Tagging (POS-Tagging), Name Entity Recognition (NER), compound analysis and lemmatization (Nzeyimana, 2020; Loubser and Puttkammer, 2020b). According to our findings, apart from the aflat (Pauw and Gilles-Maurice de Schryver, 2009) which will have ended in 2019, and (Hiroshi Yamaguchi and Tanaka-Ishii, 2012), the corpus projects in African languages, all others are in progress, according to their own calendar.

Future research should be directed towards building annotated corpus of languages that are not included in Table 1, and/or improving the performance of existing resources in that table, to equate

the diversity of annotated and unannotated corpus of South African languages.

4 Conclusions

There has been research carried out in African languages for some time, however the ratio between researchers and languages without technological resources is enormous, although the barrier of scarcity of funding (Siminyu et al., 2021), scarcity of content, the lack of a grammatical standard and spelling rules, especially in languages spoken in more than one country (Sands, 2017), there is a need to study them to build the basis of technological development and reduce the difference with languages with more resources. The Table 1 presents some of the main NLP tools existing in African languages. Development cannot occur where there are linguistic barriers (Kanana Erastus and Erastus, 2013), as languages are present in the daily lives of their speakers and serve as a work, educational tool, when seeking medical appointments, etc. and are also a means of technological inclusion for their speakers (Kandybowicz and Torrence, 2017). In this regard, African languages lagged behind, with few human resources motivated to research in this area. The only reason for the technological delay of African languages is not the economic aspect, but also those related to cultural appreciation, political interest and other aspects (Kanana Erastus and Erastus, 2013). Swuhalli, a language spoken by more than fifty million people in Africa (Kandybowicz and Torrence, 2017; Pauw and Gilles-Maurice de Schryver, 2009), would have much more resources if other governments of countries that have territories inhabited by speakers of this language had the same concern as the government of Rwanda, Uganda and South Africa, which, following these efforts, built a corpus of all its official languages and beyond, some of which are licensed for free use for non-commercial purposes (du Toit and Puttkammer, 2021; Agić and Vulić, 2019) and some cross-language translator tools (Loubser and Puttkammer, 2020b). Kikongo spoken in around four countries, with a universe of around 7 million native speakers, has one or another technological tool without due attention from institutions and researchers in these countries (Schryver et al., 2015). It appears that there is some technological development of a given African language in countries that adopt them as official languages. For example, since 2013, Google Translator has

translated from English to Zulu and Facebook has provided an interface in Chichewa, Kiswahili, Zulu and Shona, which can also be found on the Ubuntu (Keet, 2016) operating system. For the linguistic diversity of the continent, these advances, although motivating, are still insignificant. However, South Africa is a model to be taken into account, especially for giving more visibility to its other official languages (Siminyu et al., 2021; Kandybowicz and Torrence, 2017; Loubser and Puttkammer, 2020b; Niekerk et al., 2017; Sands, 2017; Schlunz, 2018).

Acknowledgements

To Projecto de Desenvolvimento de Ciência e Tecnologias (PDCT) – Angola for funding the scholarship and INESC-ID and GAIPS, the research laboratories where I am a guest during my PhD.

References

- Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroué, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. *Breaking the unwritten language barrier: The bulb project*. *Procedia Computer Science*, 81:8–14. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.
- Željko Agić and Ivan Vulić. 2019. *JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Association for Computational Linguistics.
- Felermimo D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malawi. 2021. *Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique*. *arXiv:2104.05753v1 [cs.CL] 12 Apr 2021*, 2021.
- Ikechukwu Ignatius Ayogu and Onoja Abu. 2021. *Automatic Diacritic Recovery with focus on the Quality of the training Corpus for Resource-scarce Languages*. In *2020 IEEE 2nd International Conference on Cyberpac (CYBER NIGERIA)*, pages 98–103, Abuja, Nigeria. IEEE.
- Sebastian Dom and Koen Bostoen. 2015. Examining variation in the expression of tense/aspect to classify the Kikongo Language Cluster. *Africana Linguistica* 21, 163-211. *Africana Linguistica*, 21.
- Jakobus S. du Toit and Martin J. Puttkammer. 2021. *Developing core technologies for resource-scarce nguni languages*. *Information*, 12(12).
- Roald Eiselen and Martin Puttkammer. 2014. *Developing text resources for ten south african languages*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).
- 2022 World Economic Forum. 2022. *Chatbots RE-SET Framework: Rwanda Artificial Intelligence (AI) Triage Pilot*. World Economic Forum, Rwanda.
- Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. 2012. *Text Segmentation by Language Using Minimum Description Length*.
- Fridah Kanana Erastus and Erastus. 2013. Examining african languages as tools for national development: The case of kiswahili. *The Journal of Pan African Studies*, 6:41–68.
- Jason Kandybowicz and Harold Torrence. 2017. *Africa's Endangered Languages: Documentary and Theoretical Approaches*. Oxford University Press.
- C. Maria Keet. 2016. *An assessment of orthographic similarity measures for several African languages*. ArXiv:1608.03065 [cs].
- Joseph Ki-Zerbo. 2010. *Historia geral da africa i: Metodologia e pre-historia da africa*.
- Hang Li and Lluís Màrquez, editors. 2010. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA.
- Melinda Loubser and Martin J. Puttkammer. 2020a. *Viability of Neural Networks for Core Technologies for Resource-Scarce Languages*. *Information*, 11(1):41. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- Melinda Loubser and Martin J. Puttkammer. 2020b. *Viability of neural networks for core technologies for resource-scarce languages*. *Information*, 11(1).
- Sozinho Matsinhe and Mbiavanga Fernando. 2008. *A preliminary exploration of verbal affix ordering in Kikongo, a Bantu language of Angola*. *Language Matters*, 42(2):332–358. <http://dx.doi.org/10.1007/s40858-017-0164-2>.
- Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. 2017. *Rapid Development of TTS Corpora for Four South African Languages*. In *Inter-speech 2017*, pages 2178–2182. ISCA.
- Antoine Nzeyimana. 2020. *Morphological disambiguation from stemming data*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4649–4660, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Guy De Pauw and Gilles-Maurice de Schryver. 2009. [African Language Technology: The Data-Driven Perspective](#). pages 79–96.
- Hermenegildo Pinto and Ana Alexandra Silva. 2022. [Língua umbundu: caminhos para a sua preservação](#). *Revista angolana de ciências*, 4(1).
- Bonny Sands. 2017. [The Challenge of Documenting Africa’s Least-Known Languages](#). In *Africa’s Endangered Languages: Documentary and Theoretical Approaches*. Oxford University Press.
- Bonny Sands. 2018. *Language revitalization in Africa*.
- Georg I. Schlunz. 2018. [Usability of Text-to-Speech Synthesis to Bridge the Digital Divide in South Africa: Language Practitioner Perspectives](#). In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–10, Plaine Magnien. IEEE.
- Gilles-Maurice de Schryver, Rebecca Grollemund, Simon Branford, and Koen Bostoen. 2015. [Introducing a state-of-the-art phylogenetic classification of the Kikongo Language Cluster](#). *Africana Linguistica*, 21:87–162.
- Tshephisho Joseph Sefara, Mahlatse Mbooi, Katlego Mashile, Thompho Rambuda, and Mapitsi Rangata. 2022. [A toolkit for text extraction and analysis for natural language processing tasks](#). In *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6.
- Bienvenu Sene-Mongaba. 2015. [The making of lingala corpus: An under-resourced language and the internet](#). *Procedia - Social and Behavioral Sciences*, 198:442–450. Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015).
- Kathleen Siminyu, Sackey Freshia, Jade Abbott, and Vukosi Marivate. 2020. [AI4D – African Language Dataset Challenge](#). ArXiv:2007.11865 [cs].
- Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David I. Adelani, Amelia Taylor, Jamiil Toure ALI, Kevin Degila, Momboladji Balogoun, Thierno Ibrahima DIOP, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. 2021. [Ai4d – african language program](#).
- Jörg Tiedemann. 2014. [Rediscovering Annotation Projection for Cross-Lingual Parser Induction](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Jörg Tiedemann. 2020. [The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT](#). ArXiv:2010.06354 [cs].
- Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. [Cross-lingual BERT transformation for zero-shot dependency parsing](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

**Third Workshop on Digital Humanities and
Natural Language Processing (3rdDHandNLP)**

Can rules still beat neural networks?

The case of automatic normalisation for 18th-century Portuguese texts

Leonardo Zilio

FAU Erlangen-Nürnberg, Germany
leonardo.zilio@fau.de

Rafaela R. Lazzari

UFRGS, Brazil
rafaelalazzari@hotmail.com

Maria José B. Finatto

UFRGS, Brazil
mariafinatto@gmail.com

Abstract

In this paper, we tested whether fine-tuned neural machine translation (NMT) models can produce better results than a rule-based method for the task of normalising historical medical documents written in 18th-century Portuguese. We used a replacement glossary as basis for the rule-based method, and tested three NMT models against it in an in-domain setting and in two out-of-domain scenarios. In-domain results showed that the rule-based method was better than off-the-shelf NMT models, and it still surpassed one of the in-domain fine-tuned models. The fine-tuned models showed their efficacy on out-of-domain settings, where only one NMT model did not surpass the rule-based method in one scenario.

1 Introduction

Working with historical documents has proven time and time again to be an enormous challenge for Natural Language Processing (NLP) (cf. [Quaresma and Finatto, 2020](#); [Vieira et al., 2021](#); [Cameron et al., 2022](#); [Zilio et al., 2022](#)). While there is progress in the field, most of the tools developed for working with natural language have modern iterations of the language as focus, and only a few studies have been dedicated to computationally process historical documents as they are, and even fewer such studies exist for historical Portuguese.

To help alleviate the historical gap between historical and modern-era texts, researchers started resorting to normalising the writing of historical documents ([Piotrowski, 2012](#); [Bollmann and Søgaard, 2016](#); [Bawden et al., 2022](#)); that is, they started updating the spelling of historical texts based on modern-day orthographic rules. However, this normalisation work is mainly done manually and is thus very

time consuming.

This study has the main objective of exploring ways of automatically normalising documents, so that less work has to be spent in converting the writing of historical documents into modern-day standards, and allowing for the mass-normalisation of larger corpora. Taking advantage of already existing normalised, available corpora, and of recently developed machine translation models, we analyse how three multilingual neural machine translation (NMT) models fare when compared to a rule-based normalisation model that uses a static glossary as main reference.

The main contributions of this paper are the following:

- The release of a dataset for fine-tuning and testing NMT on the task of normalising historical medical texts written in Portuguese.
- The release of scripts for automatic normalisation of historical documents¹. These scripts are fairly simple to use and can also be applied in other tasks related to sequence-to-sequence translation.
- A comparison of three off-the-shelf NMT models and their fine-tuned version in the task of normalising historical texts, both in and out of domain.
- An error analysis that shows what might still pose problems for fine-tuned NMT models in this context.
- A support for the further analysis of historical medical documents, such as the

¹These scripts and datasets can be found on the following repository: https://github.com/uebelsetzer/automatic_normalisation_of_historical_documents.

one carried out by [Lazzari and Finatto \(2023\)](#).

- The advancement of the project *Corpus Histórico da Linguagem da Medicina em Português do Século XVIII* [Historical Corpus of Medical Language in 18th-century Portuguese]²

The remainder of the paper is organised as follows: Section 2 discusses other work related to the normalisation of historical texts; Section 3 briefly describes our historical corpus and the four methods used for automatically normalising historical sentences; Section 4 presents the results of the automatic normalisation experiments; in Section 5, we discuss some key issues detected when analysing what went wrong with the automatic normalisation; Section 6 describes an experiment with out-of-domain historical documents, to evaluate the robustness of in-domain fine-tuned and rule-based methods; Section 7 sums up and discusses some aspects of the experiments with in- and out-of-domain normalisation, and discusses future work.

2 Related work

Several studies have been dedicated to the normalisation of historical documents in various languages, including Portuguese. As for automatic normalisation, most of the approaches seem to have stopped before the advent of transformer models, which make this study unique in applying the most recently developed NMT models based on the transformers architecture.

Most studies involving NMT use an encoder-decoder, character-based architecture based on long-short term memory (LSTM) models (cf. [Bollmann and Søgaard, 2016](#); [Domingo and Nolla, 2018](#); [Domingo and Casacuberta, 2019](#)). While these studies make sense, by modelling the spelling normalisation problem as a character-based replacement, much similar to what rule-based systems have done, [Tang et al. \(2018\)](#) have already hinted that subword tokens can provide a better solution to

²For more information about this project, please visit the following website (in Portuguese): <https://sites.google.com/view/projeto38597>. The project website also contains more information about the historical medical corpus that we use in this study.

character-based models. This would naturally lead to the use of transformer-based architecture. However, as far as we could verify, the study by [Tang et al. \(2018\)](#) is the only one testing transformers for this task to date, and Portuguese is not among the tested languages.

In terms of languages, the focus of studies on automatic normalisation have been on European languages. [Bollmann \(2019\)](#) developed a large comparison of automatic normalisation methods for English, German, Hungarian, Icelandic, Portuguese, Slovene, Spanish, and Swedish. Studies with less languages involve the work of [Domingo and Casacuberta \(2019\)](#) for Slovene and Spanish, [Bawden et al. \(2022\)](#) for French, and [Robertson \(2017\)](#) for English, German, Icelandic, and Swedish. For Portuguese, we could only find the above-mentioned work of [Bollmann \(2019\)](#), who used a corpus of letters from the 15th to 19th century that was made available by the Post Scriptum project ([CLUL, 2014](#)).

More recently, researchers at the University of Évora started working with text normalisation. [Cameron et al. \(2023\)](#) propose a categorisation of variants, which can support the normalisation of historical Portuguese texts, and [Olival et al. \(2023\)](#) present and discuss the normalisation of six documents that belong to the *Parish Memories*. There are also some papers that use normalised versions of Portuguese documents for different NLP tasks, such as named entity recognition ([Zilio et al., 2022](#)) and textual complexity ([Zilio et al., 2023](#)).

Considering the work that has been done, this study is the first to present an automatic approach for normalising historical medical documents in Portuguese, and possibly the first to leverage existing multilingual, transformer-based NMT models for the normalisation task.

3 Methodology

In this section, we briefly describe the corpora that were used for in-domain fine-tuning of NMT models and for glossary extraction, and also for in- and out-of-domain testing. We also describe our baseline rule-based method and present the multilingual NMT models.

3.1 Corpora

In this study, we used a total of three corpora, all of them written in Portuguese in the 18th century: a historical medical corpus, which is the focus of this study and was used for fine-tuning and testing NMT models, and for extracting a glossary for the rule-based approach; a historical corpus of censal information collected by priests in different Portuguese parishes; a historical corpus of letters collected within the Post Scriptum project (CLUL, 2014). All corpora were semi-automatically aligned at the sentence level using OmegaT’s aligner tool³. This process allowed the generation of TMX files, which were then used for further preprocessing the aligned texts for the different tasks.

Our historical medical corpus was originally transcribed from three books written in the 18th century: *Observações Medicas Doutrinaes de Cem Casos Gravissimos* [Medical and Doctrinal Observations of a Hundred Severe Cases] (Semedo, 1707), *Postilla Religiosa, e Arte de Enfermeiros* [Religious Postil, and Art of Nurses] (de Sant-Iago, 1741) and *Aviso a’ Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health] (Mauran, 1794). Since we needed to manually normalise each of the texts used in this study, we only selected a few documents from each of the books, aiming at a balanced corpus.

Some documents from the *Parish Memories* corpus have recently undergone a normalisation process (Olival et al., 2023), so we took advantage of this fact and used this corpus as an out-of-domain test for our automatic normalisation systems. For this task, we used the six documents related to Vila Viçosa (a location in Portugal) that are currently available in normalised format⁴. Each document was written by a different author, and each refers to a parish in Vila Viçosa: Nossa Senhora das Ciladas, Nossa Senhora da Conceição, Pardais, Santa Ana de Bencatel, São Bartolomeu, and São Romão.

³OmegaT is an open-source tool used for computer-assisted translation. It is available at: <https://omegat.org/>.

⁴The original texts are available on CIDEHUS’s website (<https://www.cidehusdigital.uevora.pt/portugal1758>), while the normalised versions are provided as annex in Olival et al. (2023).

While the *Parish Memories* provide an out-of-domain test set, it is still a somewhat structured type of text, in which each paragraph contains very specific information about a census that was carried out in 1758 in Portugal. To provide an even less structured test to our automatic normalisation models, we resorted to a selection of handwritten letters from the Post Scriptum collection (CLUL, 2014). The full corpus from the 18th century contains 758 letters⁵. However, due to the semi-automatic nature of the sentence-alignment process, we randomly selected 10 letters from the corpus (taking care of selecting five from each of the two available subcorpora). Here is the list of letters that were used in this study: CARDS1082, CARDS1089, CARDS2108, CARDS2707, CARDS3148, PSCR0515, PSCR0613, PSCR1648, PSCR2526, and PSCR4643.

A very important caveat needs to be presented here: our historical corpus of medical documents was normalised having modern Brazilian Portuguese as reference, while the other two corpora were normalised having European Portuguese as reference. As such, for instance, while in our corpus we normalised words like “cousa” to “coisa” [thing], this was not done in the other two corpora, as “cousa” can still be found in European Portuguese. This certainly had an impact in the results of the experiments and should be kept in mind when observing the results that we present in this study.

Table 1 presents the data information for each corpus. As can be seen in the table, our historical medical corpus has a total of 5,584 types, while the version with modernised spelling has 5,341 types. This gives us an idea of how much spelling variation there was in the original corpus: we have 1.05 type for each type in the normalised corpus. This variation is larger in both other corpora, and a possible reason for this is that they are both based on handwritten documents by several different authors, while our medical corpus was published in printed form and was the work of three authors. The medical corpus also clearly

⁵All files can be freely downloaded from the Post Scriptum website: <http://teitok.clul.ul.pt/postscriptum/index.php?action=downloads>.

distinguishes itself from the others by the amount of tokens per sentence, with around 54 T/S against ~39 and ~17 for the Parish Memories and the Post Scriptum, respectively. The medical corpus is marked by a constant use of semi-colons, where in a modern writing probably a full stop would be used. The much smaller sentence size in the Post Scriptum corpus is mostly due to the genre, but the normalisation probably also contributed to this: many of the original letters have little to no punctuation, and the normalisers added punctuation, including full stops, in the normalisation process, which might have led to a more modern use of punctuation.

The medical corpus was further split into train, development (dev) and test sets, for fine-tuning NMT models. Table 2 shows the number of tokens, types and sentences in each split, considering original and normalised versions of the corpus. An important detail in the design of the splits is that the texts used in the train and dev sets were different from the ones used in the test set. The train and dev sets were a random selection of sentences (90% for train and 10% for dev) from these texts:

- *Aviso*: chapters 2, 8, and 13, all from the second part of the book.
- *Observaçoes*: observations 42, 88, and 92.
- *Postilla*: chapters 17, 22, 29, 30, 32, 33, 34, 40, 41, 42, 43, 44, 46, 47, 48, and 58, all from the second part of the book⁶.

For the test set, we used one text from *Aviso* (chapter 5, also from the second part of the book) and from *Observaçoes* (observation 92), and two chapters from the *Postilla* (chapters 1 and 7, also from the second part of the book).

3.2 Rule-based method

The rule-based normalisation method was planned as a baseline for the automatic normalisation process. We used a glossary of aligned original and normalised words that was automatically extracted from the combined train and development corpus.

⁶The chapters in the *Postilla* are smaller, so we had to select more chapters than in the other two books in order to balance the dataset.

To extract this glossary, we first had to use a word-level aligner, to identify the pairs of historical-normalised words that actually underwent any change. For this, we used SimAlign (Sabet et al., 2020), along with the recently released Albertina model (PT-PT) (Rodrigues et al., 2023), and we carried out a semi-automatic alignment, in which instances of no alignment or of many-to-one alignments were validated manually. However, there might still be a few one-to-one wrong alignments in the dataset.

From this word-aligned dataset, we observed that 1,228 types in the original texts had a different spelling when compared to their normalised counterparts. This indicates that almost a third (31.46%) of the types needed to be normalised, reinforcing the importance of automatising the normalisation process.

The word-aligned dataset was used as input for the glossary. We also manually removed the entry “as” = “às”, because “as” might simply be the plural form of the feminine definite article “a” [the_{feminine}], and not the merge of preposition “a” and the plural form of the feminine definite article, as it was represented in the automatically extracted glossary. After this cleaning, the resulting glossary was then used as a replacement dictionary.

The first step in the process for the rule-based normalisation was to tokenise each sentence with NLTK’s⁷ word tokeniser. Then each token was checked against the glossary to verify if any replacement was needed. Phrases longer than one token were processed separately in a similar way. If a word or phrase was present in the historical text, then it was replaced with its normalised form. The rules also ensured that punctuation was correctly rendered in the output (for instance, by removing space between a word a comma, which is very common in historical documents).

3.3 Neural machine translation models

We used three multilingual neural machine translation (NMT) models:

- **opus-mt-tc-big-itc-itc (OPUS)**⁸: this

⁷NLTK’s website: <https://www.nltk.org/>.

⁸<https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-itc-itc>.

	Medical		Parish Memories		Post Scriptum	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	24504	24815	9561	9661	2547	2549
Types	5584	5341	2381	2027	950	852
Type Ratio	-	1.05	-	1.17	-	1.12
Sentences	453	453	244	244	147*	147
T/S	54.09	54.78	39.18	39.59	17.33	17.34

Table 1: Corpus information. Type Ratio = division of types in the original by types in the normalised corpus; T/S = tokens per sentence.

* The number of sentences in the PS original corpus was based on the normalised version, as there are very few or no instances of punctuation in some of the original letters.

	Train		Dev		Test	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	18047	18286	2038	2067	4419	4462
Types	3386	3213	826	803	1372	1325
Type Ratio	-	1.05	-	1.03	-	1.04
Sentences	342	342	38	38	73	73
T/S	52.77	53.47	53.63	54.39	60.53	61.12

Table 2: Information about the individual data splits. Type Ratio = division of types in the original by types in the normalised corpus; T/S = tokens per sentence.

model was originally trained in the scope of the OPUS-MT project (Tiedemann and Thottingal, 2020; Tiedemann, 2020). It comprises 19 languages from the Italic family, including Portuguese, and it was trained with all possible language combinations. This model is by far the smallest, as the final folder of the fine-tuned model has a size of only around 2.3GB, while the other two have a size of almost 7GB each.

- **mbart-large-50-many-to-many-mmt (mBart)**⁹: this model was originally developed by Tang et al. (2020) and comprises 50 languages, including Portuguese, trained in a many-to-many fashion, *i.e.* all possible language pairs are included in the training set.
- **nllb-200-distilled-600M (NLLB)**¹⁰: the NLLB paper (Team, 2022) caused much stir in the machine translation community, as it offers a huge combination of languages, including low-resource languages.

⁹<https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt>.

¹⁰<https://huggingface.co/facebook/nllb-200-distilled-600M>.

This model builds on the idea of leveraging higher resourced languages for the automatic translation of low resourced ones. Because it involves so many languages, it is also a less focused model, and while it works in advancing the machine translation state of the art for some low resourced languages, it might not perform as brilliantly for highly resourced ones, such as Portuguese.

These models were first tested as they are provided by their developers, to set some baselines for the models themselves, and then they were also used in a fine-tuning pipeline, where our training and development datasets were used to adapt these models to our normalisation task. For fine-tuning, we used the standard Transformers library, as provided by Huggingface¹¹ (Wolf et al., 2020). All models were fine-tuned with the same parameters, except for batch size, as the larger models were simply too large for our single graphics card NVidia RTX 4090 (with 24GB RAM) to handle: learning rate of 2e-5, weight decay of 0.01, and 100 epochs; batch size was 16 for OPUS, 6 for

¹¹<https://github.com/huggingface/transformers>.

NLLB and 4 for mBart. All other parameters were left as default. At the end of the fine-tuning process, the best model was selected based on BLEU scores (Papineni et al., 2002), as evaluated in the default implementation of SacreBLEU (Post, 2018) in Huggingface’s Evaluate library¹².

None of these models differentiate between Brazilian and European Portuguese, the two varieties that we are working with, so we simply used the tags “por” (European Portuguese) and “pob” (Brazilian Portuguese) for reference, but none of the models were actually trained to differentiate between the two. As such, we actually fine-tuned the models to translate from Portuguese into Portuguese, but using a dataset that was specifically curated for this normalisation task.

3.4 Evaluation

We evaluated all models using the BLEU score metric (Papineni et al., 2002). Several papers criticise the use of BLEU scores, including the paper that proposes SacreBLEU (Post, 2018), which is the implementation that we used via the Evaluate library from Huggingface, as explained in the previous subsection. BLEU is a metric that compares the number of n-grams in the target text with reference text(s), and produces a score from 0 to 100. Because any size of n-gram can be used, it is a metric that has to be well-detailed in the methodology to be reproducible, something that SacreBLEU addresses very well.

Another downside of BLEU is that it bases the correctness of a target text on the basis of reference texts. These references may or may not be good target texts themselves, and they do not necessarily invalidate other alternative, equally correct translation options for a given source text. As such, a low BLEU score might be just a reflex of different translation choices in the reference texts. While this issue can be mitigated by using several reference texts for each test sentence, several references are not always available. In our case, however, most of the time, there is no alternative correct option for a given token in the normalisation pipeline. Most historical words can only be normalised to one single form in the modern

Model	SacreBLEU
Baseline models	
OPUS*	47.57
mBart	30.73
NLLB	40.64
Rule-based model	
Replacement glossary	83.26
Fine-tuned models	
OPUS	75.05
mBart	88.20
NLLB	83.65

Table 3: SacreBLEU scores based on our test set.

* We prepended >>pob<< to the source text, as required by the system. Without prepending the language ID, the model translated the source text into Spanish, and it achieved a BLEU score of 7.77. Prepending >>pob<< actually made the fine-tuned system perform around 2 BLEU points worse on both test sets, so we did not prepend >>pob<< for the fine-tuned model.

spelling paradigm, so the issue of multiple references will rarely apply, making BLEU a perfectly sound choice for evaluating a spelling normalisation task. The choice of SacreBLEU also ensures that any researcher can use the exact same format of BLEU when trying to reproduce this study, as we used the default parameters of the metric.

4 Results of the automatic normalisation

As it can be seen in Table 3, the baseline models (without fine-tuning) perform very poorly on our data, with the highest BLEU score being achieved by OPUS at 47.57. Meanwhile, our simple rule-based system achieved 83.26 in the BLEU scale. Surprisingly, even after fine-tuning, the rule-based system remained very competitive, and was still better than the OPUS model by more than 8 points and was only barely surpassed by the NLLB model, giving us an initial answer to our main question in the title of this paper. However, mBart showed a great improvement with fine-tuning (an increase of more than 54 points) and achieved the highest score, almost five points higher than the second-best model.

Although mBart was able to beat the rule-based model with some margin, the results seem to show that a well-curated glossary

¹²<https://github.com/huggingface/evaluate>.

could actually be a better option for automatic normalisation, depending on the needs of the researchers and on the equipment available, as fine-tuning mBart is not a computationally cheap task: it requires a computer with a high-end graphics card, even for a small train and development dataset such as ours. The NMT models are also much slower at the inference phase (*i.e.*, when they are producing the normalised text): while the rule-based method is almost instantaneous for our test set, the NMT models take a few minutes on a good GPU, and up to an hour on a 12th Gen Intel Core i7-1260P CPU.

5 Error analysis

To better understand what types of errors were prevalent in the fine-tuned NMT models, we conducted an error analysis, focusing on sentences that had a low BLEU score¹³. We analysed the sentences checking for missing translations, overtranslations, hallucinations, and any common pattern that we could identify that helped bring down the scores.

In the OPUS model, we noticed that several normalised sentences missed portions of the source text, and we also noticed that the system was producing several hallucinations. One example of hallucination from OPUS is the following:

cuja verdadeira descarga se deve fazer por via de purga, & naõ de sangria, & por estas razões as purguei com felicidade ¹⁴

mas sim com os quais se descarregou a verdadeira felicidade, e purgaram as sangrias, e por isso só por causa de que os reis purgaram purgas ¹⁵

¹³For our reference in the error analysis, we computed the sentence-level BLEU score separately from the one presented in Table 3. We considered low BLEU scores the ones that deviated by at least one standard deviation from the model’s mean in the test set. This means that those considered as bad sentences in one model could actually be better than some “good” sentences in another model.

¹⁴Free translation: *whose true elimination should be done via purge, and not bleeding, and because of this I happily purged them.*

¹⁵The text does not make much sense, so we tried to keep a more literal translation: *but actually with those that true happiness eliminated itself, and purged bleedings, and that’s why only because of that the kings purged purges.*

The NLLB model had much less salient issues, as they were more focused on single tokens, and involved miss-normalisations or lack of normalisation, and the substitution of historical words with synonyms. A similar error pattern was observed for mBart, but it presented only a few cases of replacement with a synonym. These usually single-token errors included lack of or non-removal of diacritics in most cases; this involved the model simply not changing the word in the source text. In the NLLB model, we also observed a few hallucinations, mostly just short repetitions of words, such as “poreis poreis poreis” [(you will) put put put] and “sumas sumas” [(that you) disappear disappear]. For mBart, one curious case was the replacement of Outubro [October] with Novembro [November] in a segment, but the rest were mostly small issues.

6 Robustness test: use in out-of-domain historical texts

As the models that we developed and fine-tuned were focused on specialised historical medical language, we wanted to check how much information had also been gained for normalisation on out-of-domain texts. This was an experiment in “knowledge” transfer, where we try to observe how much of the information that was gathered from medical texts can be transferred to the normalisation of texts from other domains. For this, we tested our models on the normalised texts from the *Parish Memories* corpus and on normalised letters from the Post Scriptum dataset, as we described in Section 3.

Table 4 presents the results for all the models, including the non-fine-tuned ones, as a comparison for how much improvement was brought about by the fine-tuning procedure, and for how difficult the task was in relation to the normalisation task in our medical corpus. We can clearly see that the Post Scriptum dataset was much harder to normalise. Some of the originally transcribed texts do not have punctuation and have many abbreviations, which are usually extended in the normalised version. This made it much more difficult for all models to achieve a good normalisation, as they were not fine-tuned to add punctuation or to extend abbreviations. In

the *Parish Memories*, with the exception of mBart, which had an almost 6-point worse BLEU score, the results of the baseline models did not vary too much from the results in the medical dataset. In both out-of-domain datasets, the rule-based method scored more than 7 BLEU points higher than OPUS, the best baseline NMT model.

When looking at the fine-tuning improvement, we see that, except for OPUS on the Post Scriptum dataset, all NMT models performed above the rule-based method. As expected, all of them performed worse than on the in-domain dataset, but the results in the *Parish Memories* were still much better than the ones achieved by their non-fine-tuned baselines, with improvements ranging from around 13 BLEU points for OPUS up to ~33 points for mBart. In the Post Scriptum dataset, the improvements were more modest, ranging from ~7 BLEU points for OPUS up to ~29 points for mBart. In this out-of-domain test, we also see that the fine-tuned NLLB model seems to really be able to draw on its information about 200 languages for keeping it robust, as it achieved the best score on both datasets, clearing more than 3 BLEU points from mBart.

Model	SacreBLEU	
	Parish Memories	Post Scriptum
Baseline models		
OPUS*	45.72	27.08
mBart	24.94	6.79
NLLB	39.55	20.80
Rule-based model		
Replacement glossary	53.70	34.56
Fine-tuned models		
OPUS	58.77	34.05
mBart	58.10	36.01
NLLB	61.41	39.34

Table 4: SacreBLEU evaluation scores on out-of-domain corpora.

* We prepended >>pob<< to the source text, as previously explained on Table 3.

7 Final remarks

In this paper, we set out the task of testing neural machine translation (NMT) models for automatically normalising historical medical documents. We compared fine-tuning meth-

ods with a rule-based implementation of a replacement method mainly based on a glossary, and the results showed that the rule-based method was indeed a strong baseline for the NMT models. It surpassed the non-fine-tuned NMT models in all scenarios, scoring up to ~52 BLEU score points higher in the in-domain test.

After fine-tuning the NMT models, as expected, all models improved over their baseline versions, but only mBart was clearly superior to the rule-based method. OPUS still scored ~8 BLEU points lower, and NLLB was only marginally superior (less than one point). As such, as a preliminary answer to the question in the title of this paper, we can say that rule-based systems can still be superior to neural-network-based methods in some scenarios, and they are certainly much less complicated to implement and less power-consuming.

The fine-tuning advantage of the NMT models was, however, clearly shown in the out-of-domain test, where all models scored at least 4 points higher than the rule-based method when tested on the *Parish Memories* dataset, and only OPUS was not able to beat the rule-based method on the Post Scriptum dataset, showing that the fine-tuned models are better able to transfer the information gathered from one domain to another. Still, when the task was too far off, as in the case of the handwritten letters of the Post Scriptum dataset, a post-editor with the task of normalising texts would probably be better served by a glossary replacement method. Such method at least would be less intrusive, as most errors would be in the form of non-changed input, rather than an erroneously changed input (such as the hallucinations produced by NMT). However, a detailed post-editing task would need to be developed to better test this hypothesis.

In terms of fine-tuning improvement over the non-fine-tuned baselines, mBart was the model that had the best result in all scenarios. On the opposite side, OPUS was the model that showed the least improvement in all tasks. The OPUS model we used was specifically trained on Italic languages, which gave it the best result in all baseline tests. However its fine-tuned version was inferior to the rule-based method both in and out of domain,

only being able to beat the rule-based method (and marginally also mBart) when tested on the *Parish Memories*.

It was interesting to see that, in the out-of-domain task, NLLB was superior to mBart in both datasets, probably due to the larger linguistic scope of the model. It is still not yet fully clear if this is caused by mBart being perhaps more prone to overfitting, and OPUS (and also NLLB) being then less prone to overfitting, or if the out-of-domain tasks rely more on the breadth of linguistic information that was used in the original training of the models. These are all questions that we plan to investigate in the future, as further tests are needed to verify them.

The work on this paper also sets out a methodology for replicating the work using other corpora, covering other time periods, other domains, and even other languages. With the scripts that are now available on Github¹⁶, it is also possible to train models for the task of translating (instead of normalising) historical documents into modern languages using a very similar methodology as we presented in this paper, so one further future task is to create data for testing these models in a diachronic intralingual translation setting.

Acknowledgements

We would like to thank the Chair of Computational Corpus Linguistics at FAU Erlangen-Nürnberg and the PPG-LETRAS at UFRGS for support, and also the following Brazilian Institutions: PROPESQ - UFRGS – CNPq, for a PIBIC undergraduate research grant; CNPq, for funding a Productivity Research Grant (06/2019, 308926/2019-6) and a research project (26/2021 – PDE – 200051/2023-7); and FAPERGS-CAPEL, for funding a research project (06/2018 – INTERNAC., 19/2551-0000718-3).

References

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. [Automatic normalisation of](#)

¹⁶Please check our repository: https://github.com/uebelsetzer/automatic_normalisation_of_historical_documents.

[early Modern French](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. 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 3885–3898.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional lstms and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139.

Helena Cameron, Fernanda Olival, Renata Vieira, and Joaquim Santos. 2022. [Named entity annotation of an 18th-century transcribed corpus: problems and challenges](#). In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 18–25. CEUR.

Helena Freire Cameron, Fernanda Olival, and Renata Vieira. 2023. Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *LaborHistórico, Rio de Janeiro, ISSN*, pages 2359–6910.

CLUL. 2014. *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*.

Fr. Diogo de Sant-Iago. 1741. *Postilla religiosa, e arte de enfermeiros: guarneçada com eruditos conceitos de diversos authores, facundos, Moraes, e escurituarios*. Oficina de Miguel Manescal da Costa, Lisboa, Portugal.

Miguel Domingo and Francisco Casacuberta. 2019. Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents. In *New Trends in Image Analysis and Processing—ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pages 59–69. Springer.

Miguel Domingo and Francisco Casacuberta Nolla. 2018. Spelling normalization of historical documents by using a machine translation approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*, pages 129–137. European Association for Machine Translation.

- Rafaela Radünz Lazzari and Maria José Bocorny Finatto. 2023. Exame do vocabulário médico no português no século xviii: contribuições da lexicometria para o desenho de um dicionário histórico. *Mandinga-Revista de Estudos Linguísticos* (ISSN: 2526-3455), 7(1):102–123.
- G. Mauran. 1794. *Aviso a' Gente do Mar sobre a sua Saude*. R. Typ. de João Antonio da Silva, Lisboa, Portugal. Translated from the French original edition and extended with some notes by Bernardo José de Carvalho.
- Fernanda Olival, Helena Freire Cameron, Fátima Farrica, and Renata Vieira. 2023. As Memórias Paroquiais (1758) do atual concelho de Vila Viçosa. *Callipole*, 29:85–128.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool Publishers.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Paulo Quaresma and Maria José Bocorny Finatto. 2020. [Information extraction from historical texts: a case study](#). In *Proceedings of the Workshop on Digital Humanities and Natural Language Processing, co-located with International Conference on the Computational Processing of Portuguese, DHandNLP@PROPOR, Evora, Portugal, March 2, 2020.*, pages 49–56. CEUR.
- Alexander Robertson. 2017. *Automatic Normalisation of Historical Text*. Ph.D. thesis, School of Informatics, University of Edinburgh.
- João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of Portuguese with transformer Albertaina PT. *arXiv preprint arXiv:2305.06721*.
- Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.
- João Curvo Semedo. 1707. *Observações Medicas e Doutrinaes de Cem Casos Gravissimos*. Oficina de Antonio Pedrozo Galram, Lisboa, Portugal.
- Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#).
- NLLB Team. 2022. [No language left behind: Scaling human-centered machine translation](#).
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.
- Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (Alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Leonardo Zilio, Maria Finatto, and Renata Vieira. 2022. [Named entity recognition applied to Portuguese texts from the XVIII century](#). In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 1–10. CEUR.
- Leonardo Zilio, Maria José B. Finatto, Renata Vieira, and Paulo Quaresma. 2023. [A natural language processing approach to complexity assessment of 18th-century health literature](#). *Domínios de Lingu@gem*, 17.

Revealing Public Opinion Sentiment Landscape: Eurovision Song Contest Sentiment Analysis

Klara Kozolić and Gaurish Thakkar and Nives Mikelić Preradović

Faculty of Humanities and Social Sciences,

University of Zagreb, Zagreb 10000

kkozolic@m.ffzg.hr, gthakkar@m.ffzg.hr, nmikelic@m.ffzg.hr

Abstract

This paper discusses the use of sentiment analysis to evaluate public opinion towards the Eurovision Song Contest of 2022 and 2023 using social media comments in English, Croatian and Spanish. The study aims to determine the sentiment expressed in the comments and analyse the distribution of positive, negative, and neutral sentiments. The paper also reviews prior research on sentiment analysis related to the Eurovision Song Contest and outlines the research questions and methodology used in this study. The authors hypothesise that there are differences in how the Eurovision Song Contest is perceived by different language speakers and on different social media platforms. The methodology involves identifying relevant social media comments, studying the Eurovision timeline, collecting and preprocessing the data, and performing sentiment analysis and named entity recognition. The paper concludes by summarising the key contributions of the work and discussing future research.

1 Introduction

Social media comments have transformed into one of the main channels for opinion expression. This information is a great indicator of the public's opinion on numerous topics or events. Quantitative information, such as the number of comments, retweets, and likes that authors can view, supplements the textual and audiovisual content. Social media has always been essential in providing a place for discussing public thoughts and expressing views on publicly broadcast entertainment events.

The analysis of sentiment in tweets has previously been explored as a means of forecasting the outcome of certain events, such as the stock market's rise and fall (Pagolu et al., 2016). Additionally, there has been exploration of the relationship between the physical performance of basketball players and the sentiment expressed in their tweets (Xu and Yu, 2015).

The Eurovision Song Contest is a competition for the best song in European countries, held annually in May since 1956. During the event, fans post on social media content related to the competition. This provides an opportunity to evaluate methods for evaluating the sentiments of the public regarding the Eurovision Song Contest. Digital humanities strive to perform research in the humanities field with the use of technology (Svensson, 2010). In that sense, this study is related to the digital humanities, as sentiment analysis is used to extract an opinion on a social event. This paper aims to perform sentiment analysis of comments in English, Spanish, and Croatian for the Eurovision Song Contests of 2022 (ESC2022) and 2023 (ESC2023) to determine the sentiment expressed in the text. We analyse the sentiment distribution in order to measure the extent of positive, negative, and neutral subjectivity conveyed in the text. This allows for a more comprehensive assessment of the collective sentiment towards the contest and related elements.

The subsequent sections of this paper are organised in the following manner: Section 2 provides an overview of prior research conducted on the data connected to the Eurovision Song Contest. Section 3 enumerates the research questions we aim to explore. Section 3 outlines the methodology utilised in our study. The details of data preparation and preprocessing steps are described in Section 5. Section 6 of the paper provides the outcomes of the experiments, offering a thorough analysis. In Section 8, the key contributions of the work are summarised, and future research is discussed.

2 Related Work

There are several previous studies related to modelling various aspects associated with Eurovision. In 2019, Demergis (2019) analysed tweets in English and Spanish collected during the 2019 Eu-

rovision Song Contest to identify sentiment and rank target performers. The ranking was then compared with the ranking derived from the televoting during the Eurovision Grand Final. [Kumpulainen et al. \(2020\)](#) utilised sentiment analysis tools to analyse more than a million tweets in order to forecast the outcomes of ESC televoting. The study established a correlation between the sentiment ratings of tweets by comparing predicted ranks with the ranks obtained from televoting. [Koski and Persson \(2017\)](#), utilised the AFINN word lexicon to examine the tweets related to *Melodifestivalen*, an annual Swedish music competition organised by the Swedish national public TV. The authors employed several ranking methodologies on the participants and compared them with the actual ranks in the real world. [García and Tanase \(2013\)](#) utilised historical data from both Wikipedia and the official website of the Eurovision Song Contest to examine the cultural connections between European countries. The authors introduced a quantitative metric known as the Friend-or-Foe coefficient. A metric that uncovers the asymmetrical positive and negative connections among European countries indicates a correlation between cultural dissimilarity and voting biases.

3 Research Questions

Sentiment analysis of social media comments about Eurovision can give insight into many aspects of the song contest and its presence on social media. The research questions that we want to answer are the following:

1. What is the distribution of sentiments in the comments for each corpus? What is the distribution of sentiments in the comments for each social media platform?
2. What is the number of mentions of each participant country in each corpus?
3. Which country or artist are the negative comments most common about for each ESC year in each corpus?
4. Which country or artist are the positive comments most common about for each ESC year in each corpus?

We hypothesise that there were differences between how ESC2022 and ESC2023 appeared to different language speakers and on various social

media platforms. By answering our research questions, we can get a clear idea of these differences and verify if they were in accordance with the turn of events for each ESC year.

4 Methodology

The method described in more detail in the paper consists of the following steps:

1. Identifying social media comments about ESC2022 and ESC2023. The first step in identifying social media comments was crucial before data collection. For Twitter, it was important to see with which hashtags we could get the most relevant data that fit our initial criteria. And for Reddit, we had to find relevant subreddits and threads about the topic of ESC2022 and ESC2023.
2. Studying the Eurovision timeline to learn when the most comments are generated. The information about the timeline of big events before, during, and after ESC helped with determining the time frame during which the comments would be collected. For example, what were the dates of the semifinals and the final, the dates of press conferences or national song selections, etc?.
3. Collecting the comments from Twitter, Reddit, and YouTube that fit the timeline and the requirements that were set.
4. Preprocessing the data. The scraped data needed to be treated differently for each social media platform. For Twitter and YouTube, we had to manually clear out the comments in other languages or comments that were not talking about the topic of ESC. Each remaining comment was manually tagged with the appropriate target language: En, Hr, or Es. For Reddit, we also had to manually clear out the comments that led to discussions that diverged too much from the topic of ESC.
5. Constructing the corpora. Corpora was constructed with Python, with the help of language tags.
6. Performing sentiment analysis on the corpora with ML models.
7. Performing Named Entity Recognition (NER) on the corpora with Python.

8. Extracting the value from and analysing the tagged data.

To answer the questions of this research, the following natural language processing and programming tasks will be performed:

- Sentiment analysis
- Named entity recognition
- Regular expressions (RegEx)

To obtain the number of mentions for each country in each corpus, we used NER and RegEx in Python. NER was performed using pre-trained pipelines for each language, using publicly available spaCy models `hr_core_news_sm` for Croatian¹, `es_core_news_sm` for Spanish² and `en_core_web_sm` for English³. We counted the identified entities found for the country using Python. We have also performed the same task using RegEx with Python to find mentions of countries. We were able to find more mentions using RegEx than NER, but the overall results for the analysis were the same.

The data for this project consists of social media comments in English, Spanish, and Croatian for the Eurovision Song Contests of 2022 and 2023. Section 5 describes the process of data extraction and preprocessing in detail.

5 Data: Extraction and Preprocessing

The data for this project was collected between February and May 2023. The data that was chosen for the research was aimed at social media comments in English, Spanish, and Croatian for the Eurovision Song Contests of 2022 and 2023.

The Croatian language was chosen as our mother tongue. The English language was chosen as the most popular language on the Internet and it's the most popular foreign language in Croatia, with the largest percentage of primary school, high school and university students studying it (Kapović, 2022). Spanish language is chosen as the language whose culture is popular in Croatia thanks to music and TV series and it has a growing number of students in formal education (Kapović, 2022; Urquijo Sánchez, 2021). The data was divided into three corpora to obtain the most correct sentiment

¹<https://spacy.io/models/hr>

²<https://spacy.io/models/es>

³<https://spacy.io/models/en>

analysis results, using a different model for each language.

5.1 Scraping tools

For the purposes of scraping the social media comments, we have used the following three scrapers in Python:

- Twitter Scraper Selenium⁴ is a Selenium-based tweet scraping tool. The tool is provided under MIT licence and allows the scraping of public tweets from a specific user profile or a hashtag.
- Universal Reddit Scrapper (URS)⁵: The tool is provided under MIT licence and allows scraping subreddits, redditors, and submission comments.
- YouTube Comment Scraper⁶: The tool is provided under Apache licence 2.0 and allows the scraping of YouTube comments under videos.

5.2 Data preparation and processing

The scraping for each social media site was done in batches, and each website required a different approach, which will be described in the following sections.

5.2.1 Twitter

Using `twitter-scraper-selenium`, we have scraped the data under hashtags `"#ESC2022"` and `"#ESC2023"` to obtain the data about the target Eurovision Song Contests. We have collected the data from January until June 2022 for ESC2022 and from January to June for ESC2023. We have collected the biggest batch of tweets for the dates of semi-final one, semi-final two, and final for each of the contests. The tweets before these dates still contained a great number of opinions and news about the national selections, the Eurovision contests, the songs, and the artists, and the tweets after these dates were a great indicator of reactions to the results of the contests. After the scraping, we needed to filter out the comments in other languages and manually tag the language of the comments in order to facilitate their classification into language corpora. The comments that used the hashtags `#ESC2022` and `#ESC2023` but were

⁴<https://pypi.org/project/twitter-scraper-selenium/>

⁵<https://github.com/JosephLai241/URS>

⁶<https://pypi.org/project/youtube-comment-scraper-python>

not talking about Eurovision were also manually deleted.

5.2.2 Reddit

Using URS, we have scraped the comments about ESC2022 and ESC2023 from the threads in the "r/croatia", "r/hrvatska", "r/eurovision" and "r/spain" subreddits.

Since the subreddits were in the respective languages, we did not have to classify the languages, but we still had to manually clear the comments in which users completely diverged from the topic of Eurovision.

5.2.3 YouTube

Using youtube-comment-scraper-python we have scraped the comments under live broadcasts of Eurovision and of the Croatian song selection festival, Dora.

Like for Twitter, we needed to manually clear the comments in other languages and tag the language of the comments we were left with for easier classification into language corpora.

5.3 Creation of the language corpora

The language corpora were generated in CSV format using Python, incorporating the language tags derived from the scraped data. At the end, we had 2778 comments in Croatian, 1811 comments in Spanish, and 7613 comments in English. These corpora were used for sentiment analysis, with models for each language.

Lang	Tokens	Types	Lang Distribution
English	251009	15829	62.39%
Croatian	46136	11217	22.77%
Spanish	55861	8762	14.84%

Table 1: The number of tokens and types and language distribution

5.4 Sentiment analysis

The purpose of the sentiment analysis is to determine the subjective classification of the given text, that is, to identify if a post is positive, negative, or neutral. The sentiment classification was performed using publicly available models from Hugging Face. We have employed a multilingual model for sentiment analysis: Twitter-XLM-roBERTa-

base⁷ for Spanish, Twitter-roBERTa-base⁸ for English, and Cro-Frida⁹ for Croatian. The models utilised are built upon state-of-the-art transformers (Vaswani et al., 2017; Conneau et al., 2020) architecture and have undergone fine-tuning specifically for sentiment analysis in their respective languages.

The sentiment classification was performed using publicly available models from Hugging Face. We have used a multilingual model: twitter-XLM-roBERTa-base¹⁰ for sentiment analysis in Spanish, Twitter-roBERTa-base¹¹ for sentiment analysis in English and Cro-Frida¹² in Croatian. The models utilised are built upon state-of-the-art transformers (Vaswani et al., 2017; Conneau et al., 2020) architecture and have undergone fine-tuning specifically for sentiment analysis in their respective languages.

The examples of sentiment classified data (negative in these cases) for each corpus are:

- *Unpopular opinion: Sweden performance doesn't deserve to win. Too many mistakes. #Eurovision #Sweden #EUROVISION #ESC2022*, negative (0.8983), neutral (0.0916), positive (0.0101)
- *E ovo je bilo dosadno. Ne ponovilo se #ESC2023*, 0
- *Loreen ha ganado pero sintiéndolo mucho no se lo merecía y no pienso discutir esto #ESC2023 #EUROVISION #EUROVISION2023*, [{"score": 0.5728, "label": "negative"}]

6 Results and Analysis

To evaluate the corpora, we have randomly selected 300 comments from each corpora and tagged their sentiment manually. Tagged corpora were compared to corpora containing sentiment predictions using Python. The Table 2 shows the precision, recall, F-1 measure and accuracy of the models used for sentiment analysis.

⁷<https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment>

⁸<https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

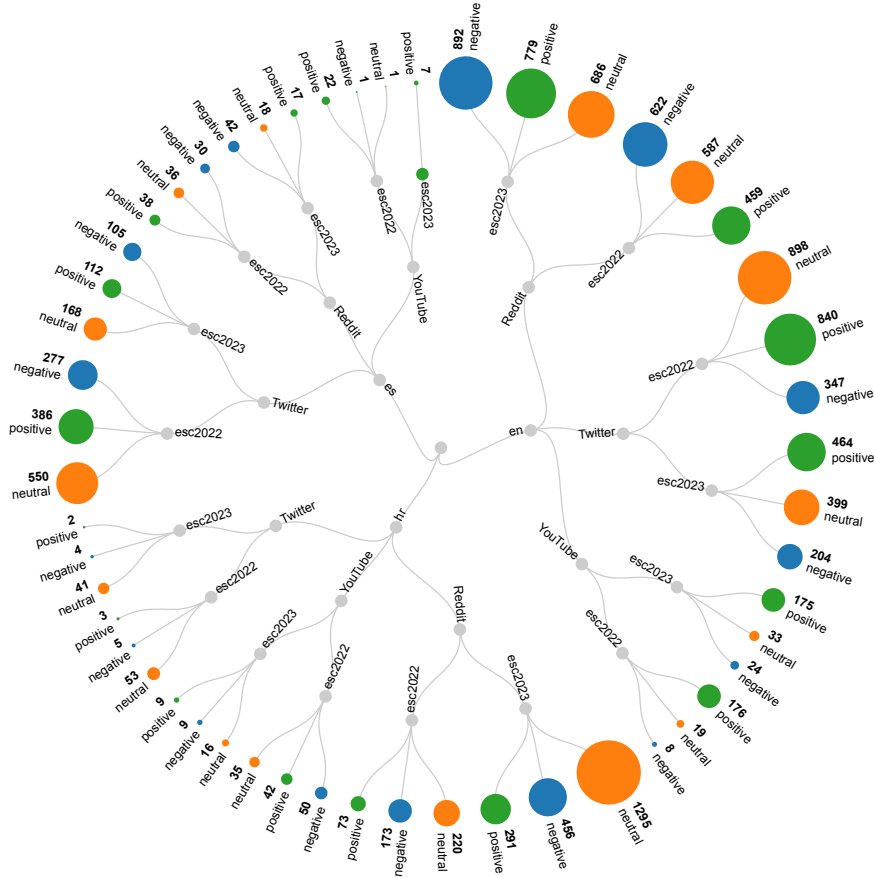
⁹<https://huggingface.co/thak123/Cro-Frida>

¹⁰<https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment>

¹¹<https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

¹²<https://huggingface.co/thak123/Cro-Frida>

Figure 1: Corpora categorised based on their language, social media source, year, and sentiment label.



Lang	P	R	F-1	Acc
English	67.21	66.78	66.53	66.78
Croatian	65.27	62.58	62.04	62.58
Spanish	0.7453	74.67	74.50	74.67

Table 2: Evaluation of sentiment analysis approaches employed for categorising tweets. P:precision, R:recall, Acc: accuracy.

Lang	Year	Pos	Neg	Neu
English	2022	1475	977	1504
	2023	1418	1120	1118
Croatian	2022	118	228	308
	2023	302	469	1352
Spanish	2022	446	308	587
	2023	136	147	186

Table 3: The distribution of sentiments in each corpus

6.1 The distribution of sentiments in the comments for each corpus

After the task of sentiment analysis, the following sentiment distribution was found:

Table 3 shows sentiment distribution for each Eurovision year in each corpus. For the English corpus, we can note that there were the most neutral comments for ESC2022 and the most positive comments for ESC2023. Taking into account that there were a lot more positive than negative comments for ESC2022, we can conclude that both of these

contests were mostly perceived as positive. For Croatian, each Eurovision had the greatest number of neutral comments. As opposed to English, we can note a higher number of negative than positive comments. Thus, we can conclude that both ESC2022 and ESC2023 were perceived as negative. For Spanish, each Eurovision had the greatest number of neutral comments as well. ESC2022 had more positive than negative comments, and ESC2023 had more negative than positive com-

ments. This coincides with the good placement of Spain in 2022 and the bad one in 2023.

6.2 The distribution of sentiments in the comments for each social media

The distribution of sentiments in the comments for each social media can be seen in Table 4.

Social Media	Year	Pos	Neg	Neu
Reddit	2022	570	825	843
	2023	1087	1390	1999
Twitter	2022	1229	629	1501
	2023	578	313	608
YouTube	2022	240	59	55
	2023	191	33	49

Table 4: The distribution of sentiments in each social media

On Reddit, each Eurovision had the most neutral comments. Each Eurovision also had more negative than positive comments, which leads to the conclusion that both years were perceived as negative. Twitter shows the opposite situation, with the most neutral but more positive comments. Thus, we can conclude that both years were perceived as more positive. On YouTube, the greatest number of comments were positive for each Eurovision. These results indicate that there was more critical discussion regarding various aspects of ESC2022 and ESC2023 on Reddit than on Twitter and YouTube. This fits the format of Reddit as one of the most popular sites for opinion exchange and discussion.

6.3 The number of mentions of each participant country in each corpus

Figure 2 shows the number of mentions of each country participating in the Eurovision in 2022 and 2023 for the English corpus. For ESC2022, the most mentioned country is Ukraine, the winner of that year’s edition. We can see Spain and the UK mentioned the most after Ukraine. The UK took second place and offered to host in place of Ukraine in 2023. Spain did well with the juries and televotes, placing third. The least mentioned countries are the ones that did not qualify for the final or were placed in lower positions, typically out of the top 10. For ESC2023, the most mentioned country is Finland, a runner-up for 2023. Finland was a fan favourite and received the most votes from the audience. The second country is Sweden, the winner of ESC2023. The least mentioned countries are the ones that did not qualify for the final

or were placed in lower positions, typically out of the top 10, except for Germany and Spain, which ended up with some of the worst results, which did not sit well with the audience. Additionally, three countries dropped out of the competition in 2023 and were not mentioned at all for that year: Montenegro, North Macedonia, and Bulgaria.

Figure 3 shows the number of mentions of each country participating in the Eurovision in 2022 and 2023 for the Spanish corpus. The Spanish corpus did not reflect the general opinion about Eurovision as much as the opinions of Spanish people. In both Eurovision song contests, Spain was the most mentioned country. Spain was mentioned more in ESC2022 when its entry placed third, and many fans thought it would have won if it hadn’t been for the war in Ukraine. The higher number of mentions of other countries for ESC2022 also indicates that ESC2022 was more viewed than ESC2023 in Spain.

Figure 4 shows the number of mentions of each country participating in the Eurovision in 2022 and 2023 for the Croatian corpus. Like in the Spanish corpus, the Croatian corpus reflects the opinions of Croatians regarding ESC2022 and ESC2023. The most mentioned countries, alongside Croatia, were the countries that placed higher in each ESC edition. As opposed to Spain, the bigger interest for Croatian data is in ESC2023, where Croatia qualified for the finals for the first time since 2017.

6.4 Country: negative comments

The extraction of mentions of each country and sentiments related to that country showed insight into which country faced the most negative comments in each corpus. Starting with the English corpus, the country with the most negative comments associated with it for ESC2022 was Ukraine. This result was not surprising, as Ukraine won in 2022. Many people believe it was a political win due to the Russian invasion of Ukraine that started in 2022 and because of which Russia was banned from participating in ESC2022 (Welslau and Selck, 2023). For ESC2023, the country with the most negative comments was Finland, the runner-up for that year. Following ESC2023, there was a lot of discussion involving Sweden and Finland, where fans were arguing that Finland was the audience favourite and should have won, saying that the juries wanted Sweden to win. These discussions led both Finland and Sweden to be mentioned in both a negative and positive light, the most. The countries with the

Figure 2: Frequency distribution of nations cited in the English corpus.

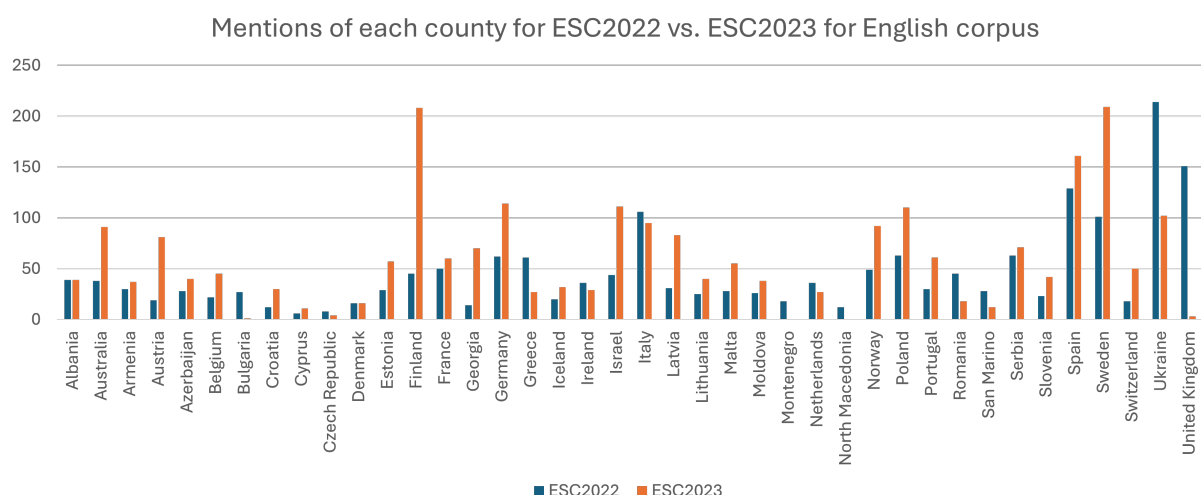
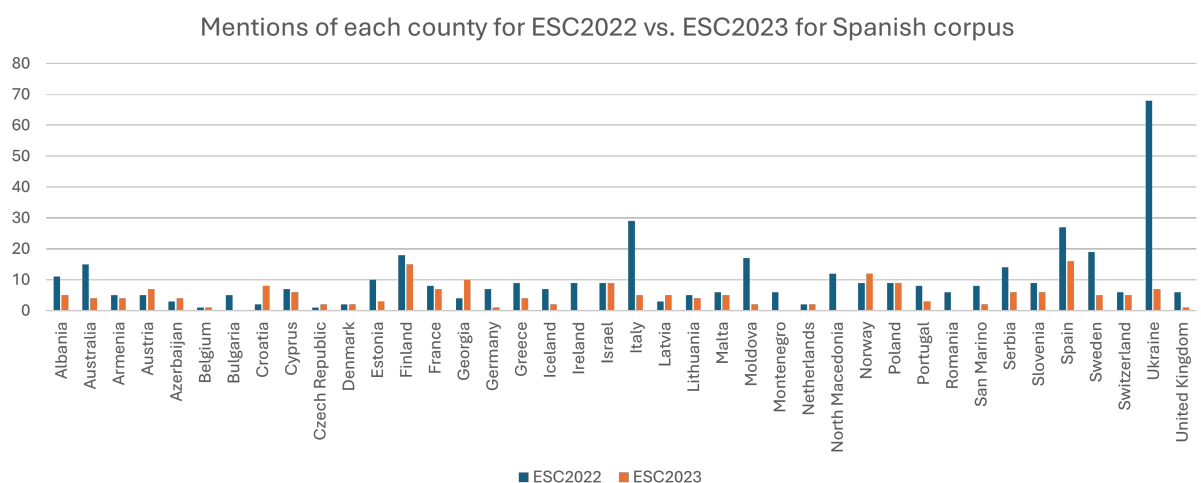


Figure 3: Frequency distribution of nations cited in the Spanish corpus



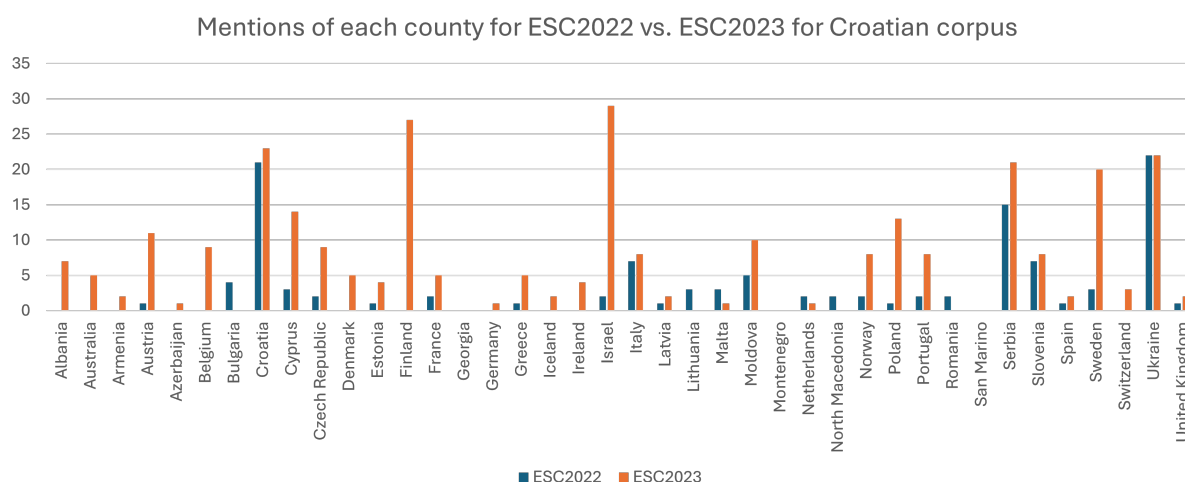
most negative mentions in the Croatian corpus for ESC2022 were Croatia and Ukraine. Croatia sent a song that didn't qualify, and the public in Croatia did not like it, while the number for Ukraine could be explained by its triumph, which many perceive as political. For ESC2023, the most negatively mentioned countries in Croatian corpus were Finland and Serbia. Finland could be explained with the same reasoning as for the English corpus: the Finland vs. Sweden discussions. As for Serbia, neighbouring countries or countries that have something in common, such as Slavic roots in this case, tend to mention one another more in both a positive and negative light. In the Spanish corpus, the most negatively mentioned country for ESC2022 was Spain. These comments can be explained by the

fact that Spain placed third, and the Spaniards believed it should have won, so they were expressing their dissatisfaction. For ESC2023, the most negatively mentioned country was Spain again. Spain received very few points from the public in 2023 and the public in Spain and worldwide didn't seem to receive the modern rendition of flamenco well.

6.5 Country: positive comments

The extraction of mentions of each country and sentiments related to that country showed insight into which country faced the most positive comments in each corpus. In the English corpus for ESC2022, the country with the most positive comments was Ukraine. While it generated the most negative comments, it also generated the most pos-

Figure 4: Frequency distribution of nations cited in the Croatian corpus



itive ones, since a lot of people expressed their solidarity with Ukraine and liked the song they performed. For ESC2023 in the English corpus, the country with the most positive mentions was Finland. However, it's important to note that Sweden had only 4 comments marked as positive, fewer than Finland, yet it achieved a higher positive average score. This demonstrates really well how people had polarising opinions and discussions on which of the two countries should have won the ESC2023. In the Croatian corpus for ESC2022, the most positively mentioned countries were Croatia and Poland. ESC2022 was not as commented on and followed as ESC2023 in Croatia, ending with Croatia being mentioned as the most positive and most negative. For ESC2023, the most positively mentioned countries were Croatia, Finland, Israel, and Ukraine. All four countries were discussed on social media and had both positive and negative comments. In the Spanish corpus for ESC2022, the most positively mentioned country was Spain, which placed third that year, and the comments were convinced of Spain's win. For ESC2023, the most positively mentioned countries were Israel and Georgia. Israel placed third in 2023 and was a very discussed entry for that year, while Georgia didn't qualify for the finals and many people believed it should have.

7 Discussion

After the analysis of the data, we saw that each corpus had a great number of comments tagged as neutral. If we take into account only positive and

negative comments, the distribution is the following: for the English corpus, the sentiment was positive for both ESC years; for the Croatian, the sentiment was negative for both ESC years; and for the Spanish, the sentiment was positive for ESC2022, and it was negative for ESC2023. The distribution of sentiments for each social media platform revealed that Twitter and YouTube had more positive comments, and that Reddit had more negative comments for each ESC year. The number of mentions for each participant country followed the outcome of each ESC year in the English corpus, with the most mentioned countries being the countries that scored higher or qualified for the final. The mentions of countries in Croatian and Spanish corpora also followed the outcome of ESC, but with more mentions of their own countries or their neighbouring countries. The negative comments about countries revealed that in each corpus, the most negatively mentioned countries were the ones whose victory was controversial or those who people felt deserved to win. In Croatian and Spanish corpora, we again saw the focus shifting more to their own or neighbouring countries. There is a similar situation with positive comments about countries, where the most discussed countries had more positive mentions than the others, with more details about their own countries in Croatian and Spanish comments. As demonstrated in the analysis, Eurovision can reflect the cultural and political situation in Europe and its entries, and the public can use it as a means of social commentary (López Zapico, 2023). Its cultural importance is manifested through the sense of unity of Europe and belonging

that it's aiming to emit to its viewers through the practices of so-called "postcards", videos before each performance that show the artist and natural beauty of each country (Coupe and Chaban, 2019). In their work (Spierdijk and Vellekoop, 2006) have found that not only geographical proximity plays a role in Eurovision votes, but also shared cultural background and cultural similarities such as a language result in genuine votes. In our data, we have found that countries with geographical proximity, language, shared culture and history tend to mention one another more in both positive and negative contexts. If we take Spain as an example, we can see that it favours Italy, but dislikes France. What it has in common with these two countries is the same language group and cultural and historical ties. The same can be said for Croatia that either favours or dislikes Serbia, its neighbouring country with shared cultural patterns and language group. The same was found in a study by (Spierdijk and Vellekoop, 2006) where, for example, Croatia was favoured by Slovenia, but disliked by Hungary, both neighbouring countries or Spain that favoured Italy.

8 Conclusion

This research gave an insight on how sentiment analysis can demonstrate public opinion towards the Eurovision Song Contest. The analysis was performed on data in three different languages: English, Spanish, and Croatian. The research found that the English corpus showed general opinion very well, while the Spanish and Croatian corpora showed opinions and sentiments more concentrated on their own entries, very popular entries, or entries that have something in common with these countries, such as language group or neighbouring geographical location. The NLP tasks of sentiment analysis and NER proved to be beneficial in painting the picture of opinion towards the Eurovision Song Contest as a whole: before, during, and after the show. The results of the analysis were in accordance with how certain entries and countries were perceived during the show and the reaction to them after the show. However, a clear limitation of our analysis are the differences between languages and language resources for the three languages we chose, cultural references and sarcasm in comments, and the bias in the data, which was shown well in Croatian and Spanish examples. Furthermore, during data collection, we noticed that

people from different countries used different social media sites to talk about Eurovision. For example, we saw that the Croatians were more active on Reddit, whereas the Spaniards were more active on Twitter. All of these limitations impact the end goal of sentiment analysis for this research: determining the sentiment about ESC in different countries. Another limitation that needs to be considered is the subjectivity of choosing both the languages and certain social media sites as sources. In order to increase the objectivity of the results, data from more social media or internet sites could be added. The mention of the artists or songs, the addition of more categories of sentiment for an even more detailed demonstration of the opinion, or the addition of comments in more languages are some of the numerous elements that could be added to expand this research. In conclusion, sentiment analysis of ESC-related comments can be beneficial for studying the social and political situation in Europe and how it reflects on ESC or for studying the entertainment value ESC brings to its fans each year.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm'an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Advances in Neural Information Processing Systems*.
- Tom Coupe and Natalia Chaban. 2019. [Creating europe through culture? the impact of the european song contest on european identity](#). *Empirica*, 47(4):885–908.
- Dimitri Demergis. 2019. [Predicting eurovision song contest results by interpreting the tweets of eurovision fans](#). In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 521–528.
- David García and Dorian Tanase. 2013. Measuring cultural dynamics through the eurovision song contest. *Advances in Complex Systems*, 16(08):1350037.
- Marko Kapović. 2022. [Strani jezici u formalnom obrazovanju u hrvatskoj](#). *Strani Jezici*, 51(2):283–309.
- Alexander Koski and Jennifer Persson. 2017. [And the winner is... : Predicting the outcome of melodifestivalen by analyzing the sentiment value of tweets](#).
- Iiro Kumpulainen, Eemil Praks, Tenho Korhonen, Anqi Ni, Ville Rissanen, and Jouko Vankka. 2020. [Predicting Eurovision Song Contest Results Using Sentiment Analysis](#), pages 87–108.

- Misael Arturo López Zapico. 2023. Europe's living a celebration? el impacto social y cultural de la europeización de españa a través de canciones y otros productos culturales. *Revista de Estudios Europeos*, 82:152–183.
- Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPE5)*, pages 1345–1350. IEEE.
- Laura Spierdijk and Michel Vellekoop. 2006. Creating europe through culture? the impact of the european song contest on european identity. *Empirica*, 47(4):885–908.
- Patrik Svensson. 2010. The landscape of digital humanities. *Digital Humanities Quarterly*, 4.
- José Ignacio Urquijo Sánchez. 2021. EL ESPAÑOL EN CROACIA, pages 385–395.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5099–5109.
- Lea M Welslau and Torsten J Selck. 2023. Geopolitics in the esc: Comparing russia's and ukraine's use of cultural diplomacy in the eurovision song contest. *New Perspectives*, page 2336825X231222000.
- Chenyang Xu and Yang Yu. 2015. Measuring nba players' mood by mining athlete-generated content. In *2015 48th Hawaii International Conference on System Sciences*, pages 1706–1713. IEEE.

Could Style Help Plagiarism Detection? - A Sample-based Quantitative Study of Correlation between Style Specifics and Plagiarism

Adile Uka
Ruhr University Bochum
adile.uka@rub.de

Maria Berger
Ruhr University Bochum
maria.berger-a21@rub.de

Abstract

The paper presents an empirical study based on samples of the correlation between an author’s writing style and the plagiarism ratio in a text. Specifically, we investigate the research question whether a correlation in style can also hint to potential plagiarism, or at least some systematic copying in two texts. To gain an understanding of the characteristics a “copied” text might pertain, we collect different sample sets reaching from chapters by the Brontë sisters, over parallel samples of a plagiarism corpus up to the ChatGPT rephrased pendants of some essay pages by the main author. We also add sections by Matthew, Mark and Luke from a Brazilian Portuguese Bible translation to our samples. Results show that there exists a moderate positive correlation between style similarities and plagiarism overlap across four different genres.

1 Introduction

Plagiarism detection has always been an immensely important task in natural language processing. First, because it is an essential task in all-day publishing, second, because research in plagiarism detection also plays an elementary role in developing information retrieval-related algorithms and operations (c.f., [Potthast et al., 2010](#); [Foltýnek et al., 2019](#); [Alzahrani and Salim, 2008](#)). However, highly-paraphrased plagiarism with many word substitutions and re-ordering still presents a challenge for many systems (c.f., [Hunt et al., 2019](#); [Alvi et al., 2021](#)).

The humanities have always been an important driver for the development of technical and digital approaches to analyse and share information across time and space. This is why we also understand the historical use-cases of authorship attribution and user or author pseudonymization (also aliasing) as a means to investigate contemporary challenges of plagiarism and third-party authorship.

In this paper, we attempt to operationalize a procedure that helps us to understand whether style similarities among two texts can give us a hint on plagiarism. Strong overlaps in plagiarism between two texts usually mean that many areas in both texts (suspicious document and sample candidate) share very literal wording. This wording typically also shares style characteristics ([Eissen and Stein, 2006](#)). However, in style analysis, the focus is not on content words, but on the distribution of function words, which typically are the most frequent n words in a language. We also want to compare the text samples towards their ChatGPT-rephrased ([OpenAI, 2024](#)) versions to make a statement on how GPT’s style might differ compared to manually written texts.

2 Related Work

Even though there exists lots of research on plagiarism detection and style analysis, the complementing of both tasks was not performed very often. However, searching for style deviations to find hints for third-party authorship is a common set-up. [AlSallal et al. \(2019\)](#) perform a study on intrinsic plagiarism detection by validating the inner parameters of an author’s text. They use most-frequent-words features to represent an author’s profile, and use a classifier to evaluate whether a text was actually written by that author. Another form of intrinsic plagiarism analysis is presented in [Oberreuter and Velásquez \(2013\)](#). The authors argue that it is not always possible to measure similarity at the document-level, because the referring documents are not necessarily available. Therefore, the authors use profiles based on linguistic features at the character and word level to compare segments of a text towards the whole text. Whether these paragraphs then are on par with the profiles of the entire document can be determined by measuring the significance. With the rise of LLMs-based

text generators, we have encountered another quality of “plagiarism”. However, we hypothesize that we can still distinguish the writing style of artificially generated text from human-written text. We already found proof for this in [Zwilling and Berger \(2024\)](#). [Gao et al. \(2023\)](#) perform a study that shows that ChatGPT generated paper abstracts are unlikely to be plagiarized from the web (while their originals are moderately likely to contain plagiarism) and, further, these generated abstract have only moderately similar texts existing in the web while their originals have very similar versions existing online. This is due to the fact that plagiarism indicates at least area-wise very close style similarities.

3 Method

3.1 Data Selected

We compile a diverse data set that comprises plagiarism (in the broadest sense) data samples from several sources.

1. Excerpts from British novels by the Brontë sisters as well as some texts under discussion of being authored by William Shakespeare versus Christopher Marlowe (henceforth, Novels samples).
2. Novels samples and their ChatGPT re-products (Novels & GPT).
3. Five pairs sampled from the Webis corpus for plagiarism detection¹ ([Burrows et al., 2013](#)) (Webis plagiarism samples).
4. Webis plagiarism samples and their ChatGPT re-products (Webis plagiarism & GPT).
5. Five pages from four different German term papers by one of our co-authors versus their ChatGPT re-product (Essays & GPT).
6. The books Mark & Luke and Mark & Matthew from the Almeida Revisada in Brazilian Portuguese language (Almeida Revisada).

Novels: We use text pairs that we assume suitable for our study because earlier work showed stylistic similarities of them. First, one chapter (Ch. 8) of Charlotte Brontë’s “Jane Eyre” together with her sister’s Anne Brontë’s “The Tenant of Wildfell Hall” (Ch. 1). These works show an especially high

stylistic overlap since the sisters’ creative writing was exposed to a close exchange among each others from the beginning (c.f., [Eder et al., 2016](#)).² We downloaded these works from GitHub³. We further use texts by Shakespeare’s “Henry VI” Part 1⁴ (act 1, scene 2, featuring Joan Pucelle) versus “Henry VI” Part 2⁵ (act 4, scene 2, featuring Cade at Blackheath). There is strong evidence that both scenes could be considered to be written by Christopher Marlowe, not Shakespeare himself or him alone (c.f., [Craig and Kinney, 2009](#); [Nance, 2017](#)).⁶

Novels & GPT: We take each text solely from the Novels dataset and rephrase a ChatGPT version of it and compare it towards this version. Since ChatGPT usually shortens and summarizes such long texts, we opted for rephrasing the novels in chunks of about 200 words to ensure a proper rephrasing of the complete texts. We use the same ChatGPT prompt for the novels and the Webis data, which is “Rephrase this text as much as possible.”

Webis Plagiarism samples: We use Random Lists⁷, an online random number generator that takes a range and a list size as input. Hence, it returns 5 numeric values within a given range of numbers (the size of the Webis corpus). This way, we could safely select samples from the Webis-CPC-11 corpus for plagiarism detection to have a representative subset to investigate correlation between style and plagiarism. See [Tab. 1](#) for an example of paraphrastic plagiarism.

Webis Plagiarism & GPT: We take each single text from the Webis Plagiarism dataset and have ChatGPT rephrase a version of it. Then, we compare the rephrased version towards its original text.

Essay & GPT: We take five pages from four different German essays by one of our co-authors and compare it towards their ChatGPT re-products. This is especially interesting, as these texts are very specific and we can assume that there is not too much redundancy of these texts around in the web—a fact that GPT would benefit from. The ChatGPT prompt for the rephrasing of the essays

²Acc. Feb. 2024 <https://alanabeeblog.wordpress.com/2013/12/19/the-bronte-sisters-a-stylometric-analysis/>

³Acc. Feb 2024 https://github.com/computationalstylistics/A_Small_Collection_of_British_Fiction/tree/master/corpus

⁴Acc. Feb 2024 <http://shakespeare.mit.edu/1henryvi/index.html>

⁵Acc. Feb. 2024 <http://shakespeare.mit.edu/2henryvi/index.html>

⁶The stylistic techniques that were applied to derive that hypothesis were later re-checked within a broader statistical test by [OTA \(2023\)](#).

⁷Acc. Feb. 2024 <https://www.randomlists.com/random-numbers>

¹Acc. Feb. 2024 <https://webis.de/data/webis-cpc-11.html>

original	paraphrastic plagiarism
The explanation of this estrangement given by my grandfather, was that there had been a disagreement about land; but perhaps he may have felt some delicacy about telling his children that his unambitious marriage had contributed to render the separation permanent.	Explanation of the disposition given by my grandfather, was that there was a disagreement on the land, but perhaps he may have felt some delicacy about his children modestly says his marriage helped make the separation permanent.

Table 1: Plagiarism example from Webis sample 4325

is: “Formuliere diesen Text soweit wie möglich um” (“Rephrase this text as much as possible”).

Almeida Revisada: We also are interested in comparing Matthew and Mark, and Luke and Mark in a Brazilian Portuguese version of the Bible. The “Synoptic Gospels” are a strong use-case for plagiarism in the digital humanities. For us, these books show-case another form of similarity and thus are worth investigating. We compare Matthew and Mark, and Luke and Mark, because each of them show a strong overlap with Mark, which is historically acknowledged and computationally confirmed (c.f. Jänicke et al., 2014; Harder, 2022). We downloaded a Brazilian Portuguese version of the Bible from the Mysword Bible repository footnote. Feb. 2024 <https://www.mysword.info/download-mysword/bibles>. Precisely the “Almeida Revisada de acordo com os Melhores Textos em Hebraico e Grego” from 1967, which is a rather modern version of Almeida’s translation. We use this version, because we think it is one of the most common Brazilian Portuguese Bible translations. A version closer to Almeida’s original version would also be interesting to study. However, it is beyond the scope of this study to analyse historical spelling modification and its effect on plagiarism detection and style analysis tools. Matthew is a much longer work than Mark, hence, we cut the chapters 1, 2, 5, 6, 7, 11, and 25 in Matthew as these have no or almost no similarity edges to Mark (Harder, 2022). In Luke, we remove chapters 1, 2, and 12-17, because these also do not have many textual overlaps with Mark.

3.2 Tools Used

We use the Stylo R package⁸ (Eder et al., 2016) for calculating stylistic relations between each pair of our sample set. We choose the cosine distance measure handing over our own code of cosine distance. Then, we subtract the resulting value from 1 (that represents the cosine similarity), and receive results that range between 0 and 1. Stylo’s default

⁸Acc. Feb. 2024 <https://github.com/computationalstylistics/stylo>

cosine distance version scales these ranges back to the numeric origin scale, which allows also negative similarities scores.⁹ A positive effect of cosine measure is that it is robust towards documents of different length (Evert et al., 2016). As we want to investigate the style of the texts—not necessarily the domain content—, we set features to most frequent words (MFW), which long proved to be a good means to identify similar style (Damerou, 1975; Hoover, 2003).¹⁰ For the books in the Brazilian Portuguese version of the Bible, we use the 200 MFW, because these documents are much longer than the others. For all the other documents, we use the 100 MFW. This setting is the most intuitive while simple and effective.

We further use WCopy find (?) to calculate plagiarism overlaps among our sample pairs. Even though there are a lot of plagiarism tools available (also for free use), many of them are not very flexible and do not enable an extrinsic comparison to a local repository. We found WCopy find a useful tool as it determines the similarity based on the partition of common sub-strings (of a given length), in a bi-directional manner (Left, and Right, c.f. Tab. 2), and it also highlights closed-reading overlaps so that we can easily find very long string-overlaps when apparent. The parameters we use for “Shortest Phrase to Match” is 3. We require the system to match 100% of these words, which is a very strict setting for verbatim plagiarism.

4 Results & Discussion

Following, we describe how we investigate the correlation between both, overlapping style characteristics in two texts, and overlapping plagiarism.

4.1 Quantitative correlation measured

To calculate correlation between the style similarities and the plagiarism overlaps, we use the average

⁹The cosine similarity can range between -1 and 1. Because the distance is 1-cosine-similarity, it can range from 0 to 2, see <https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded>, acc: Jan 2024

¹⁰Please note that the most frequent top n words of a text do not necessarily contain domain content words

Sample set	cos sim.	pla. %	example of longer overlaps
Novels samples			
Jane_8 & Tenant_1	.95	01L, 01R	"would not be"
HenryVII_Pucelle & HenryVI2_Cade	.78	00L, 00R	-
Novels & GPT			
Jane_8 & Jane_8_gpt	.98	36L, 39R	"the swelling spring of pure, full, fervid eloquence? Such was the characteristic of Helen's discourse on that"
Tenant_1 & Tenant_1_gpt	.98	31L, 36R	"she is known to have entered the neighbourhood early last week, she did not make her appearance at church on Sunday; and she - Eliza, that is - will"
Webis Plagiarism samples			
3895-orig & 3895-para	.93	30L, 29R	"for me to impugn their honesty if"
4099-orig & 4099-para	.96	45L, 49R	"causes of discontent than they would naturally have independent of this circumstance"
4325-orig & 4325-para	.97	62L, 66R	"is difficult for me to realize the simple fact that she was niece to an uncle"
4475-orig & 4475-para	.82	02L, 04R	" frequently mistaken for"
7804-orig & 7804-para	.97	15L, 13R	"like the bottom of a well with"
Webis Plagiarism & GPT			
3895-original & 3895_gpt	.95	21L, 25R	"detected by one o f the three judges in the ring"
4099-original & 4099_gpt	.93	26L, 32R	"since the earliest settlement o f the country, would"
4325-original & 4325_gpt	.93	34L, 41R	"impression is that the child was asked to describe the vision more minutely"
4475-original & 4475_gpt	.97	38L, 46R	"This species is one of the most graceful birds"
7804-original & 7804_gpt	.94	25L, 34R	"from Shih-tien delivered his official dispatch at"
Essays & GPT (DE)			
text1 & text1_gpt	.97	62L, 67R	"umfasst jeden noch so kleinen intertextuellen Bezug im Text, was bedeuten würde, dass Intertextualität eine zentrale Eigenschaft von Texten ist."
text2.1 & text2.1_gpt	.93	46L, 44R	"indem sie zugibt, dass sie bei polnischen Gospelsongs eher an eine kulturelle Aneignung von schwarzer Kultur in den USA"
text2.2 & text2.2_gpt (5)	.97	41L, 47R	"die Ansichten zu diesem Thema, die Hand in Hand mit den Wahrnehmungen gehen, erarbeitet."
text3 & text3_gpt	.94	18L, 22R	"ein Zielllexikon mit manuell normalisierten Wort f ormen"
text4 & text4_gpt	.90	35L, 35R	"Ursprung, aber alles Neue in Natur und Kultur kann als Ergebnis der Schöpf ung durch Übernatürliches"
Almeida Revisada (PR)			
Matthew & Mark	.99	21L, 27R	"em verdade vos digo que de modo algum perderá a sua recompensa."
Luke & Mark	.99	16L, 18R	"Ora, para que saibais que o F ilho do homem tem sobre a terra autoridade para perdoar pecados (disse ao paralítico)"

Table 2: Overview o f style similarities and plagiarism “overlap” in our samples: style similarity represents the cosine similarities between the distribution o f the most f requent 100 words o f two texts; plagiarism (in %) represents the token-wise overlap with respect to all f ive-word-windows that overlap between two texts;

of the plagiarism percentage detected (towards the left and towards the right-hand sided texts), but keep them in the table separately to not lose any in formation (see Tab. 2).

We find a moderate positive correlation coefficient (c.f. Pearson, 1895) of 0.52 between style

similarities and plagiarism overlap in our samples (excluding couples with ChatGPT-generated texts). For the texts coupled with ChatGPT texts only, this value is 0.32 (weakly positive). This is especially attributed to the fact that ChatGPT naturally does not use a specific style. Instead it might makes

heavily use of the texts input's style. The correlation of all texts amounts to a value of 0.5 indicating still a moderate positive correlation.

4.2 Qualitative correlation measured

Novels: The novels samples show very little to no plagiarism across both sets of texts, while both reveal similarities in the use of style with a maximum cosine similarity of up to .95. The novels especially show the case where the domain vocabulary obviously is very different in both texts. We find very strong style overlap, but only very little plagiarism.

Our novels samples, compared with their ChatGPT re-phrased version show very similar styles with a cosine similarity of 98%. This very identical style could hint to the fact that ChatGPT is not very creative in formulating its own wording and style.

Webis Plagiarism data: Plagiarism is defined by verbatim copying which also goes strongly together with a high stylistic similarity. The Webis data set is the most interesting one for us, because it ensures the domain overlaps and helps us to make predictions on the texts' style. Leaving aside text 4475 with a plagiarism percentage close to zero, the results show a plagiarism percentage ranging between 13% and 66%. In text 4325, we can find the highest result of detected plagiarism of 66% while simultaneously showing very similar style. Text 4475 does not show a meaningful plagiarism ratio, but it also ships with a significantly lower style similarity. The results generally show the correlation of the style similarity and the plagiarism detected: The more similar the writing style of two texts compared, the higher the percentage of plagiarism overlap can be. In comparison with the Webis plagiarism data, the GPT-rephrased versions show less plagiarism detected ranging relatively close between 21% and 46%. This is on par with the study by [Gao et al. \(2023\)](#) where the authors found that GPT-produced texts are less likely to be plagiarised. We still also observe a narrow range in the use of style, ranging between .93 and .97. Looking at samples, we find that GPT typically replaces content words with similar ones, but the overall sentence structure stays rather similar. We find that, although less obvious, the same style-plagiarism correlation is visible.

Essays & GPT: The essays also show a high ratio of plagiarism overlap and a very strong correlation with the relating style similarities. The observations are comparable with those from the Webis Plagiarism & GPT. Again, very similar style

can be owed to the fact that ChatGPT does not do a good job in rephrasing sentences, it simply replaces words and phrases.

Almeida Gospels: The results show an identical cosine similarity of .99 while the detected plagiarism is between 16% and 27%, higher in the Matthew & Marc comparison than in the Luke & Marc comparison. These samples are possibly comparable with the Webis Plagiarism & GPT samples, which also show high stylistic similarity while also showing a meaningful plagiarism overlap.

5 Conclusion

We showed that there is a moderate correlation between the text samples coming from the English literature period, the Webis Plagiarism corpus together with their paraphrased versions and the books of the Brazilian Bible translation. We carefully selected the features that we utilize to measure style similarities considering function words distribution that do reliably represent style characteristics. In future work, we will fine-tune the procedure and have a closer look at how different register ranges affect our style similarities, and how pruning the most frequent n words (depending on the language) affects these correlation. We also found lower correlation between the style employed and the plagiarism detected in the texts re-phrased by ChatGPT. Which leads us to the conclusion that it copies the author's style, especially because it can sample style from pre-existing texts for German only with some effort.

References

- Muna AlSallal, Rahat Iqbal, Vasile Palade, Saad Amin, and Victor Chang. 2019. An integrated approach for intrinsic plagiarism detection. *Future Generation Computer Systems*, 96:700–712.
- Faisal Alvi, Mark Stevenson, and Paul Clough. 2021. Paraphrase type identification for plagiarism detection using contexts and word embeddings. *International Journal of Educational Technology in Higher Education*, 18(1):42.
- Salha Mohammed Alzahrani and Naomie Salim. 2008. Plagiarism detection in arabic scripts using fuzzy information retrieval. In *Student Conf. Res. Develop., Johor Bahru, Malaysia*, pages 281–285.
- Steven Burrows, Martin Potthast, and Benno Stein. 2013. [Paraphrase Acquisition via Crowdsourcing and Machine Learning](#). *Transactions on Intelligent Systems and Technology (ACM TIST)*, 4(3):43:1–43:21.

- Hugh Craig and Arthur F Kinney. 2009. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press.
- Fred J Damerau. 1975. The use of function word frequencies as indicators of style. *Computers and the Humanities*, pages 271–280.
- Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. *Stylometry with R: A Package for Computational Text Analysis*. *The R Journal*, 8(1):107–121.
- Sven Meyer zu Eissen and Benno Stein. 2006. Intrinsic plagiarism detection. In *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings 28*, pages 565–569. Springer.
- Stefan Evert, Fotis Jannidis, Friedrich Michael Dimpel, Christof Schöch, Steffen Pielström, Thorsten Vitt, Isabella Reger, Andreas Büttner, and Thomas Proisl. 2016. "delta" in der stilometrischen autorschaftsatriebution. In *DHd*.
- Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, 52(6):1–42.
- Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, 6(1):75.
- Douglas Wilhelm Harder. 2022. Plagiarism in the gospels. Acc: Jan 2024 <https://www.dwharder.org/plagiarism-in-the-gospels>.
- David L Hoover. 2003. Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3):261–286.
- Ethan Hunt, Ritvik Janamsetty, Chanana Kinares, Chanel Koh, Alexis Sanchez, Felix Zhan, Murat Ozdemir, Shabnam Waseem, Osman Yolcu, Binay Dahal, et al. 2019. Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104. IEEE.
- Stefan Jänicke, Annette Geßner, Marco Büchler, and Gerek Scheuermann. 2014. Visualizations for text re-use. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 59–70. IEEE.
- John V. Nance. 2017. “we, john cade”: Shakespeare, marlowe, and the authorship of 4.2.33–189 2 henry vi. *Shakespeare*, 13(1):30–51.
- Gabriel Oberreuter and Juan D Velásquez. 2013. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications*, 40(9):3756–3763.
- OpenAI. 2024. Chatgpt. Acc: Jan 2024 <https://chat.openai.com>.
- Kazuaki OTA. 2023. Was marlowe shakespeare’s collaborator?: Computational stylometry and the authorship of the three parts of henry vi.
- Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.
- Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005.
- Lukas Zwilling and Maria Berger. 2024. Chatgpt does not speak style!

Authorship attribution in translated texts: a stylometric approach to translator style

Ana Pagano

UFMG / Brazil

anapagano.ufmg@gmail.com

Carlos Perini

UFMG / Brazil

perini@ufmg.br

Evandro Cunha

UFMG / Brazil

cunhae@ufmg.br

Adriana Pagano

UFMG / Brazil

apagano@ufmg.br

Abstract

This paper presents an exploratory study of stylometry for authorship attribution in translated texts based on characteristics of translator style. The study aimed to assess to what extent stylometric methods were successful in attributing a translated text to a particular translator (classification task) and clustering translated texts by the same translator (clustering task). To that end, a corpus of eighteen texts was compiled, including novels and short stories, originally written in English and translated into Brazilian Portuguese. Six different translators were included, each of them having authored three translated texts. The classification task was performed using a Python script for three stylometric methods: Mendenhall's Characteristic Curves of Composition; Kilgarriff's Chi-Squared Method; and Burrows' Delta Method. The clustering task was carried out in the R programming environment using various parameters available in the *stylo* package. The results were partially successful, with authorship of some of the translated texts correctly attributed to their translators in both the classification and clustering tasks. The study also found that texts by translators contemporary to each other were clustered as more similar to one another and that some translated texts were clustered due to being translations of the same original text, regardless of being authored by different translators. Our findings are in line with previous stylometric studies of translated texts, which point to the original text's style as bearing an impact on both classification and clustering tasks of translated texts.

1 Introduction

Within digital humanities, stylometry has been pursued for a variety of tasks, among them author verification, plagiarism detection, author profiling or characterization, and detection of stylistic inconsistencies (Stamatatos, 2009). Additionally, disciplines such as forensic linguistics and translation

studies have also resorted to stylometric methods: the former in order to explore stylistic aspects of texts that can support correct authorship attribution for forensic purposes, and the latter to investigate characteristics of the so-called "translator style" (Saldanha, 2011).

With regard to stylometric approaches to translator detection, Rybicki (2012, 2013) carried out studies of translated texts, having obtained, at least for the language pairs he considered, partially successful and inconclusive results regarding the potential of stylometry in authorship attribution to a translated text. For the English-Brazilian Portuguese language pair, no studies, to the best of our knowledge, have reported results of stylometric techniques for the investigation of translator style. This paper seeks to fill this gap by reporting on a stylometric analysis of translated literary texts. To this end, it draws on a set of translations published in Brazil by translators who were very actively engaged in translation activities in two different historical periods: between 1930 and 1955 and between 1990 and 2015. Considering that choices made by translators reveal ways in which they see their role as cultural mediators, there is a characteristic translation style of the time, which manifests itself in traces of each translator style (Baker, 2000).

The aim of this study is thus to contribute to digital humanities and to the disciplinary field of translation studies by (i) exploring the concept of translator style from the perspective of stylometry and (ii) inquiring into how author attribution based on stylometry can be applied to translated texts in the English-Brazilian Portuguese language pair.

The remainder of this paper is organized as follows. Section 2 presents a review of the main concepts that guide our analysis. Section 3 presents the methodology for corpus compilation and analysis. In Section 4 we report results obtained. Section 5 discusses our results with respect to the available

literature. Finally, Section 6 draws conclusions from our study and presents the limitations of our work as well as perspectives for further research. Sources for our corpus and the bibliography supporting our study are provided in the References.

2 Review

2.1 Translator style

According to Baker (2000), a translator's style is a "*a kind of thumb-print*" that can be mapped based on non-linguistic and linguistic characteristics. Non-linguistic characteristics are the choices made by a translator regarding the type of text and the authors a translator decides or agrees to translate. Linguistic characteristics are recurring grammatical and lexical choices, which may or may not be conscious on the part of the translator.

Saldanha (2011) further complements Baker's definition by adding that style is a set of recurring patterns in different texts translated by the same translator, which occur regardless of the style of the original text. Saldanha (2014) also highlights the connection between the concept of translator style and that of "audience design" (Mason, 2000), which posits that the way in which translators see their role as cultural mediators and represent their readers has an impact on their translation choices, which contribute to characterize their style.

Both Baker (2000) and Saldanha (2011, 2014) use corpus linguistics concepts, such as word frequency, collocations and keywords in context, to study translator style.

2.2 Stylometry

Stylometry is an established field that explores the style of texts from a quantitative perspective, generally through computational methods. The assumption is that every author has a particular and consistent way of writing that can be recognized based on their use of lexical words (nouns, verbs, adjectives, adverbs), grammatical words (articles, prepositions, conjunctions), length of sentences used, use of punctuation marks, among other features (Stamatatos, 2009).

The first quantitative studies of style date back to the 19th and early 20th centuries (Mendenhall, 1887; Yule, 1939, 1944; Zipf, 1932). In subsequent decades, several studies were developed with a view to ascertaining which textual characteristics were most productive in author attribution of

a work within the scope of what we today call stylometry (Holmes, 1994, 1998).

With developments in computer processing, stylometry (or computational stylometry) began to be approached by natural language processing (NLP) as a form of natural language understanding, with a view to extracting both knowledge and metaknowledge about texts, the latter referring to knowledge about the author of the text as a kind of psychological and sociological profile (Daelemans, 2013).

According to Laramée (2018), the lexicon that a person uses is a particular characteristic of each human being: some authors make use of a more limited lexicon than others. A writer, especially a literary one, is expected to have a more extensive and fine-grained vocabulary. However, a renowned writer such as Ernest Hemingway is frequently referred to as an author who makes use of a relatively small number of unique words when writing (Rice, 2016). This does not implicate lesser value in terms of his writing; it is deemed a matter of style.

In stylometric studies, unlike in NLP approaches and disciplines like corpus linguistics, function words, such as articles, prepositions and conjunctions, are particularly important. Stamatatos (2009) presents a review of stylometric methods and highlights the use of function words as being important as they are "used in a largely unconscious manner by the authors, and they are topic-independent" (p. 540). For stylometric analyses, function words within a corpus of works by the same author tend to vary less than lexical words.

2.3 Author attribution

Within the scope of computational stylometry, Rybicki (2012) suggests that author attribution implicates a machine learning approach for a classification task. In this process,

the traceable differences between texts in a corpus are first used to produce a set of rules – a classifier – for discriminating authorial "uniqueness". The second step is to use the trained classifier to assign other texts samples to the authorial classes established by the classifier. (Rybicki, 2012)

For classification tasks in stylometric studies, three well established methods are explored by Laramée (2018), briefly described in the three following subsections.

2.3.1 Mendenhall's Characteristic Curves of Composition

Mendenhall (1887) proposed characterizing an author's style by a curve that expresses the distribution of the length of the words used. This is accounted for by the idea that in an author's writing, certain personal characteristics become recurrent throughout their career and these have to do with the frequency of use of short and long words. Thus, a person's writing can be characterized by counting the size of the words they use and how often this size varies.

Mendenhall compared several authors from the same historical period, counted the number of characters in each word they used and calculated the number of words with the same length. He started by counting the first 1000 words and then took random excerpts from their works. He observed that there was a pattern in word length that was repeated across different samples from the same author. Mendenhall asserted that curves generated from word sets extracted from various works by the same author will closely resemble the characteristic curve of this author.

In his proposal, a set of n words is taken from a text and, from there, a graph is created showing the frequency and size of the words in a curve.

2.3.2 Kilgarriff's Chi-Squared Method

Kilgarriff proposed using the chi-squared statistic to measure the "distance" between the lexicon used in two sets of texts. Unlike Mendenhall, whose method relied on word length distribution, Kilgarriff relies on word frequency distribution.

His method requires two corpora and selecting the n most common words in the larger corpus. He stated that the number of words to be considered is a matter not yet solved, the literature pointing to numbers between 100 and 1,000 of the most common words.

In Kilgarriff's method, the smaller the chi-squared value obtained, the more similar two texts will be and the more certain we can be that both texts were written by the same author. The assumption is that word usage patterns and a person's lexicon are very constant in an author's career.

2.3.3 Burrows' Delta Method

Burrows proposed a statistic delta value to express the distance between a text to which authorship must be attributed and a set of other texts whose authorship is already known within a corpus. Un-

like Mendenhall and Kilgarriff, Burrows focuses on function word frequency and his delta is calculated by comparing the relative frequencies of function words.

The method receives this name because it measures the difference between a sample text of an author to be discovered and the other works compiled in a corpus by a known author, generating a delta value.

From this delta value, it is possible to rank candidate authors of the sample text in terms of probability of authorship. The author that is most similar will be the one whose delta has the lowest value.

2.4 Stylometry and translation

Rybicki (2012) reports a study in which he seeks to verify whether stylometric techniques are efficient to correctly attribute an author to a translated text, that is, whether translations done by the same translator are correctly identified as having the same author. Rybicki analyzes different corpora of translations of novels in two different language pairs (English-Polish; English-French). His results show that, regardless of the language pair, stylometric techniques group translated novels according to the author of the original texts instead of the translator. Instead of Burrows' Delta, Rybicki suggests using his Zeta and Iota methods, which are based, respectively, on words with intermediate frequency and the least frequent or most singular words used by a translator. Rybicki (2013) complements the results of his studies in Rybicki (2012). In studies of translated texts, stylometric methods can more successfully detect the author of original texts rather than the translator. As Rybicki highlights, the style of translated texts of the same original seems to bear similarities despite the fact that the translated texts were authored by different translators.

3 Methodology

3.1 Corpus compilation

The corpus used in our study is monolingual and comprises 18 texts translated into Brazilian Portuguese, authored by 6 Brazilian translators, each translator being the author of 3 texts.

The criteria for compiling the corpus were: (i) texts should be translations of novels or short stories originally published in English; (ii) texts should be first translations and/or retranslations into Brazilian Portuguese published in Brazil between 1930-1955 and 1990-2015; and (iii) texts

Translator name	Title and publication year of original text	Title and publication year of translated text	Label assigned	# Tokens
Monteiro Lobato	The adventures of Huckleberry Finn (1884)	As aventuras de Huck (1934)	Lobato1	82,355
	A farewell to arms (1929)	Adeus às armas (1942)	Lobato2	77,201
	The thin man (1934)	A ceia dos acusados (1936)	Lobato3	47,031
Érico Veríssimo	Of mice and men (1937)	Ratos e Homens (1940)	Verissimo1	28,470
	Point Counterpoint (1928)	Contraponto (1943)	Verissimo2	183,001
	They kill horses, don't they? (1935)	Mas não se mata cavalo? (1947)	Verissimo3	25,285
Mário Quintana	Lorde Jim (1900)	Lorde Jim (1939)	Quintana1	97,840
	God's men (1951)	Debaixo do céu (1955)	Quintana2	151,883
	Tales from Shakespeare (1807)	Contos de Shakespeare (1943)	Quintana3	83,690
Julieta Cupertino	Lorde Jim (1900)	Lorde Jim (2002)	Cupertino1	251,661
	The end of the tether (1902)	O fim das forças (2000)	Cupertino2	51,388
	Bliss and other short stories (1920)	Felicidade e outros contos (1991)	Cupertino3	34,443
Rubens Figueiredo	I married a dead man (1948)	Casei-me com um morto (1996)	Figueiredo1	64,405
	The circle (2013)	O círculo (2013)	Figueiredo2	147,033
	The thin man (1934)	O homem magro (2002)	Figueiredo3	60,670
Renato Pompeu	They kill horses, don't they? (1935)	A noite dos desesperados (2000)	Pompeu1	50,919
	No pockets in a shroud (1937)	Mortalha não tem bolso (2002)	Pompeu2	53,599
	The friends of the friends (1896); The country of the blind (1911)	Os amigos dos amigos (2004); Em terra de cego (2004)	Pompeu3	20,545
Total				1,511,419

Table 1: Corpus composition and token distribution

should make up three sets of translations by the same translator.

Monteiro Lobato, Érico Veríssimo and Mário Quintana fulfilled our criteria for the period from 1930 to 1955, whereas Rubens Figueiredo, Julieta Cupertino and Renato Pompeu met our criteria for the period from 1990 to 2015.

Two works translated by each translator were used to characterize their style in the training set, and a third one was used as a testing set for the classification task to verify whether the techniques used allowed correctly inferring author attribution based on the degree of similarity of each work in the testing set with the style of each translator as characterized on the basis of the training set. Table 1 shows the texts compiled in our corpus, their label and number of tokens¹. The whole corpus totaled 1,511,419 tokens.

Texts were converted from their epub or pdf editions to UTF-8 encoded txt files. File preparation procedures were performed as follows: (i) assigning file name labels; (ii) clearing metadata (author name, title, title page, pagination) and additional metatext in the text; (iii) clearing symbols, spaces and blank lines. Due to pending copyright clearance for some of the texts, access to the corpus is available for research purposes upon request.

3.2 Stylometric analysis

The study comprised two stages. In the first one, a classification task was performed using a script the Python programming language. The analysis was

¹Texts with lower number of tokens were used for testing and appear in the 'Label assigned' column with number '3'.

based on the methodology presented by Laramée (2021) and included the three methods introduced in Section 2.3: (i) Mendenhall's Curves; (ii) Kilgarriff's Chi-Squared; and (iii) Burrows' Delta.

In the second stage, a clustering task was conducted using the stylo package in R (Eder et al., 2016). This package enables the customization of text grouping parameters, including language selection, unit consideration (token or character), n-gram size, and the establishment of minimum and maximum frequency of words. Additionally, choices such as the inclusion or exclusion of pronouns can be made based on the selected language. Clustering was executed for each parameter outlined in Table 6. In our study, texts were clustered using the various parameters and results compared as reported in the Results section.

4 Results

4.1 Classification task

Our first analysis explored Mendenhall's Curves of Composition. To assess which curves were closest to one another, a confusion matrix was generated as seen in Table 2, where, for each line, the lower the value (highlighted in red), the greater the similarity between two texts.

	Lobato	Verissimo	Quintana	Cupertino	Figueiredo	Pompeu
Lobato (test)	0.564330	0.744281	0.448944	0.517613	0.345221	0.491614
Verissimo (test)	0.419614	0.674223	0.657576	0.718682	0.585002	0.400529
Quintana (test)	0.525755	0.400119	0.449070	0.517624	0.735718	1.012819
Cupertino (test)	0.324554	0.581069	0.404144	0.410844	0.262137	0.499184
Figueiredo (test)	0.916860	1.080957	0.754031	0.833025	0.498228	0.730975
Pompeu (test)	0.526212	0.715731	0.416241	0.471150	0.218603	0.627120

Table 2: Confusion matrix for Mendenhall's Curves of Composition method results

A heatmap was outputted as shown in Figure 1, where the closer the result to the blue shades of color, the greater the similarity.

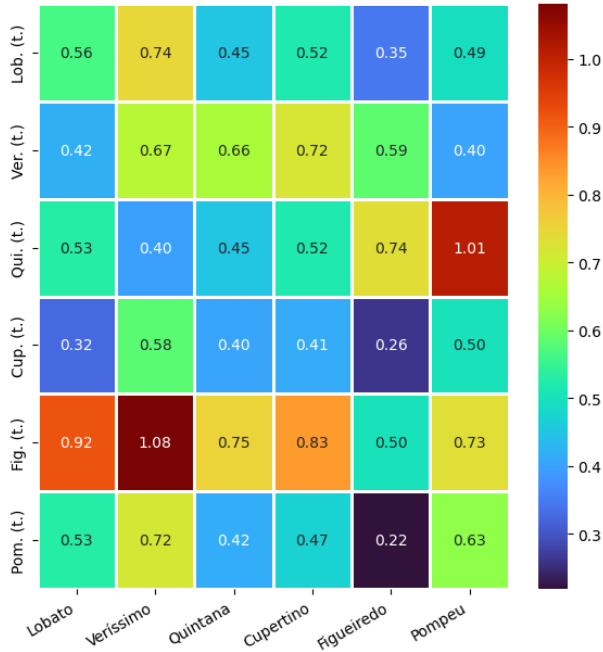


Figure 1: Heatmap of confusion matrix for Mendhall's Curves Method results. Author names on the y axis shortened as: Lobato (test) as *Lob. (t.)*, Veríssimo (test) as *Ver. (t.)*, Quintana (test) as *Qui. (t.)*, Cupertino (test) as *Cup. (t.)*, Figueiredo (test) as *Fig. (t.)* and Pompeu (test) as *Pom. (t.)*.

As we can see, the method correctly attributed authorship to Figueiredo's text (test). The method attributed Cupertino (test) and Pompeu (test) to Figueiredo. This result could be accounted for by the fact that Cupertino, Pompeu and Figueiredo are contemporary authors (1990-2015) and hence may have more similar styles. The method also attributed Quintana (test) to Veríssimo, both authors being contemporary (1930-1955). Moreover, there are low values (higher similarity) for texts translated by translators of the same original novel, namely Érico Veríssimo and Renato Pompeu (*They shoot horses, don't they?*) and Monteiro Lobato and Rubens Figueiredo (*The thin man*).

Our second analysis implemented Kilgarriff's Chi-Squared method, with three parameters for the number of most frequent words: 500, 1000 and 5000. In this method, the lower the chi-squared result, the greater the similarity between two texts, as highlighted in Table 3.

Among all the samples, the 500-word sample was the most successful². The method yielded

²With the parameter of 1000 and 5000 words, the method

	Lobato	Veríssimo	Quintana	Cupertino	Figueiredo	Pompeu
Lobato (test)	18601.03330	24766.78745	26889.55810	29758.62721	24399.01371	16960.88682
Veríssimo (test)	13144.97659	14865.49822	15389.20671	15725.75842	13834.01081	9220.405519
Quintana (test)	21592.12369	21491.86867	18402.43422	23239.12784	26385.17238	26627.86092
Cupertino (test)	14362.48120	15754.84795	16738.18427	18026.77286	14374.43532	12695.68699
Figueiredo (test)	19544.93223	19544.93223	29465.92648	32305.87924	22741.30033	14912.87123
Pompeu (test)	7218.145580	5743.387775	5766.852190	4557.430660	5470.142603	7324.708148

Table 3: Confusion matrix for Kilgarriff's Chi-Squared method results

greater proximity between texts by contemporary authors: Figueiredo, Pompeu and Cupertino (1990-2015), and Lobato, Quintana and Veríssimo (1930-1955), and texts translated by translators (Veríssimo and Pompeu) who translated the same original text (*They shoot horses, don't they?*).

Rank	Word	# Ocorrences
1	de	47,360
2	a	38,361
3	que	36,772
4	o	35,161
5	e	35,061
6	não	20,244
7	um	18,970
8	para	15,619
9	uma	13,924
10	se	13,679
11	com	12,732
12	ele	12,312
13	do	11,822
14	em	11,214
15	os	9,831
16	da	9,321
17	eu	8,156
18	é (gram.)	7,946
19	por	7,881
20	como	7,524
21	mas	7,492
22	no	7,326
23	na	6,907
24	as	6,750
25	era (gram.)	6,503
26	sua	5,986
27	mais	5,936
28	ela	5,806
29	você	5,566
30	seu	5,280

Table 4: Thirty most common words from all sets of texts sorted by decreasing frequency

Our third analysis explored John Burrows' Delta method. The method first extracts the thirty most frequent words in all sets, as seen in Table 4.

Table 4 shows that the most frequent words common to all sets of texts are function words, that is, pronouns, conjunctions, prepositions and articles, did not correctly attribute authorship to any of the translated texts.

including contracted forms in Portuguese (preposition plus article). Inflections of the verb "ser" (*to be*) are also part of the list.

The results of Burrows' method are displayed in Table 5. In this method, the lower the Delta score, the more similar the texts under comparison. As we can see highlighted in red, the method correctly detected the authorship of the texts translated by Quintana and Lobato. This method also yielded proximity between contemporary authors: Cupertino, Pompeu and Figueiredo, and Lobato and Quintana. Again, proximity was yielded between texts translated by Veríssimo and Pompeu, which are translations of the same original novel (*They shoot horses, don't they?*).

	Lobato	Veríssimo	Quintana	Cupertino	Figueiredo	Pompeu
Lobato (test)	1.105159	1.447027	1.540591	1.688335	1.564369	1.141621
Veríssimo (test)	0.944607	1.331801	1.254815	1.500599	1.209773	0.662313
Quintana (test)	1.631374	1.269202	1.227610	1.582967	1.675187	2.073649
Cupertino (test)	1.074294	1.172532	1.209249	1.218953	1.067089	1.289313
Figueiredo (test)	1.192101	1.682147	1.628778	1.704506	1.448851	0.827124
Pompeu (test)	1.438636	1.025698	1.174371	1.204885	1.094245	1.762203

Table 5: Confusion matrix for Burrows' Delta method results

4.2 Clustering Task

Results of the clustering task corroborate what was pointed out by Rybicki (2012, 2013). This can be seen in Figure 2, for instance, which shows a dendrogram for the results obtained for the Delta Classic distance parameter. Some of the translated texts were correctly grouped as having been authored by the same translator (Cupertino1 and Cupertino2; Lobato1 and Lobato2). Interestingly, the two translated texts authored by Cupertino were written by the same author, Joseph Conrad. In this case, clustering may be due to both translator style and original text style.

All clustering methods are shown on Table 6. For each parameter in the first column, the second column shows clustered texts which are authored by the same translator, whereas the third column shows clustered texts which are translations of the same original text. As we can see, all parameters clustered Lobato's and Cupertino's translated texts while some of them clustered Quintana's translated texts. Lobato is the only translator whose three translated texts were clustered by some of the parameters.

In addition, all parameters clustered translated texts of the same original text: Veríssimo's and Pompeu's translations of *They shoot horses, don't they?* and Quintana's and Cupertino's translations

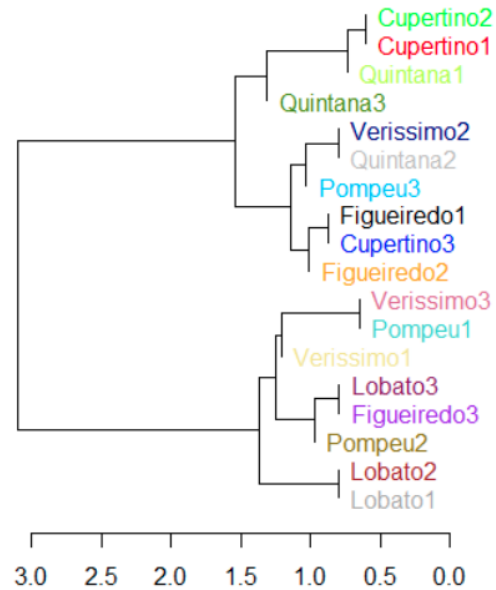


Figure 2: Classic Delta Distance Dendrogram

of *Lord Jim*, with some of them clustering Lobato's and Figueiredo's translations of *The thin man* as well.

5 Discussion

Our classification results evidence partial success, correctly classifying only one translator in the first two methods (Mendenhall and Kilgariff), and two in Burrows'. The methods showed greater proximity between texts by translators who were contemporary, particularly within the period between 1930 and 1955, which is considered a period when translators had a more similar style (Laviosa et al., 2017).

Additionally, in all analyses, the methods associated at least one pair of texts by different translators of the same original text, which points to the impact of the original text on translation style.

Our clustering results, using different parameters, showed strong tendencies towards grouping some texts based on two aspects: (i) authorship attribution to texts by the same translator and (ii) authorship attribution to translated texts of the same original text. In other words, texts translated by the same translator cluster; so do translations of the same original by different translators. This corroborates what was pointed out by Rybicki (2012, 2013) regarding stylometric analyzes of translated texts: the algorithms are not always successful in correctly grouping texts by the same translator and there is an impact of the original text on the group-

Parameter	Clustering by translator	Clustering by original text
<i>Classic Delta</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Cosine Delta</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2; Quintana3 and Quintana2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Eder Delta</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Eder Simple Delta</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Entropy</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Manhattan</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Canberra</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Euclidean Distance</i>	Lobato1, Lobato2 and Lobato3; Cupertino1 and Cupertino2.	Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>Cosine Distance</i>	Lobato1, Lobato2 and Lobato3; Cupertino1 and Cupertino2; Quintana2 and Quintana3.	Verissimo3 and Pompeu1; Quintana1 and Cupertino1.
<i>MinMax</i>	Lobato1 and Lobato2; Cupertino1 and Cupertino2; Quintana2 and Quintana3.	Lobato3 and Figueiredo3; Verissimo3 and Pompeu1; Quintana1 and Cupertino1.

Table 6: Texts clustered according to each parameter applied

ing of translated texts.

Among the texts translated by the same translator that were most successfully grouped are the texts translated by Monteiro Lobato (the only translator whose three translated works were grouped by some of the parameters), followed by Mário Quintana and Julieta Cupertino. These results may indicate more marked style traits in these translators than in Rubens Figueiredo, Renato Pompeu and Érico Veríssimo.

The results obtained partially corroborate those obtained by Laviosa et al. (2017). The author carried out a manual analysis of characteristics noted in samples of the corpora, which grouped the translators from the periods 1930-1955 and the translators from the periods 1990-2015 into two distinct classes. The stylometric analysis carried out in our study grouped texts by the same translator, which corroborates Laviosa et al. (2017), but it also grouped translations of the same original made by translators at different times.

6 Conclusions

This study contributed to digital humanities and the disciplinary field of translation studies by pursuing research that investigated the concept of translator style from a stylometric perspective. To the best of our knowledge, the study is the first that explored this topic for the English-Brazilian Portuguese language pair.

Classification and clustering tasks were performed for authorship attribution of translated texts and the results confirmed what was observed by other stylometric studies of translated texts, with emphasis on the impact of the original text on the grouping results. We verified that stylometric methods are partially successful, both in a classification task for author attribution of a translated text and in a task of clustering texts by translator.

The main limitations of our study are: (i) variation in our corpus in terms of subgenres within the narrative genre – that is, different types of novels and short stories were used; and (ii) the predominance of male translators. These limitations

have to do with the availability and access to texts translated into Brazilian Portuguese fulfilling the inclusion criteria. Perspectives for further research include pursuing Rybicki (2012)'s suggestion to use his Zeta and Iota methods to verify whether words with intermediate frequency and less frequent or more singular words used by a translator could be indicators with more potential for author attribution of a translation.

7 Acknowledgments

The authors would like to thank three anonymous reviewers for their valuable comments. Adriana S. Pagano holds a research productivity grant awarded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (Processo CNPq 313103/2021-6).

References

- Mona Baker. 2000. [Towards a methodology for investigating the style of a literary translator](#). *Target*, 12(2):241–266.
- Pearl Buck. 1955. *Debaixo do Céu*. Livros do Brasil, Lisboa. Transl. by Mário Quintana. *Corpus Reference*.
- Joseph Conrad. 1971. *Lorde Jim*. Globo, Porto Alegre. Transl. by Mário Quintana. *Corpus Reference*.
- Joseph Conrad. 2000. *O Fim das Forças*. Revan, Rio de Janeiro. Transl. by Julieta Cupertino. *Corpus Reference*.
- Joseph Conrad. 2002. *Lord Jim; um romance*. Revan, Rio de Janeiro. Transl. by Julieta Cupertino. *Corpus Reference*.
- Walter Daelemans. 2013. [Explanation in computational stylometry](#). In *Computational Linguistics and Intelligent Text Processing*, pages 451–462, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. [Stylometry with r: A package for computational text analysis](#). *R J.*, 8(1):107.
- Dave Eggers. 2013. *O Círculo*. Companhia das Letras, São Paulo. Transl. by Rubens Figueiredo. *Corpus Reference*.
- Dashiell Hammett. 1984. *A Ceia dos Acusados*. Abril, São Paulo. Transl. by Monteiro Lobato. *Corpus Reference*.
- Dashiell Hammett. 2002. *O Homem Magro*. Companhia das Letras, São Paulo. Transl. by Rubens Figueiredo. *Corpus Reference*.
- Ernest Hemingway. 2013. *Adeus às Armas*. Bertrand Brasil, Rio de Janeiro. Transl. by Monteiro Lobato. *Corpus Reference*.
- David I. Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.
- David I. Holmes. 1998. [The Evolution of Stylometry in Humanities Scholarship](#). *Literary and Linguistic Computing*, 13(3):111–117.
- Aldous Huxley. 2014. *Contraponto*, 7th edition. Globo, São Paulo. Transl. by Érico Veríssimo. *Corpus Reference*.
- William Irish. 1996. *Casei-me com um Morto*. Companhia das Letras, São Paulo. Transl. by Rubens Figueiredo. *Corpus Reference*.
- Henry James. 2004. Os amigos dos amigos. In Italo Calvino, editor, *Contos Fantásticos do Século XIX*, pages 600–640. Companhia das Letras. Transl. by Renato Pompeu. *Corpus Reference*.
- Charles Lamb and Mary Lamb. 2013. *Contos de Shakespeare*, 8th edition. Globo, São Paulo. Transl. by Mário Quintana. *Corpus Reference*.
- François Dominic Laramée. 2021. [Introdução à estilometria com Python](#). *Programming Historian em português*, (1).
- François Dominic Laramée. 2018. [Introduction to stylometry with Python](#). *Programming Historian*, (7).
- Sara Laviosa, Adriana Pagano, Hannu Kemppanen, and Meng Ji. 2017. [A Contextual Approach to Translation Equivalence](#), pages 73–127. Springer Singapore, Singapore.
- Katherine Mansfield. 2000. *Felicidade e Outros Contos*, 3rd edition. Revan, Rio de Janeiro. Transl. by Julieta Cupertino. *Corpus Reference*.
- Ian Mason. 2000. [Audience design in translating](#). *The Translator*, 6(1):1–22.
- Horace McCoy. 1982. *Mas Não Se Mata Cavalo?* Abril Cultural, São Paulo. Transl. by Érico Veríssimo. *Corpus Reference*.
- Horace McCoy. 2000. *A Noite dos Desesperados*. Sá Editora, São Paulo. Transl. by Renato Pompeu. *Corpus Reference*.
- Horace McCoy. 2002. *Mortalha Não Tem Bolso*. Sá Editora, São Paulo. Transl. by Renato Pompeu. *Corpus Reference*.
- Thomas C. Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214S):237–249.
- Justin Rice. 2016. What makes Hemingway Hemingway? A statistical analysis of the data behind Hemingway's style. *LitCharts*.

- Jan Rybicki. 2012. [The great mystery of the \(almost\) invisible translator: Stylometry in translation](#). In *Quantitative Methods in Corpus-Based Translation Studies*, page 231–248. John Benjamins Publishing Company, Amsterdam.
- Jan Rybicki. 2013. The translator's other invisibility: stylometry in translation. SLE 2013 Annual Meeting. Croatia, Split University.
- Gabriela Saldanha. 2011. [Translator style](#). *The Translator*, 17(1):25–50.
- Gabriela Saldanha. 2014. *Style in, and of, Translation*, chapter 7. John Wiley & Sons.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- John Steinbeck. *Ratos e Homens*. Livros do Brasil, Lisboa. Transl. by Érico Veríssimo. *Corpus Reference*.
- Mark Twain. 2005. *As Aventuras de Huck*. Companhia Editora Nacional, São Paulo. Transl. by Monteiro Lobato. *Corpus Reference*.
- H. G. Wells. 2004. Em terra de cego. In Italo Calvino, editor, *Contos Fantásticos do Século XIX*, pages 687–722. Companhia das Letras. Transl. by Renato Pompeu. *Corpus Reference*.
- George Udny Yule. 1939. [On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship](#). *Biometrika*, 30(3/4):363–390.
- George Udny Yule. 1944. *The Statistical Study of Literary Vocabulary*. CUP Archive.
- George Kingsley Zipf. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge/London.

Support Verb Constructions in Medieval Portuguese: Evidence from the CTA Corpus

M.^a Inês Bico^{id}, Esperança Carneira^{id} Jorge Baptista^{id}

Univ. Lisboa
Faculdade de Letras
Centro de Linguística
Lisboa, Portugal
mariainesb1@edu.ulisboa.pt
ecardeira@edu.ulisboa.pt

Univ. do Algarve
Faro, Portugal
INESC-ID Lisboa,
Portugal
jbaptis@ualg.pt

Fernando Baptista^{id}
ISCTE, Inst. Univ. Lisboa
INESC-ID Lisboa, Portugal
fernando.batista@iscte-iul.pt

Abstract

This paper conducts a systematic survey of Support Verb Constructions (SVC) in Medieval Portuguese using the Corpus de Textos Antigos (CTA) corpus. SVC involves verb-noun combinations, where the noun serves as the main predicative element, and the verb conveys grammatical values. Limited historical evidence exists for SVC in earlier stages of Portuguese, with previous studies lacking digital access and Natural Language Processing tools. This study utilizes a subset of the CTA, comprising around 0.5 million tokens, annotated for part-of-speech and lemmata. Out of 175 candidate support verbs in Modern Portuguese, 81 were identified in the corpus, totaling 27,645 occurrences. Manual analysis of concordances revealed a little more than 3,000 SVC instances and more than 900 predicative nouns (types), uncovering several linguistic phenomena. The paper delineates the adopted procedure and explores essential linguistic properties of identified SVC in the CTA corpus, emphasizing the significance of leveraging NLP tools for a comprehensive linguistic description of Old Portuguese.

1 Introduction

This paper presents a systematic survey of *support verb constructions* (SVC) from Old Portuguese (OP) and Middle Portuguese (MP) found in a corpus. Old Portuguese is an early stage of the Portuguese language that was written from the 12th century to the late 14th century or early 15th century, when Middle Portuguese starts. This period of the language goes until the mid-16th century (Castro, 2006).

Support verb constructions (Harris, 1964, 1991; Gross, 1981, 1996) are elementary sentences, consisting of verb-noun combinations where the main predicative element is the noun, called a *predicative noun* (*Npred*), and the verb, called a *support verb* (*Vsup*), has mostly a grammatical auxiliary

function, conveying the tense and person-number agreement values.

Next, there is a clear example of a SVC.¹

(1) [OP] *e nõ etendedes quanta ãJuria fazedes ao vosso criador* [mPT] *‘e não entendeis quanta injúria fazeis ao vosso criador’* [EN] *‘and you do not understand how much injury (= insults, offenses) you do to your creator.’* [HdE-A:142892]

In this instance extracted from the ms. A of *Horto do Esposo* (HdE-A), the noun *ãJuria* ‘injúria’ ‘injury’ is supported by the verb *fazedes* ‘fazeis’ ‘do-present-2nd-person-pl.’. The syntactic structure, as determined by the predicative noun, reveals a *dative* complement *ao vosso criador* ‘idem’ ‘to your creator’, along with the distributional constraints placed on the human subject. The dative type of complement can be further substantiated by examining other occurrences of the noun in the corpus.

(2) [OP] *lhe perdõ das ãJurias que lhe fezera* [mPT] *‘lhe perdoou das injúrias que lhe fizera’* [EN] *‘forgave him/her for the insults that had been done to him/her’* [HdE-A:140338]

The identification of the SVC involving this noun in Old Portuguese texts is facilitated not only by its persistence in Modern Portuguese (albeit with potentially nuanced differences in meaning) but also by its consistent display of the same syntactic construction and distributional constraints. However, in this context, particular em-

¹Conventions: In the examples, a color code indicates the language: *purple* for *Old Portuguese* (OP), *‘red’* for the *‘Modern Portuguese’* (mPT) gloss, and *‘blue’* for an *‘English’* gloss (EN); occasionally, *gray* will be used for *Modern Portuguese* (PT) example. For clarity, examples and glosses are also preceded by the code of each language. Targeted words, usually the support verb and the predicative noun are highlighted in bold typeface. The alphanumeric codes at the end of the examples identify the text and the token number where the verb occurred in the corpus. The list of texts in the corpus considered in this paper is presented in (3.1). In accordance with the transcription standards adopted for the corpus, words are presented in their original spelling, including the use of uppercase/lowercase letters as well as diacritical marks.

phasis should be placed on the observations put forth by Ranchhod (1999, p. 3) regarding the use of corpora in studying the earlier stages of language evolution:

When studies are conducted on earlier stages of the evolution of a language, the knowledge derived from it comes from what is evident in the texts of the period or periods under investigation. Nothing that is not present in the texts can be presumed; it is not possible to make judgments of acceptability about constructions and usages that are not attested there. Only the analysis of these data and theoretically grounded argumentation can fill the gap in these refutation criteria. (our translation)

2 Related Work

The study of support-verb constructions in old stages of Portuguese, namely Medieval Portuguese, is scarce. The two principal works are those of Chacoto (1996) and Ranchhod (1999), both studying the existence of these constructions in different texts. Though different corpora were used, the conclusions are similar: i) SVC are a patrimony of the Portuguese language, going as far back as the medieval period; ii) there are a lot of similarities between Medieval and Modern Portuguese. The more significant difference noted by both authors are the position of the constituents of SVC: in certain types of constructions, not possible today, Medieval Portuguese permitted the placement of the predicative noun before the support verb. For example, Ranchhod (1999) identifies this different word order in *Crónica Geral de Espanha de 1344* [*General Chronicle of Spain from 1344*].

(3) [OP] *depois que ambas amor ouvemos*
[mPT] *'depois que ambas amor houvemos'*
[EN] *'after we both love had'*

Chacoto (1996) studies SVC with the verb *fazer* 'to do', in *Demanda do Santo Graal* [*The Quest of the Holy Grail*]. The author distinguishes three uses of the verb: full or distributional verbs, causative operator-verbs e support-verb constructions. Following a brief description and characterization of these types of constructions, the focus turns to cases wherein exists nominalization ('*fazer pecado*' = '*pecar*' 'to do a sin' = 'to sin'), giving also examples for when this relation does not exist and what reasons may explain this.

Ranchhod (1999) in its turns studies SVC in fifteen chapters of *Crónica Geral de Espanha de 1344*, noting the high frequency of these constructions. Defining the SVC in an elementary phrase, the author also discusses SVC as nominal groups, nominalizations and autonomous predicative nouns. The most frequent verbs were found to be '*haver*' 'have' and '*fazer*' 'do', while the verbs '*ser*' and '*estar*' 'be' are not yet used in SVC.

More recently, Pante and Ortega (2009) and Pante (2012) studied SVC with the verb '*tomar*' 'take' in different medieval texts. Pante and Ortega (2009) uses two late 15th century texts (*Leal Conselheiro* [*Loyal Adviser*] and *Livro da Enseñança de bem cavalgar toda a sela* [*Book of the lesson on how to ride well every saddle*]) to shine light on the reasons why SVC are used: to allow the characterization of the predicative noun and to reduce the valence of the verbs. Pante (2012) studies this verb in *Horto do Esposo*, which is also present in our corpus. However, only instances in which it is possible to discern a case of nominalization were included, in order to identify the possible reasons for choosing a SVC instead of a distributional verb. Nine pairs of SVC and nominalizations were found, and results were similar to those already mentioned above.

3 Methods

This paper utilizes the part-of-speech and lemmatized partition of the CTA corpus to conduct a systematic survey of Medieval Portuguese SVC. In this Section, we elaborate on the methodology employed.

3.1 The CTA Corpus

The Old Portuguese Texts Corpus (Corpus de Textos Antigos - CTA)² comprises 32 texts (as of January 2024) written until 1525. Adhering to a principle of high fidelity to the source documents, the editions maintain the original graphic forms, lacunae, and errors. Utilizing the web-based framework TEITOK (Janssen, 2016; Vaamonde and Janssen, 2020), the corpus combines text with linguistic annotation.

Established in 2015, the CTA corpus has been subject to a data enhancement project, employing Natural Language Processing tools and meth-

²<https://teitok.clul.ul.pt/teitok/cta/index.php?action=home>

ods (Bico et al., 2022). One of the first tasks done was the enrichment of the data with lemmata and part-of-speech tags, with the respective morphosyntactic category and inflexion values. As all original graphic forms are maintained in the corpus, graphic variance poses a challenge for search queries. This is overcome by the existence of lemmas aggregating different forms and granular POS tags which allow the optimization of sophisticated lexical queries. For SVC, filtering both lemmatized nouns and verbs enables a more extensive cross-search spanning the entire corpus. For instance, in examples (4)-(6), both the noun *avito* and *religiã* exhibit various graphic forms.

(4) [OP] *tomasse ho avyto da Relligjã* [mPT] ‘*tomasse o hábito da religião*’ [EN] ‘[he/she] would take the habit of religion’ [MISJ:17755]

(5) [OP] *tomou ho nosso sancto avito da Religiã* [mPT] ‘*tomou o nosso santo hábito da religião*’ [EN] ‘[he/she] took our holy habit of religion’ [MISJ:21060]

(6) [OP] *tomara ho avito da Relligiam* [mPT] ‘*tomara o hábito da religião*’ [EN] ‘[he/she] had taken the habit of religion’ [MISJ:23066]

As described in (Bico et al., 2022), after a manual annotation task of the 150K tokens that make the ms. A of *Horto do Esposo* (HdE-A), two automatic annotation experiments were done using TreeTagger (Schmid, 1994, 1999). In the first experiment, a model trained with the batch of 150K tokens was applied on a set of two new texts: fragments D, C and E of *Horto do Esposo* (HdE-DCE), and ms. G1 of *Vida e Milagres de Santa Senhorinha de Bastos* (VMSSB-G1). The results varied between 91% and 67% of precision. Incorrectly annotated data was then manually revised and fed to a new, second model, which now comprised c. 167K tokens for training. The second experiment, which followed the premises of the first, applied the second model to four new, larger texts: *História de mui nobre Vespasiano*, *Memorial da Infanta Santa Joana*, *Livro dos Mártires*, and ms. E of *Vida e Milagres de Santa Senhorinha de Bastos* (respectively, Vespasiano, MISJ, LdM, VMSBB-E). Results slightly increased, varying between 77% and 82%. The manual revision task is of the utmost importance as it ensures the data correction and refinement for following annotation models and further work with the data. In the end, the corpus currently has seven annotated texts (one manually, six semi-automatically), whose combined totals

Text	Tokens	Precision
HdE-A	154,891	
HdE-DCE	2,694	90.97%
VMSSB-G1	13,944	67.41%
VMSSB-E	14,747	76.96%
Vesp[asiano]	19,134	82.36%
LdM	253,277	80.45%
MISJ	51,679	77.22%
Total	510,366	

Table 1: Total of annotated tokens in CTA and precision rates of the automatic annotation.

surmounts more than half a million of annotated tokens (Table 1).

3.2 Selected Texts

Horto do Esposo was first written in Portuguese, at the turn of the 14th to the 15th centuries (1390-1437), in the *scriptorium* of Mosteiro de Santa Maria de Alcobaça. Manuscript HdE-A dates from the same period. HdE-DCE is made of three fragments from a manuscript from Mosteiro do Lorvão. Spiritually didactic, the text uses *exempla* to illustrate several subjects, often recurring to the *auctoritas* and the Holy Bible.

Of the *História de mui nobre Vespasiano* there is only one known witness: the 1496 incunable printed in Lisbon by Valentim Fernandes. It tells the novelistic story of Roman emperor Vespasian who, to be cured from leprosy, seeks relics from Jerusalem.

The *Livro dos Mártires* was first written in Castilian by Bernardo de Briuega, at the mandate of King Alfonso X of Castille. It is the third part of a five book project and deals with the life of saints and martyrs from the beginning of Christendom. The text was translated into Portuguese during the reign of King Denis of Portugal (1279-1325), and it is known that a 15th century manuscript existed. However, the only surviving complete witness is the 1513 incunable printed in Lisbon by João Pedro Bonhominy.

The *Memorial of Infanta Santa Joana* tells the story of Infanta Joana of Portugal (1452-1490), daughter of King Afonso V of Portugal, and her entrance and life in the Mosteiro de Jesus of Aveiro. The text is supposed to have been first written between 1513 and 1525, and the copy is posterior to 1525.

The *Vida e Milagres de Santa Senhorinha de Basto* is a hagiographic tale of the life of Saint Senhorinha, a Portuguese abbess from the 10th century. There are four known witnesses of this text, all present in CTA, but only VMSSB-G1 and VMSSB-E have been lemmatized and annotated so far. The text was first written in the 13th century (1248-1284); ms. G1 was produced in the 17th century (1620-1645) and ms. E between 1692 and 1705.

3.3 Corpus processing

From an existing (though non-exhaustive) list of 175 support verbs previously identified in Modern Portuguese (Baptista and Mamede, 2020a,b), concordances were automatically extracted from the annotated partition of the corpus. The verbs on the list are not exclusively used for SVC and can also be used as full verbs, auxiliary verbs, etc. A 7-word window was applied for both the left and right context. A list of 27,645 instances of these verbs was automatically retrieved for 81 verbs of that initial list. Then, a manual identification of SVC was made, distinguishing them from other uses of these verbs. In this analysis, the predicative noun is indicated next to the respective concordance, along with the signs < and >, depending on whether the noun appears to the left or to the right of the support verb. Table 2 displays the distribution of some of these candidate support verbs, simultaneously, the most and the least frequently occurring, along with the number of SVC identified. A total of 2,982 SVC have been identified. This number is, in fact, slightly higher, as predicative nouns often appear coordinated, e.g. *fe e Confyāca* ‘fé e confiança’ ‘faith and confidence’. A total of 112 instances of this coordinated noun phrases have been found, raising the total to 3,094 SVC found (949 different types). For 2,423 instances, the predicative noun is on the right of the support verb, and for 559 are on the left (often as antecedent of a relative subclause). To the best of our knowledge, this is the largest collection of SVC in Medieval Portuguese available.³ The identification of SVC is still ongoing: verbs ‘*ser*’ (11,153 occurrences) and ‘*estar*’ (1,424 occurrences) ‘*be*’ were not considered at this stage.

The most frequently occurring Vsup verbs are ‘*haver*’ and ‘*fazer*’ (‘to have’, ‘to do’), with

³Table 3 in the Appendix A shows a sample of the concordances with SVC. Appendix B shows the frequency of the predicative nouns types.

Verb	Total	SVC
haver (<i>have</i>)	4,163	1,556
fazer (<i>do</i>)	3,613	1,216
ter (<i>have</i>)	1,117	61
tomar (<i>take</i>)	667	91
...		
usar (<i>use</i>)	105	20
reinar (<i>reign</i>)	85	4
falecer (<i>die</i>)	83	4
cobrir (<i>cover</i>)	74	7
mover (<i>move</i>)	73	1
manter (<i>maintain</i>)	32	7
cessar (<i>cease</i>)	31	3
cometer (<i>commit</i>)	22	3
firmar (<i>firm</i>)	8	1
conceder (<i>concede</i>)	5	2
incorrer (<i>incur</i>)	5	5
render (<i>yield</i>)	5	3
gozar (<i>enjoy</i>)	3	2
proferir (<i>utter</i>)	1	1

Table 2: Distribution of SVC in the corpus

1,470 and 1,243 instances as Vsup in SVC, respectively. While not reaching these high frequencies, it’s worth noting that the verbs ‘*ter*’ and ‘*tomar*’ (‘to have’, ‘to take’) also appear, with 61 and 91 occurrences, respectively. All other verbs exhibit a residual presence in the corpus, contributing only a few instances of SVC each.

4 Results: SVC in the CTA

Not being possible to analyse all of the cases found, only note-worthy observations will be made in this Section.

4.1 Nominalizations

Numerous SVC, featuring predicative nouns, find their equivalent in transformationally related verbal constructions, as noted by various scholars (Harris, 1964, 1991; Gross, 1981). Moreover, these instances of *nominalization*, a paraphrastic equivalence between the full verb and the predicative noun constructions, exhibit identical meanings and distributional constraints. Consequently, they hold a transformational status within the grammatical structure (Harris, 1964, 1991).

This phenomenon is also apparent in Old Portuguese, as the corpus presents instances of SVC alongside the corresponding verbal predicates. An illustrative example is *alguém fazer injúria a al-*

guém = *alguém injuriar alguém*, which translates to ‘someone doing/giving offense/insult to someone’ = ‘someone offending/insulting someone’. The SVC has already been shown in examples (1)-(2). Additionally, there is evidence of the corresponding verbal construction in (7), which even encompasses a passive construction in (8).

(7) [OP] *alugue quẽ me ãJuriasse*. [mPT] ‘alguém que me injuriasse.’ [EN] ‘someone who would insult/offend me’ [HdE-A:95923]

(8) [OP] *enfadou de seer ãJuriado per cinco ãnos* [mPT] ‘enfadou-se de ser injuriado por cinco anos’ [EN] ‘became fed up with being insulted/offended for five years’ [HdE-A:95976]

4.2 Standalone predicative nouns

However, the presence of a nominalization is not a prerequisite for establishing the predicative nature of a noun. Indeed, numerous predicative nouns exist as morphologically isolated lexical units, lacking any corresponding verbal or adjectival equivalent. These *standalone* or *autonomous predicative nouns* (Gross, 1981; Ranchhod, 1990; Baptista, 2005; Baptista and Mamede, 2020b), can be identified through various means: (i) their retention of the same function in Modern Portuguese, often accompanied by a similar syntactic construction; this is the case of the noun *ãJuria* ‘*injúria*’ ‘insult/offense’, in the examples above; (ii) there is supporting evidence from corpus, showcasing syntactic patterns alongside other predicative nouns from the same text; this can be seen in example (9):

(9) [OP] *ella tijnha fe e Confyãca ã deus* [mPT] ‘ela tinha fé e confiança em deus’ [EN] ‘she had faith and trust in God’ [MISJ:26966]

In this example, the predicative nouns *fe* ‘fé’ ‘faith’ and *Confyãca* ‘confiança’ ‘trust’ appear coordinated under the support verb *tãer* ‘ter’ ‘have’, with the same subject *ella* ‘ela’ ‘she’ and the same complement *deus* ‘deus’ ‘God’. This example allows one to consider both nouns as predicates expressed by SVC. On one hand, both nouns persist in Modern Portuguese, maintaining the same construction and conveying similar meanings, as per criterion (i). On the other hand, the coordination of these nouns under the same support verb in (5) implies, in accordance with criterion (ii), that they also serve a similar function. Indeed, *Confyãca* serves as a predicative noun, evident in its standalone occurrences with the same support verb, as in (9). Additionally, there is evidence of the cor-

responding nominalization in texts from the same period, as seen in (10)-(11).

(10) [OP] *pois confiãça tẽ em deos* [mPT] ‘pois confiança tem em deus’ [EN] ‘for [s/he] has trust in God’ [LdM:27891]

(11) [OP] *e todos que cõfiam ã elle* [mPT] ‘e todos os que confiam nele [= em_ele]’ [EN] ‘and all those who trust in him.’ [HdE-A:48062]

In turn, in spite of not having a verbal equivalent construction, the occurrence of *fe* (or the variant *fee*) ‘fé’ ‘faith’ both in sentences like (9), coordinated with another predicative noun, and in standalone mode in sentences like (12), allows one to consider *fe* a predicative noun and *tãer* ‘ter’ or *aver* ‘haver’ ‘have’ its support verbs.

(12) [OP] *e a fe que eles tijnhã* [mPT] ‘e a fé que eles tinham’ [EN] ‘and the faith that they had.’ [LdM:216921]

In certain cases, the support verb construction (SVC) has disappeared from use, despite the continued presence of the individual words—both the support verb and the noun—in the modern language. This is the case of the SVC *ter mentes* (13):

(13) [OP] *Hũũ sabedor teue mẽtes ãnas cousas do mũdo* [mPT] ‘Um homem sábio teve mentes nas coisas do mundo’ [EN] ‘A wise man had minds on the things of the world/on worldly matters’ [HdE-A:57369]

whose meaning could be defined as ‘to consider, to reflect upon’ or similar expressions. A systematic analysis of the usage of this expression, facilitated by the annotated corpus used here, can contribute to a more precise understanding of both its meaning and syntactic properties.

4.3 Variant support verbs

One of the defining properties of SVC is that the same predicative noun can select, besides its elementary *Vsup*, the one with the least semantic load and broader distribution, a variety of other support verbs, which introduce nuances of aspect (Fotopoulou et al., 2021), modality and style. This variation is also abundantly exemplified in the corpus, as seen in examples (14a)-(14d) with the noun *fe* ‘fé’ ‘faith’. This nouns displays several SVC both with the elementary support verbs *tãer* ‘ter’ ‘have’ (14a) and ‘aver’ ‘haver/ter’ ‘have’ (14b); the durative/permansive variant *manter* ‘id.’ ‘keep/maintain’ (14c); and the terminative variant *perder* ‘id.’ ‘loose’ (14d), respectively.

(14a) [OP] *a ajuda de deus . ã que tinha toda sperãca e ffe* [mPT] ‘a ajuda de Deus, em quem tinha toda a esperança e fé’ [EN] ‘help of God, in whom he had all hope and faith’ [MISJ:26289]

(14b) [OP] *por sempre fostes mentirosos e que no auedes fee nemhuã* [mPT] ‘porque sempre [vós] fostes mentirosos e porque não haveis fé nenhuma’ [EN] ‘because [you_pl.] have always been liars and because [you_pl.] have no faith.’ [LdM:93832]

(14c) [OP] *E o que mantem aquesta fe pola boã obra sera parceiro* [mPT] ‘E o que mantém esta fé pela boa obra será parceiro’ [EN] ‘And the one who maintains this faith through good deeds will be a partner’ [LdM:64555]

(14d) [OP] *homẽs maos e desleaes que auiaã perdida a sua fe.* [mPT] ‘homens maus e desleais que tinham perdido a sua fé’ [EN] ‘evil and disloyal men who had lost their faith’ [LdM:193826]

These examples do not exhaust the full syntactic diversity of the SVC involving the noun *fé* and other constructions in which this noun appears in the corpus.⁴

As mentioned in the literature, we also observed a differential distribution between the more commonly used *aver* and the less frequent *teer*. This

⁴Due to space constraints, we cannot explore them further, but we can briefly mention a few, providing a tentative description: (i) Permansive SVC ‘*ter-se na fé*’:

(15) [OP] *huã martire que se teve firmemente na fé de nosso senhor* [mPT] ‘um mártir que se manteve firmemente na fé de nosso senhor’ [EN] ‘a martyr who remained steadfast in the faith of our Lord’ [LdM:296]; (ii) Terminative SVC with the predicative noun as the subject:

(16) [OP] *Mas eu roguei por ti que tua fé não falhasse* [mPT] ‘Mas eu roguei por ti [para] que a tua fé não falhasse’ [EN] ‘But I have prayed for you [so] that your faith may not fail’ [LdM:23046]; (iii) Another SVC with a kind of echo complement introduced by *em*:

(17) [OP] *Se ouverdes em vós fé tanto como grão de sebe* [mPT] ‘Se tivéreis em vós tanta fé como grão de sebe’ [EN] ‘If you had within you so much faith like a grain of mustard seed’ [LdM:74610]; (iv) An (arguably) SVC with *ficar em*:

(18) [OP] *E muitos ficaram em na fé de nosso senhor Jesus Cristo* [mPT] ‘E muitos ficaram na fé de nosso senhor Jesus Cristo’ [EN] ‘And many remained in the faith of our Lord Jesus Christ’ [LdM:65467]; (v) Another (arguably) SVC with *tornar-se a*:

(19) [OP] *E aconselhavam-lhes estes santos todo o dia que se tornassem à fé de nosso senhor Jesus Cristo* [mPT] ‘E aconselhavam-lhes estes santos todo o dia que regressassem à fé de nosso senhor Jesus Cristo’ [EN] ‘“And these saints advised them every day to return to the faith of our Lord Jesus Christ.”’ [LdM:68776]; (vi) Finally, an autonomous (unrelated) SVC with *dar*, still existing in Modern Portuguese:

(20) [OP] *E eu dão fee aas palauras delles* [mPT] ‘E eu dando fé às palavras deles’ [EN] ‘And I, giving credence to their words’ [LdM:29731]. Other constructions exist but cannot be further detailed.

trend may have eventually resulted in ‘*ter*’ assuming the role of support verb that the former held.

4.4 Conversion

A distinctive type of SVC involves *Conversion*, a transformation introduced by Gross (1981) and extensively described for Modern French by Gross (1989, 2022). This transformation is akin to a verbal active/passive transformation, but is specific to SVC. During this operation, the positions of the arguments of the predicative noun are altered: the agent-subject in the standard, active-like SVC becomes the prepositional complement in the converse, passive-like SVC. Simultaneously, the patient/object-complement becomes the grammatical subject of the converse, and the support verb in the standard construction is replaced by a converse support verb.

This transformation has been extensively examined in Modern Portuguese Ranchhod (1990); Baptista (2005); Chacoto (2005); Rassi et al. (2016); Rassi (2023), among others. Evidence of it is also present in the CTA corpus, as illustrated by examples (21)-(22), with *mercee* ‘*mercê*’ ‘favor/gift’:

(21) [OP] *Fazé me tanta mercee* [mPT] ‘Fazei-me tanta *mercê*’ [EN] ‘Do me such a favor’ [MISJ:26593]

(22) [OP] *eu reãebẽdo delle merãee* [mPT] ‘eu recebendo dele *mercê*’ [EN] ‘I receiving from him [a] favor’ [LdM:230811]

However, it’s important to note that the same predicative noun can determine multiple SVCs, each characterized by different support verbs and structures, and denoting distinct meanings. An example of this is the noun *mercee*, which conveys a different notion of ‘ *piedade*’ ‘pity/mercy’ in another SVC, as demonstrated in (23) with the support verb *aver* ‘*haver/ter*’ ‘have’.

(23) [OP] *bõa mulher aue merãee de mi* [mPT] ‘boa mulher, *tende piedade de mim*’ [EN] ‘good woman, have mercy on me’ [LdM:80162]

The attentive reader would likely have observed that, in these examples, the noun appears in three different spellings. Consequently, these nuanced distinctions can only be discerned through a systematic examination of the distribution of a given noun in the corpus. This is made feasible by querying it through the same (modern) lemma, ‘*mercê*’, as it was done here.

4.5 Operator-verb constructions

During the process of identifying SVCs, care is taken to avoid confusing support verbs with other syntactic constructions. Indeed, certain verbs, besides their role as support verbs, display a range of syntactic constructions, including both *full* or *distributional verbs* (24) and (tense) *auxiliary verbs* (25).

(24) [OP] *Como sam sadornim mādou fazer hũa casa muy pequena açerca da igreja* [mPT] ‘*Como São Sandornim mandou fazer uma casa muito pequena próximo da igreja*’ [EN] ‘*How Saint Sandornim ordered to have a very small house built near the church.*’ [LdM:10525]

(25) [OP] *e que lhe tinham os inimigos cercado o castello d aguiar*⁵ [mPT] ‘*e que os inimigos lhe tinham cercado o castelo de Aguiar*’ [EN] ‘*and that his enemies had surrounded the castle of Aguiar*’ [VM-G1:11353]

In this context, and due to its potential confusion with SVC, this paper emphasizes a specific verb-noun construction: *operator-verb constructions* (*Vop*). Operator-verbs, a theoretical construct introduced by Gross (1981), serve to describe a complex construction formally resembling SVC but incorporating a new element of **cause**, while maintaining the distributional constraints between the predicative noun and its arguments. Consequently, they are also referred to as *causative operator-verbs* (*Vopc*). To be precise, *Vopc* operate on an elementary sentence, introducing a new element and establishing a causal relationship between this element and the sentence they modify. Typically, *Vopc* induce some form of restructuring in the sentence, normally ‘absorbing’ the support verb and altering the syntactical function or case of the predicate noun and its arguments within the sentence. These operations may result in surface sequences where the *Vopc* and the predicative noun formally resemble those of SVC. Example (26) provides an illustration of the *Vopc* *fazer* ‘id.’ ‘do/make’ operating on the predicative nouns *temor* ‘id.’ ‘fear’ and *espanto* ‘id.’ ‘astonishment’.

(26) [OP] *faziam me temor e graue espanto* [mPT] ‘*faziam-me temor e grave/muito espanto*’ [EN] ‘*[they] filled me with fear and great astonishment*’ [HdE-A:32599]

⁵We do not enter here in the discussion of this periphrastic verbal tense and the agreement of the part participle with the sentence subject. Please see below the concept of *Vopl*.

The SVC involving *temor* and *espanto* are otherwise well-documented in the corpus (12)-(13), employing the verb *haver* ‘have’, which would be replaced in Modern Portuguese by *ter* ‘have’, as seen from examples (27)-(28):

(27) [OP] *emristeçer e hauer temor e angustia* [mPT] ‘*enrister e ter temor e angústia*’ [EN] ‘*to sadden/become sad and to have fear and anguish*’ [LdM:23327]

(28) [OP] *grande foy o espãto que ouue porque* [mPT] ‘*grande foi o espanto que teve porque*’ [EN] ‘*great was the astonishment that he/she had because*’ [LdM:248295]

Another complex construction introduced by Gross (1981) is the *linking operator-verb* (*Vopl*). In this scenario, an element also is added to an elementary sentence, but this element is already an argument of the predicative noun, hence the term *linking*. This is indicated by a constraint of coreference with the predicative noun, typically accompanied by a possessive determiner that is obligatorily coreferent to the *Vopl* subject. *Vopl* constructions can thus be seen as a linguistic device for restructuring an SVC elementary sentence. This is achieved by extracting the linked element and placing it in the more prominent syntactic position of the subject in the sentence. Example (29) illustrates a *Vopl* construction.

(29) [OP] *e aquele entenda que tem a sua nobreza ãteyra.* [mPT] ‘*e aquele entenda que tem a sua nobreza inteira*’ [EN] ‘*and let that person understand that he has his nobility intact*’ [HdE-A:78625]

In this example, *nobreza* is modified by a secondary predication, *ãteyra* ‘intact’, and there is constraint coreference between the subject of *tẽer* ‘ter’ ‘have’ and the possessive determiner of the noun (*sua* ‘id.’ ‘his’). In Modern Portuguese a similar *Vopl* construction exists, as shown by the gloss. However, in Medieval Portuguese, the noun *nobreza* selects the support verb *haver* ‘id.’ and not *tẽer* (30). Similarly, in this stage of the language, the copula verb selected by the adjective *ãteyro* is *seer* ‘ser’ ‘be’ and not *estar* ‘id.’ ‘be’ (31).

(30) [OP] *que ha grande nobreza do teu linhagem* [mPT] ‘*que tem a grande nobreza da tua linhagem*’ [EN] ‘*that has the great nobility of your lineage*’ [LdM:242119]

(31) [OP] *ca nõ es ãteyro ãno spiritu* [mPT] ‘*pois [tu] não estás inteiro no espírito*’ [EN] ‘*for [you] are not whole in spirit*’ [LdM:144083]

In this case, one could arguably analyze (31) as a *Vopl* construction of *ter* on the adjectival construction of *ẽteyro*, which predicates on the noun *nobreza*, as illustrated by the following structure (32):⁶

(32) [OP] *aquele_i tem # a nobreza de_aquele_i ẽteyra* [mPT] ‘aquele_i tem # a nobreza de_aquele_i ẽ inteira’ [EN] ‘that person has # the nobility of that person is intact’

In spite of having identified many of these instances of *Vopc* and *Vopl*, they were not considered at this stage of the annotation, and were left for future work.

4.6 Contrasting constructions

Several impersonal constructions, expressing *meteorological* predicates were identified (33)-(34).

(33) [OP] *Porque fazia entom frio* [mPT] ‘Porque fazia entãõ frio’ [EN] ‘because it was cold’ [LdM:24447]

(34) [OP] *e tu poseste as tebras e fez sse a noyte* [mPT] ‘e tu puseste as trevas e fez-se a noite’ [EN] ‘and you put the darkness and night was made’ [LdM:48130]

These constructions closely resemble those found in Modern Portuguese. The grammatical status of verbs such as ‘fazer’ ‘do’ is unclear, as well as its two contrasting constructions: the pronominal structure of this verb with *noyte* ‘noite’ ‘night’, against the non-pronominal structure with *frio* ‘id.’ ‘cold’.

Additional contrasting constructions were identified. Certain verbs, such as *ministrar* ‘ministrate’, do not seem to be utilized as support verbs in Medieval Portuguese. Out of the 13 occurrences of this verb, one can find as its direct complement the generic noun *cousas* ‘coisas’ ‘things’ (5 instances) and with the noun *riquezas* (1 instance), in the sense of ‘administrate’ (35):

(35) [OP] *que auia satẽta e tres que lhe menis-trauã suas riquezas . e suas herdades* [mPT] ‘que havia sete e trẽs [anos] que lhe ministrava suas riquezas e suas herdades’ [EN] ‘that there were seven and three [years] that he/she administered his/her wealth and his/her estates.’ [LdM:116992] The so-called ‘relative subclause without explicit antecedent’ (Veloso, 2013), has also been found (1 instance), for example (36):

(36) [OP] *cõ a dita Senhora steuesse e lhe minystrasse e aparelhasse o que avija de comer* [mPT] ‘com a dita Senhora estivesse e lhe ministrasse e aparelhasse o que houver de comer’ [EN] ‘that he/she be with the said Lady and serve and prepare what there is to eat’ [MISJ:18795].

This full/distributional verb ‘ministrar’ is no longer in use in Modern Portuguese. Instead, the verb *administrar* presents a very similar construction. This later verb also often occurs as a variant *Vsup* of *ministrar*. It is also worth mentioning an apparently intransitive construction, eventually with a dative beneficiary complement (37):

(37) [OP] *E logo sse ella leuãtou sãã e mynjstrou a todos.* [mPT] ‘E logo se ela levantou curada, serviu a todos’ [EN] ‘And she immediately got up, [fully] healed, and she served [to] everyone’ [LdM:78346]

In contrast, in Modern Portuguese SVC, *ministrar* is a relatively common stylistic variant of *dar*, as seen in examples like *ministrar – assistência, curso, medicamento, sacramento, tratamento* ‘to provide – assistance, a course, medicine, a sacrament, treatment’ (Baptista, 1997; Rassi et al., 2016; Rassi, 2023). Naturally, more precise insights on the constructions of this verb ‘ministrar’ may be gained through further investigation, potentially by incorporating more annotated data.

Likewise, in the CTA corpus, the verb *cometer* seldom functioned as a *Vsup*, as in (38)

(38) [OP] *Arrepẽdẽdo se do crime que tinha cometido* [mPT] ‘arrepẽdendo-se do crime que tinha cometido’ [EN] ‘regreting the crime [he/she] had committed’ [LdM:25447]

This is the exclusive usage of *cometer* in Modern Portuguese, as seen in examples like *cometer – crime, infração, pecado* ‘to commit a crime, infraction, sin’. Its prevailing role in the corpus, however, denotes ‘movement’, as seen in (39), aligning with the Modern Portuguese counterpart *acometer* ‘to attack, approach’:

(39) [OP] *e cõ uirtude do cumo dela uay cometer o basilico e vẽçe o e mata* [mPT] ‘e com virtude, do cimo dela vai acometer (=atacar) o basilisco e vence-o e mata’ [EN] ‘“And with valor, from the top of it, he will assail the basilisk, defeating and killing it.”’ [HdE-A:67550]

From a different perspective, since predicative noun can be employed with various support verbs, yielding different meanings, some constructions persist from Medieval Portuguese, while others fell

⁶The ‘#’ separates the *Vopl* from the elementary sentence on which it operates; the indices *i* denote coreference.

in disuse.⁷ For instance, in example (42), *fazer deuacões* ‘fazer devoções’ ‘make devotions’ conveys the notion of offering **prayers**. Conversely, in example (43), *têer deuacõ* ‘ter devoção’ ‘have devotion’ expresses the idea of experiencing a **sentiment** towards somebody/something. In Modern Portuguese, both meanings continue to exist.

(42) [OP] *E per todo ho Regno se fazijã muitas deuacões e oracões* [mPT] ‘E por todo o reino se faziam muitas devoções e orações’ [EN] ‘And throughout the kingdom, a lot of devotions and prayers were done’ [MISJ:33714]

(43) [OP] *ẽ algũas festas e dias ẽ que tinha mais spicial deuacõ* [mPT] ‘em algumas festas e dias em que tinha mais especial devoção’ [EN] ‘in some feasts and day in which [she/he] had more special devotion’ [MISJ:3650]

In a different context, both *aver doo* in (44) and *tomar doo* in (45) signify having a feeling of pity towards something. In contrast, *fazer doo* in (46) corresponds to an expression of lamentation. In contemporary Portuguese, the latter meaning no longer exists.

(44) [OP] *Doo ey eu da tua morte* [mPT] ‘Dó hei eu da tua morte’ [EN] ‘Pity I have of your death’ [LdM:215464]

(45) [OP] *e tomar doo por ho fallecimẽto desta sancta Senhora* [mPT] ‘e tomar dó pelo falecimento desta santa senhora’ [EN] ‘and take pity for the passing of the holy lady’ [MISJ:45727]

(46) *e disto fezerõ as donas grande doo polla morte de ...* ‘e disto fizeram as donas grande dó pela morte de ...’ ‘and from this, the ladies made a great lamentation for the death of.’ [Vesp:11659]

5 Conclusion and Future Work

This paper systematically explored support-verb constructions (SVC) in the *Corpus de Textos Antigos* (CTA). With a corpus annotated with lemmata and morphosyntactical categories, a little more than

⁷It is also possible to find SVC in which the predicative noun is the *subject* of the sentence (40)-(41) (Baptista, 2022):

(40) [OP] *Ca ẽnos presuntuosos reyna a soberua* [mPT] ‘Porque nos presuntuosos reina a soberba’ [EN] ‘because in the presumptuous reigns pride/arrogance’ [HdE-A:83107]

(41) [OP] *supytamẽte a tomou hũu leue sõpno* [mPT] ‘subitamente, um leve sono tomou-a’ [EN] ‘suddenly, a light sleep overtook her’ [HdE-A:83107]

The Vsup involved in this less studied ‘predicative noun as subject’ construction are the same and support the same type of predicates as found in Modern Portuguese, as the examples above show.

27,000 instances were automatically extracted, involving 81 different candidate verbs, also serving as support verbs. Manual analysis was undertaken to differentiate SVCs from other verb uses, resulting in the identification of a total of 3,094 SVC instances (949 different types). To the best of our understanding, this compilation represents the largest collection of SVCs in Medieval Portuguese currently available. SVCs featuring the verbs *ser* and *estar* (‘be’) are currently under investigation.

From the studied SVC, cases of nominalizations and standalone predicative nouns were identified, along with various support verb variants and types of converse SVC. Distinctive constructions, including operator-verb constructions (both causative and linking operator-verbs), were also identified. Subsequent efforts should aim to systematize the identification of operator-verbs. The most commonly employed support verbs in SVC within this corpus were *fazer* ‘to do/make’ and *haver* ‘to have’. These verbs, being less semantically loaded compared to other support verbs, exhibit a broader distribution concerning the predicative nouns that select them, thereby contributing to their higher frequency of use. Numerous instances of SVC identified in the CTA corpus extend beyond Medieval Portuguese, persisting seamlessly into Modern Portuguese. This sustained continuity solidifies the enduring status of SVC as an integral component of the linguistic heritage within the language. The linear order of SVC constituents remains comparable between both periods, albeit with a few instances in Medieval Portuguese allowing the inversion of Vsup and Npred, a phenomenon that is currently blocked.

Future efforts should prioritize the extraction and analysis of instances (a) based on the list of predicative nouns found at this stage, to explore Vsup variation and discover unseen word senses; and (b) expand it to other predicative nouns determined in Modern Portuguese (Baptista and Mamede, 2020b). Attention should also be directed towards complex noun phrases that arise from the reduction of SVC (Gross, 1981). These phrases are headed by the predicative noun and retain their arguments throughout the reduction process. Furthermore, operator-verb constructions, both causative and linked Vop, related to SVC should be systematically described as well. An examination of the intersection between data related to support verbs and the results obtained for predicative nouns would enhance the diachronic study of Portuguese SVC.

References

- Jorge Baptista. 1997. *Sermão, tarefa e facada: uma classificação das expressões conversas dar-levar*. *Seminários de Linguística 1*, pages 5–37.
- Jorge Baptista. 2005. *Sintaxe dos Predicados Nominais com ser de*. Fundação para a Ciência e a Tecnologia & Fundação Calouste Gulbenkian, Lisboa.
- Jorge Baptista. 2022. Support verb constructions with predicate noun in subject position. *Bulletin de linguistique appliquée et générale*, 40:379–397.
- Jorge Baptista and Nuno Mamede. 2020a. *Dicionário Gramatical de Verbos do Português Europeu*. Universidade do Algarve.
- Jorge Baptista and Nuno Mamede. 2020b. Syntactic Transformations in Rule-Based Parsing of Support Verb Constructions: Examples from European Portuguese. In *9th Symposium on Languages, Applications and Technologies (SLATE 2020)*, volume 83 of *OpenAccess Series in Informatics (OASICs)*, pages 11:1–11:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.
- Maria Inês Bico, Jorge Baptista, Fernando Batista, and Esperança Cardeira. 2022. Early experiments on automatic annotation of portuguese medieval texts. In *Linking Theory and Practice of Digital Libraries*, pages 442–449, Cham. Springer International Publishing.
- Ivo Castro. 2006. *Introdução à História do Português, 2.ª edição*. Edições Colibri.
- Lucília Chacoto. 2005. *O Verbo Fazer em Construções Nominais Predicativas*. Ph.D. thesis, Universidade do Algarve, Faro.
- Lucília Chacoto. 1996. Predicados nominais com *fazer* no português medieval. In *Actas do XII Encontro da Associação Portuguesa de Linguística*, volume 2, pages 69–77.
- Aggeliki Fotopoulou, Eric Laporte, and Takuya Nakamura. 2021. Where do aspectual variants of light verb constructions belong? In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 2–12. Association for Computational Linguistics.
- Gaston Gross. 1989. *Les construction converses du français*. Droz, Genève.
- Gaston Gross. 2022. *Manuel d’analyse linguistique: approche sémantico-syntaxique du lexique*. Presses universitaires du Septentrion.
- Maurice Gross. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 15(63):7–52.
- Maurice Gross. 1996. Lexicon-grammar. In Keith Brown and Jim Miller, editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge.
- Zellig Sabettai Harris. 1964. The elementary transformations. In Henry Hiz, editor, *Papers on Syntax*, pages 211–235. D. Reidel Pub. Co.
- Zellig Sabettai Harris. 1991. *A Theory of Language and Information. A Mathematical Approach*. Clarendon Press, Oxford.
- Maarten Janssen. 2016. **TEITOK: Text-faithful annotated corpora**. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).
- Maria Regina Pante. 2012. O verbo *tomar* como verbo-suporte no português arcaico. *Línguas & Letras*, 13(24).
- Maria Regina Pante and Érica Fernanda Ortega. 2009. O verbo *tomar* como verbo-suporte no livro de ensinância de bem cavalgar toda a sela. *Revista Philologus*, 125:144–152.
- Elisabete Ranchhod. 1990. *Sintaxe dos predicados nominais com estar*. Instituto Nacional de Investigação Científica (INIC).
- Elisabete Ranchhod. 1999. Construções com nomes predicativos na *Crónica Geral de Espanha de 1344*. In I. H. Faria, editor, *Homenagem ao Homem, ao Mestre e ao Cidadão*, pages 667–682. Edições Cosmo.
- Amanda Rassi, Nathalia Calcia, Oto Araújo Vale, and Jorge Baptista. 2016. Estudo contrastivo sobre construções conversas em PB e PE. In O. Nadin, A. Ferreira, and C. Fargetti, editors, *Léxico e suas interfaces: descrição, reflexão e ensino*, pages 199–218. Cultura Acadêmica.
- Amanda Pontes Rassi. 2023. *O verbo dar em português brasileiro. Descrição, classificação e processamento automático*. Letraria, São Paulo.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 154–163.
- Helmut Schmid. 1999. *Improvements in Part-of-Speech Tagging with an Application to German*, pages 13–25. Springer Netherlands, Dordrecht.
- Gael Vaamonde and Maarten Janssen. 2020. *Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK*, pages 271–292.
- Rita Veloso. 2013. Gramática do português. volume 2, chapter Subordinação relativa, pages 2061–2134. Fundação Calouste Gulbenkian.

Appendix A - SVC from the CTA corpus

ID	Right context	Vsup	Left Context
MISJ:34147	afeytauas afeytou sse ela . Quando tu	aauias	fame ouue ela fame e quando tu
MISJ:38305	que nos chamamos filho de deus .	auêdo	mjsericordia sobre os homêes
HdE-A:117731	muitas lides , e contendas as de	auer	com o inimigo , ca sei sarta
LdM:117037	que os que pecã nõ	auerem	pena . E estes maaes sobreditos
LdM:38822	e preguar , pera as gentes	auerem	notiçia , e conhocimento a sua vida
MISJ:13404	o mestre salla preguntou a jacob onde	aueria	pousada e jacob respondeu lhe e dixe
LdM:124182	E porque algũ per vêtura	aueria	sabor de os aprêder teue por bẽ
LdM:251724	mas empero, que grande fiuza ella	auia	em Deos , nenhũ nõ sabe pero
LdM:147565	comgrande alegria porque	auia	reçeo do padre encorrer na sua ira
LdM:189484	pera fazerem festa , assi como	auiam	custume de fazer quada sabodo
LdM:66406	sobre esto cõ hũũ sancto bispo que	auja	nome alexandre
MISJ:14442	Mais porque o sancto bispo nõ	auja	aquella arte nem tal sçiencia
MISJ:25178	Exemplo Hũũ homẽ	auja	grande odio mortal a outro
MISJ:9536	assy . que este homẽ que asy	auja	este malquerença
LdM:133001	contra ele com grande sanha que	auya	cuydãdo em que gujsa
VM-G1:4363	Aÿda muito minyno pera Casar e	aver	erdeyro e ser sempre doente
MISJ:7006	neste moesteiro . pera que	averia	êteyro poder
LdM:18116	papa e mestre da ordẽ .	Avido	maduro cõsselho . e vêedo seer
HdE-A:91771	lenco na mão e Rosto . nõ	cessãdo	de lagrimas e salucos e vrrros mÿy
Vesp:338	suplicou ao sancto padre lhe	cõcedesse	graca de plenaria yndulgẽcia
LdM:104132	aquelle era seu filho . E mãdou	cometer	casamêto cõ elle pera sua filha
LdM:191166	Arrepêdêdo se do crime que tinha	cometido	bolueo as trinta peças de prata
LdM:82187	soo filha do dito Conde . lhe	Cometyã	grãdes e altos Casamêtos cõ duques
VM-G1:492	pera lhe seer per elles	Concedido	e outorgada licẽca e liberdade pera
Vesp:3975	daquella terra treeuosa e	cuperta	d' escuridade e de morte perdurauel
LdM:197570	E porque leyxou a sabedorya	êcorreo	ẽ ygnorancia . que he abetamento
LdM:27972	sustancia . nõ departida .	emcorreo	ẽ cobijça que he corrupcõ que
HdE-A:30929	da mête e da alma e faz	falecer	a ujsta do verdadeyro conhecimêto
HdE-A:122634	que aquesto fosse . aynda aly nõ	faleçia	a ajuda . nem a merçe de
LdM:74342	os homês . E com aquelle ujnho	fara	o senhor deus cõujte ao seu pouoo
MISJ:18828	grãde tempo que o nõ vio .	faz	grãde alegria . Assy se alegrauã eles
LdM:106925	esto faz aquelle que nõ cuyra nẽ	faz	conta que os outros cuyrem delle .
LdM:40977	todos aqueles que	fazẽ	a sua uõtade . Elle destruy cruelmête
VM-G1:10999	sã çebriã começou a chorar	fazêdo	cõfissã de todolos peccados
LdM:17088	se ally gastã aquelle dia . E	fazem	muytas outras alegrias
HdE-A:145171	e seus amjgos . tomã prazer e	fazem	grandes convites . E esto ham elles
MISJ:30459	todos a fogir asanhãdo se . e	fazendo	muy grãde arroydo . e diserom todos
HdE-A:107979	o bispo de yterãna	fazendo	bõa vida mereçeo de se gozar as
MISJ:8335	gente . chorando e	fazendo	chanto a muy grandes vozes . por
HdE-A:94185	aa sua madre muy sancta . e	fazendo	conparaçõ antre ty e elles acharas que
Vesp:2264	toda ajuda que fisicos lhe podiã	fazer	E metera sse a handar per
MISJ:39328	Outrossy nõ deue homẽ deixar de	fazer	a caridade aos Jrmãos pello studo da
MISJ:6126	seruiã outrossy a ella . sẽ	fazer	cerymonyas nẽ deferẽca
HdE-A:58119	a castigar . e afaagar . e	fazer	lhe muytos viços . e a falar
HdE-A:120357	ẽ aquela festa começou de	fazer	tã grãdes virtudes e tã grãdes marauilhas
VM-G1:3723	a qual disse . molher boas obras	fazes	: leuãta te pella manhaã e sairas
LdM:137425	ante que fosse abade	fazia	grandes abstêẽças
VM-G1:13442	homẽ bebedor . e que	fazia	adulterio . e amigo de totalas cousas
Vesp:8701	he casado cõ ela aquele que lhe	fazia	muytos falsos afagamêtos primeyro
LdM:211701	que lhe derõ pollo bẽ que lhes	fazia	Assy senhor que eu nõ sey
MISJ:1210	os tormentos que sofriam . Mas pero	fazia	lhes deos muyto bem e muyta merçee
HdE-A:54198	assi todos na igreja pella chuiua que	fazia	, hũa bõa molher auendo doo de
HdE-A:48044	atee os mesmos caualleiros	faziã	bulrra dele . e se chegauã a
MISJ:24022	de sam fracisco . os frades	faziã	ujda estreyta e aspera ẽ grande pobreza
VM-G1:5367	a el rrey d' ugaria . e foy	fecto	o casamêto da madre cõ o filho
MISJ:1682	que vio o grãde aluoroço que era	feito	ẽno poboo mandou o de cabo emçarrar
MISJ:7766	E em aquele logar som muytos beês	feitos	por ele a louuor de nosso senhor
HdE-A:68016	Estas tres batalhas forã	feytas	ẽ tres ãnos . en que forã
HdE-A:92056	Como se solêpnemête teuesse	feyto	voto de proffysam e obediencia
MISJ:45134	Ja entõ nõ podendo mais	fez	ho padre a absoluycã
HdE-A:61653	de çiriaco a carpasio o vigairo e	fez	banho ẽna fonte do bauptismo
HdE-A:31394	das brauezas que o adiantado	fez	contra o seu corpo . e todolos
HdE-A:150910	aquelle que a muytos parêtes se nõ	fezer	sua uõtade e seu prazer delles .
LdM:143707	totalas chagas que [. . .] maximiano	fezera	em my . e esto com o
LdM:164715	e door . screueo lhe que nõ	fezesse	aballo nẽ partisse . atee elle lho
LdM:164095	pallauras . e discretas .	ffez	hũa aRenga ante el Rey . princepe

Semantic Exploration of Textual Analogies for Advanced Plagiarism Detection

Elyah Frisco Andriantsialo
GLoRe, Madagascar
elyahfrisco7@gmail.com

Volatiana Marielle Ratianantitra
GLoRe, Madagascar
volatianamarielle@yahoo.fr

Thomas Mahatody
GLoRe, Madagascar
tsmahatody@gmail.com

Abstract

This study explores textual analogies in French within the context of plagiarism detection, adopting a semantic approach. By combining traditional methods with advanced models such as BERT and GPT, the paper proposes a hybrid model to enhance detection efficiency. Comparative evaluation highlights the model's ability to detect subtle similarities and paraphrases. The approach represents a significant advancement in accurate plagiarism detection by leveraging deep contextual understanding and the reformulation capabilities of integrated models.

1 Introduction

Plagiarism detection, constantly evolving, remains a crucial challenge in the field of digital content management. The emergence of new copying methods and circumvention of traditional systems necessitates the exploration of more advanced and adaptive solutions. In this perspective, our research positions itself at the forefront of innovation by focusing on Semantic Exploration of Textual Analogies.

The current landscape, marked by the sophistication of plagiarism practices, underscores the urgency of adopting more complex and sophisticated approaches. Our research capitalizes on the latest advancements in natural language processing (NLP), thus laying the groundwork for a more robust plagiarism detection system.

2 Basic Theory

2.1 Plagiarism

Plagiarism is a term with moral and aesthetic connotations, used in literature to describe the act of incorporating, in an undisclosed and more or less faithful manner, textual elements from another author. This term is not commonly used in legal contexts, where one would rather refer to

infringement and violation of copyright law (Vandendorpe, 1992).

A document is considered plagiarized when it is produced by applying a series of transformations to an original document. The plagiarized document should retain the same function as the original but may take on a different form. There are several types of plagiarism, including copy-paste, paraphrasing, the use of false references, and plagiarism of ideas. (Mostafa, 2016)

2.2 Natural Language Processing

Natural Language Processing (NLP) is a multidisciplinary field involving linguistics, computer science, and artificial intelligence (AI) with the aim of creating NLP tools capable of automatically processing linguistic data for various applications.

Some of the most well-known applications include automatic translation, information extraction, text summarization, spell checking, automatic generation, voice synthesis, speech recognition, and the detection of specific topics (sentiment analysis, etc.) (Ratianantitra, 2023).

One outcome of the progress in NLP is GPT (Generative Pre-trained), a language model employing deep learning to generate text resembling human speech. In simpler terms, it's a computational system created to produce sequences of words, code, or other data from an input source known as the prompt. GPT finds applications in various tasks like machine translation, where it predicts word sequences statistically. The model is trained on an unlabelled dataset comprising texts from sources like Wikipedia, available mostly in English but also in other languages. This computational approach serves diverse purposes, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and more (Floridi, Luciano, Massimo Chiriatti, 2020).

3 Literature review

The literature review emphasizes the importance of detecting plagiarism by examining similarities between documents. Different approaches, can be explored. Table 1 summarizes these plagiarism detection methods, ranging from simple algorithms to advanced approaches.

Method	Description
Fingerprinting	Represents the document in the form of fingerprints (n-grams) and utilizes algorithms such as "Rabin-Karp" for plagiarism detection.
String Matching	Compares documents word by word using algorithms such as "Brute Force"
Bag of Words	Utilizes a vector space model with vectors representing documents, calculating cosine similarity to measure the similarity between texts.
Citation Analysis	Analyzes citations within texts to detect similar patterns in citation sequences, adapted for academic and scientific texts
Stylometry	Utilizes statistical methods to quantify and analyze the writing style of an author based on features such as word frequencies.
Rule-Based Algorithms	Simple to implement and quick, but limited to predefined rules and less suitable for different languages and sentence structures.
Neural networks	Achieves high performance on complex texts, detects paraphrases and similarities, but requires a high level of implementation complexity and massive amounts of data for training.
Bidirectional Encoder Representations from Transformers (BERT)	Utilizes a pre-trained model on a large corpus of text, provides a deep understanding of the text but requires high computational power and is not designed for text generation.

Table 1: Overview of Plagiarism Detection Methods

When comparing documents to detect plagiarism, the search for similarities is crucial. Word-for-word comparison, while effective in identifying "copy and paste" instances, becomes insufficient in the face of sophisticated paraphrasing and rephrasing. The work of Barron-

Cedeño et al. (2013) highlights the challenges posed by these practices. Detecting paraphrases and rephrases requires distinct approaches, although they are semantically related (Harris, 1957; Martin, 1976; Duclaye, 2003).

To overcome these challenges, alternative methods can be explored:

- Stylometric Approaches, this method employs statistical techniques to analyze various aspects of writing style, focusing on features such as word frequencies, sentence lengths, punctuation usage, and syntactic structures. By quantifying these features, the method aims to capture unique patterns and characteristics specific to each author's writing style.
- Neural networks, this method achieves high performance in detecting plagiarism, especially in identifying paraphrases and subtle similarities within complex texts. It utilizes advanced techniques such as deep learning models, which have demonstrated superior capabilities in capturing intricate patterns and nuances in language.
- BERT is a pre-trained natural language processing (NLP) model developed by Google. It uses Transformer architecture and is trained on large unlabeled text corpora. BERT is designed to understand the context of words in a sentence by looking at both preceding and succeeding words, allowing it to capture nuances of meaning and context.

Methods for detecting paraphrastic rephrasing are common, using alignment methods (Callison-Burch et al., 2008) or more advanced techniques (Shen et al., 2006). The work of Fenoglio et al. (2007) emphasizes fundamental elementary transformations, while Mel'cuk's Sense-Text theory (1967) is often adopted.

These advancements in NLP and models like BERT contribute not only to the efficiency of plagiarism detection but also to a more nuanced understanding of language use.

4 Methodology

Our approach is to combine the traditional method, which is Direct Textual Comparison, with Natural Language Processing techniques such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) to provide a more robust and accurate approach.

We will proceed as follows:

Step Name	Description
Phase 1 Text Preprocessing: Tokenization	Using text cleaning techniques to eliminate irrelevant elements such as whitespaces, punctuation, etc. Tokenizing the texts to prepare them for processing by models.
Phase 2 Direct Textual Comparison	Using traditional methods such as string comparison to detect direct copy-pasting.
Phase 3 Semantic Analysis with BERT	Converting texts into embeddings (vector representations) using BERT and comparing the embeddings to evaluate semantic similarity between texts. If the embeddings are very similar, it could indicate paraphrasing or plagiarism
Phase 4 Generation and Comparison with GPT	Using GPT to rephrase one of the texts and comparing the rephrased text with the other. If GPT generates a text very similar to the other text, it could indicate plagiarism.
Phase 5 Stylometric Analysis	Using stylometric analysis techniques to compare the writing style of both texts. If the styles are very similar, it could indicate plagiarism, especially if the content is also similar.
Phase 6 Evaluation and Decision	Combining results from all methods to make a final decision on plagiarism. For example, if direct textual comparison, semantic analysis with BERT, and stylometric analysis all indicate plagiarism, you can be reasonably certain that the text is plagiarized.

Table 2: Different phases of designing the stages of the multi-level combination method.

By combining these models, we can leverage multiple architectures and achieve better results, obtaining superior performance compared to each individual model. However, this requires careful planning and implementation.

5 Evaluation

We assess the effectiveness of our plagiarism detection approach by applying various methods. The tests encompass diverse datasets containing authentic texts and examples of plagiarism with varying levels of complexity.

Evaluation Metrics: Precision, recall, and F-measure are employed for a balanced assessment of the model.

Test Dataset: Various texts representing different styles are utilized, including simulated cases of plagiarism to test the model's sensitivity.

Comparison with Other Methods: Our performance is compared to traditional methods and others.

Let's take a look at two text extracts::

- Text A: " Les avancées technologiques ont révolutionné notre quotidien."
- Text B: " Les progrès technologiques ont bouleversé la vie quotidienne."

For testing, we used two sentences in French, as it is the most widely used language for publishing articles or writing theses in Africa. However, it can also be used with various languages such as English and Spanish, as BERT and GPT already support multiple languages.

Method	Precision	Limitation
Textual Comparison	0.85	Limited to copy-paste cases, less effective on longer texts.
Semantic Analysis with BERT	0.92	High computational costs, requires large amounts of training data.
Generation and Comparison with GPT	0.88	It can generate text, unlike BERT, but it's not primarily designed for plagiarism detection and requires finesse in hyperparameter tuning
Stylometric Analysis	0.80	May be sensitive to intentional stylistic variations.

Table 3: Results of the approaches used

To obtain the result, we followed a rigorous evaluation procedure using diverse datasets, including authentic texts and plagiarism examples of varying levels of complexity. Each method was evaluated based on its precision performance, taking into account its specific advantages and limitations.

The textual comparison method was applied to authentic texts and copy-paste cases, evaluating accuracy and identifying limitations on lengthy texts. Semantic analysis with BERT converted texts into embeddings, measuring semantic similarity with paraphrase examples while assessing computational costs.

Generation and comparison with GPT involved rephrasing a text and adjusting hyperparameters, evaluating accuracy and detecting creative similarities. Stylometric analysis assessed writing style with tests sensitive to stylistic variations, measuring accuracy.

The overall process encompassed a comparison of the performance of each method, identifying and analyzing the limitations of each approach for a comprehensive evaluation.

Following the evaluation of these results, it was observed that the plagiarism detection approach combined with the natural language processing methods Bert and GPT reflects effectiveness in several key aspects: improved accuracy, detection of complex plagiarism patterns, scalability, and generalization.

6 Conclusion

Our innovative semantic approach, integrating BERT and GPT, has demonstrated increased effectiveness in detecting various forms of plagiarism, including subtle paraphrasing. Despite challenges related to complexity and computational costs, significant benefits, such as accurately detecting paraphrased content, make our model a promising solution for meeting the requirements of plagiarism detection in diverse digital contexts. Moreover, our approach is designed to be easily implementable, utilizing programming languages compatible with OpenAI's library and BERT. Ongoing research is necessary to optimize the model and explore emerging domains, underscoring our commitment to evolving plagiarism detection tools and preserving the integrity of digital content.

7 References

- Vandendorpe, C. (1992). *Le plagiat*.
- Hambi El Mostafa, Faozia Benabbou, El Habib Ben LahMar. (2016, June). *Comparaison Des Techniques De Détection Du Plagiat Académique*.
- Ratianantitra, V. M. (2023, December). A State of the art review on Natural Language Processing applied to the Malagasy Language. In *International Conference on Artificial Intelligence and its Applications* (pp. 1-5).
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681-694.
- Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4), 917-947.
- Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. *Language*, 33(3), 283-340.
- Fenoglio, I., Lebrave, J. L., & Ganascia, J. G. (2007). *EDITE MEDITE: un logiciel de comparaison de versions*.
En ligne: <http://www.item.ens.fr/index.php>.
- Martin, R. (1976). *Inférence, antonymie et paraphrase: éléments pour une théorie sémantique*.
- Duclaye, F. (2003). *Apprentissage automatique de relations d'équivalence sémantique à partir du Web* (Doctoral dissertation, Télécom ParisTech).
- Callison-Burch, C., Cohn, T., & Lapata, M. (2008, August). Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 97-104).
- Shen, S., Radev, D., Patel, A., & Erkan, G. (2006, July). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 747-754).

Creating datasets for emergent contact languages preservation

Dalmo Buzato

Faculty of Letters
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
dalmobuzato@ufmg.br

Átila Vital

Faculty of Letters
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
atilavital@ufmg.br

Abstract

The Venezuelan socioeconomic crisis increased the immigration process in Latin America. In Brazil, the Warao ethnic group, from Northeast Venezuela, has arrived in search of jobs and better social conditions, speaking a homonymous isolated language and mostly having Spanish as a second language. The communities have contact with the Brazilian Portuguese, intensifying the possibility for the appearance of an emergent contact language. This paper presents a dataset for the description and preservation of that emergent language. Based on previous works about multimodal data compilation, the dataset will be fed with written and spoken texts, sociolinguistic information, and morphosyntactic annotation. As soon as possible, it will be freely available for web consultation, following precepts of the Open Science Framework and the Digital Humanities paradigm.

1 Introduction

Language contact occurs when speakers of different languages interact with each other in communicative situations. Depending on social variables, such as the intensity of the contact, the prestige position of the languages and speakers involved, and the need for a mutual means of communication, a contact language may emerge.

Not all language contact situations lead to the emergence of contact languages. In some cases, there is linguistic borrowing between the languages involved, and changes accumulated over several generations. There are also cases in which a language is used as a mediator of contact but is not necessarily a contact language, in which case we have what linguists call a *lingua franca*. For a systematic discussion of the various linguistic possibilities in a language contact situation and the differences between them, we cite [Holm \(2000\)](#) and [Matras \(2020\)](#).

Much has been discussed about the preservation and ecology of contact languages ([Mufwene, 2003](#)).

Many contact languages have an unstable status of existence, becoming extinct when the contact situation between speakers of different languages ends (such as business situations, migrations, etc.). In addition, when there is a creole language, i.e. when the contact language is the mother tongue of a generation of individuals in a contact ecology, this language usually suffers from low social prestige and is usually not taught in schools, with no other instruments of social stimulation (literature, media use, government use).

In addition to the extremely productive dialogue between corpus linguistics and contact linguistics ([Nagy, 2011](#); [Mello, 2014](#); [Adamou, 2016](#); [Léglise and Alby, 2016](#)), relevant discussions have emerged about the creation of corpora and the use of the web for language preservation and documentation. According to [Cunha \(2020\)](#), "in the face of the effective threat of disappearance that thousands of languages around the world are currently suffering, all instruments for the conservation of linguistic diversity must be explored¹". For the author, the internet has a paradoxical role in this context, because while it contributes to the dissemination of majority languages to more individuals and communities, it can also help to amplify the voice of minority language speakers.

Intending to document and preserve emerging contact languages, this study reports on the ongoing development of a dataset with spoken and written data produced by Venezuelan refugees in Brazil. Most of the data was produced by indigenous refugees of the Warao ethnic group, as we will detail later in the text.

We believe that this work falls within the field of Digital Humanities because it promotes the documentation and maintenance of a language through

¹Original text: "[...] diante da efetiva ameaça de desaparecimento que sofrem, na atualidade, milhares de línguas ao redor do mundo, todos os instrumentos para a conservação da diversidade linguística devem ser explorados."

digital resources. Much more than simply storing audio and text files in a digital database, this paper uses data collected on the web (mostly photographs and video interviews available on the internet produced by the news media) to document the linguistic variety that emerges when Venezuelan Warao refugees arrive in Brazil.

An almost countless amount of data is produced every day on the internet, whether in media outlets, on social networks, or on websites. This data, even though some of it is currently produced with the help of artificial intelligence, is extremely valuable to linguists because it allows access to an exorbitant amount of data full of linguistic phenomena in an accessible and relatively simple way.

Another justification for the development of this paper lies in the very nature of contact languages. Many of them, including extinct ones, have little documentation as they are primarily transmitted orally and have an unstable survival status, such as pidgins, which emerge as emergency languages. Furthermore, if there is significant pressure for social integration, succeeding generations of contact language speakers may abandon it or incorporate various elements from the prestigious language in a process of language planning.

The subsequent sections will be organized as follows: in the upcoming section, we will provide a concise introduction to the Warao migration to Brazil and the language contact resulting from this migration. Sections 3 and 4 will elucidate the methodological details of the dataset, encompassing the nature and processes involved in storing, transcribing, and annotating both the written and spoken data. The former will include a brief description of the Universal Dependencies framework and its use for annotating linguistic phenomena. The oral data section will present a brief description of the C-ORAL-BRASIL's (Raso and Mello, 2012) transcription criteria, positioned before the audio section of this work. Section 5 will outline potential linguistic phenomena discerned during the previous analysis, while Section 6 will serve as the concluding remarks on the future of the dataset.

2 Warao migration to Brazil

Since the migration of Venezuelans to Brazil began in mid-2017, this phenomenon has been documented by researchers in law, anthropology, sociology, and linguistics. Since the first records, the presence of indigenous refugees has been noted,

mainly from the Warao ethnic group.

Research in linguistics has emerged since the beginning of the migration and has mostly been concentrated in the field of applied linguistics, such as in the areas of language policy and foreign language teaching and learning.

Research into language contact has emerged very recently, mainly analyzing the written productions of Venezuelan refugees asking for help, examples of research taking this approach are Mesquita (2020); Buzato and Vital (2023); Buzato (2023).

Points of relevance for research into language contact in the case of Venezuelan indigenous migration to Brazil is the fact that the Warao are speakers of a homonymous native language with no known linguistic relatives as L1, and are speakers of Spanish as L2 at different levels of proficiency, with a significant percentage of migrants having a low level of schooling and literacy.

In addition, according to anthropological studies (UNHCR, 2021; Soneghetti, 2017), the Warao were not a people with nomadic characteristics before their growing status of subalternity, which began with the loss of land for extractive activities in Venezuela, and with their migration to Brazil.

The refugees have not just stayed in the border regions between Brazil and Venezuela, or concentrated in the north of the country, which is closest to the neighboring country. On the contrary, they've moved inland and made long, independent journeys through towns and cities, always with the help of local citizens, to reach regions they believe are best for them to settle in.

For example, the distance between the city of Boa Vista, the capital of the Brazilian state of Roraima (the main initial concentration of refugees after leaving Venezuela), and the city of Belo Horizonte (where some of the photographs were taken) is over 3,000 kilometers, a route traveled independently by the refugees with their families and belongings.

3 Written signs and the dataset

This section will discuss some linguistic aspects of the written signs produced by the refugees, to ask the Brazilian population for help. As will be described below, the written dataset is of a mixed nature, with a percentage of photographs collected from news sites on the internet, and the other part of the photographs of the signs were collected by the researchers, since March 2022, in a fieldwork car-

ried out in the city of Belo Horizonte and metropolitan region, in the state of Minas Gerais, located in the southeast region of Brazilian territory.

Figure 1 below represents an example of a sign written by indigenous refugees. Although the signs collected in the city of Belo Horizonte in almost two years of fieldwork represent an important part of the data, we believe that the photographs collected from the web represent greater quality and representativeness of the phenomenon, since we have reported signs from 2018 to the present year 2024, and collected in several Brazilian cities of different population sizes and regions.

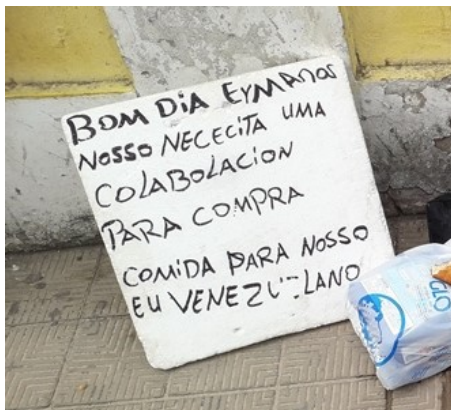


Figure 1: Example of a sign in the dataset

Currently, the photographs of the written signs, whether they come from the web or the researchers' fieldwork, are also transcribed and annotated according to the criteria of the Universal Dependencies (UD) project (Nivre et al., 2016). The transcriptions were made in a standard txt file, and the morphosyntactic annotations are in CoNLL-U format, the standard format of the UD project², as can be better elucidated in Buzato (2023). The choice of the UD framework for the written signs is based on its typological proposal and its growing use for annotating non-Indo-European and minority languages, in the spoken and written modalities of language.

Below is an example of how the sign shown in Figure 1 was transcribed. As we can see, no spelling or linguistic corrections have been made to the text produced since they can contain contact phenomena. Furthermore, due to the textual and writing context of the signs, as well as the socio-economic variables of the refugees, most of our signs do not have any graphic punctuation marks.

²<https://universaldependencies.org/format.html>

We also decided not to include them, as the absence of punctuation is an important aspect for our research.

Transcription: bom dia ermanos nosso
nececita uma colabolacion para compra
comida para nosso
eu venezuelano

3.1 Universal dependencies (UD) and language contact phenomena

The UD framework is increasingly developing treebanks to document contact languages and varieties. Currently, it has treebanks of Creole languages and varieties of spoken and written code-switching, derived from diverse texts such as comments on websites (Seddah et al., 2020) or radio interviews (Braggaar and van der Goot, 2021). However, the documentation of pidgin or mixed languages is still underdeveloped. A proposal was recently presented by Buzato (2023) whose annotations will form part of the dataset described here.

The presence of minority/low-resource languages is essential for any typological project, which certainly includes varieties emerging from language contact, especially in initiatives that promote the use of computational tools for typological analyses and the use of large amounts of data from different languages to improve models and tasks in computational linguistics and natural language processing. For this reason, documenting the variety presented here employing UD also contributes to the framework's objective and explores its potential for morphosyntactic annotation.

As can be seen in the annotation guidelines of UD for phenomena of foreign expressions and code-switching³, it essentially covers lexical borrowing and code-switching phenomena. These phenomena are typically considered to emerge from language contacts of lesser intensity between two or more communities. The annotation of these phenomena depends on the nature of the corpus (if it is a code-switched corpus or a monolingual one), and which specific phenomenon is under consideration. In such instances, multilingual material is annotated in features like Lang (language), Foreign, and OrigLang (Original language).

Since the annotation methods mentioned above do not encompass the phenomena found in our corpus, we have decided not to fully adopt them. In the example in Figure 1, for instance, we have

³<https://universaldependencies.org/foreign.html>

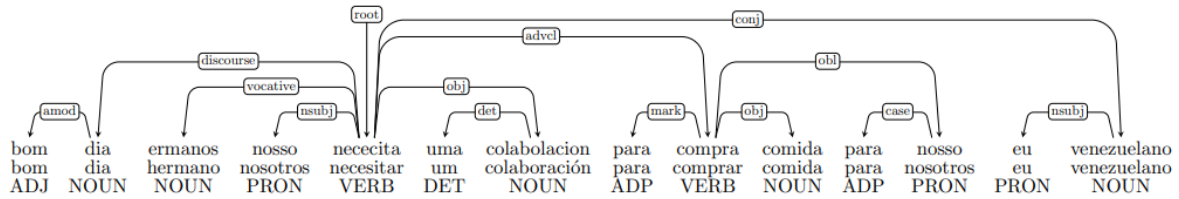


Figure 2: Example of a passage present in the written corpus annotated according to the UD framework

words like 'ermanos' borrowed from the Spanish 'hermanos', 'colabolacion' borrowed from 'colaboración', and 'nosso' borrowed for 'nosotros'. As a regular occurrence in mixed languages, there is a more productive blend between the repertoires of the languages in contact. Unlike borrowing, an element is never precisely derived from a language, there are often innovations (addition or loss) present in the linguistic element. Taking into account the aspects presented above, Figure 2 illustrates how the content presented in Figure 1 was annotated following the guidelines of UD.

3.2 Future steps on written section

We know, for example, of the existence of computer tools that allow multimodal texts to be annotated synchronously with the image, which is widely used in research into historical linguistics, such as the TEITOK tool (Janssen, 2016). Furthermore, the use of this tool is increasingly common for documenting minority or low-resourced languages such as Judaeo-Spanish (Quintana, 2020) and Galician (Blanco and Seoane, 2020), or varieties of languages spoken by second language (L2) speakers. This is a work in progress and certainly, one of the next steps in the project, which involves annotating and documenting the written texts produced by the refugees, most of which are multimodal texts - recorded by photographs, includes the use of TEITOK to unify the photograph and the linguistic annotation made by us.

4 The audio section

As a multimodal dataset, there are planned 20 recordings and transcriptions of spontaneous speech. In the first step of the audio compilations, it is important to perform tests to elaborate specific transcription criteria. The basic methodology to be applied is the same used in the C-ORAL-BRASIL's project (Raso and Mello, 2012), adapted by essential changes that will be elaborated in terms of the potential grammaticalization phenomena to be represented in the dataset.

4.1 A brief look to the C-ORAL-BRASIL's criteria

One of the most important aspects when dealing with the speech is the package of information conveyed by the prosody (Izre'el et al., 2020). Unlike written texts, the orality does not use punctuation to integrate the discourse into its morphosyntax parameters. Much more than that, through the combination of the fundamental frequency, length, and intensity, the spoken discourse integrates form and function, indicating the way the morphosyntax and the semantic/pragmatics relationship work across the sequence of words (González Ledesma et al., 2004; Raso and Mello, 2012).

Based on prosodic parameters, the transcriptions criteria used in the C-ORAL's corpora follow the segmentation of speech flow in terms of utterances and tonal units. In other words, the utterance is considered the minimal unit of the speech that conveys a complete communicative function (Izre'el et al., 2020). Along with that, two types of prosodic brakes (terminal and non-terminal brakes) give us perceptual clues about the compositionality and the non-compositionality of linguistic sequences.

The transcription criteria were adapted to the spontaneous spoken Brazilian Portuguese based on the C-ORAL-ROM's (Cresti and Moneglia, 2005). To provide consistent guidelines for the transcription crew, the authors organized several pilot studies. Those studies helped the establishment of a series of semi-orthographic criteria capable of capturing cliticizations, apheretic forms, erasing of verbal morphology, new pronominal paradigms, disfluencies, and many others. If the transcriptions followed purely orthographic criteria, many relevant lexical and grammatical phenomena would be hidden for future research.

In 1, there is an example of utterance recorded by C-ORAL-BRASIL I. The double bars “//” indicate the end of an utterance. Simple bars “/” indicate intonational units that do not convey a complete communicative function.

Example 1 (bfamnn06)

JOR: aonde a gente tem muito problema de liquidez / até em empresas que têm / &he / formação de família / na segunda pa terceira geração / já começa a dar problema e &f [2] e [1] e fecha //

Considering the prescriptive orthography, the utterance in 1 has lots of particularities. In a brief look at it, we can identify errors of pronunciation in the word “problema”, which would be written “problema” according to the grammatical prescriptivism. The choice to represent faithfully the way the speaker spoke is important for studies in variational linguistics that have been done with Brazilian Portuguese. Additionally, the phonetic contraction of the preposition “pra” (pronounced and transcribed as “pa”) can reveal the complex topic of prepositions and their forms in romance languages.

Just like the mispronunciation, the transcription developed for the C-ORAL-BRASIL represents disfluencies (self-corrections and fragmented words) and time-taking units (entire conversational turns with only “he” and “uhn”). The letter “&” represents a filled pause (&he) or an incomplete word (&f). The mark “[/n]”, in which “n” is a natural number, represents a retracting, a very common disfluency in spontaneous speech, a.k.a. self-correction; it happens when the speaker produces a word or a part of it and immediately corrects himself. The number inside the brackets means the number of canceled words (Raso and Mello, 2012).

4.2 First application of the transcriptions criteria to the immigrants’ spoken language

The first applications of the transcriptions give us important inferences about the richness of linguistic phenomena presented in a new spoken dataset. The goal of this subsection is to validate the conventions created for C-ORAL-BRASIL to the application in the emergent immigrant’s language. To do this, there were transcribed some audio parts collected from Warao’s documentary available on YouTube.

Example 2 (documentary_VAR)

VAR: lá / passava muita / dificuldade / por falta de / &m [1] da medicamento / porque / muita [1] muita criança // &he / muito / homem / mulher / vovó / &fa [1] faleciam / porque / faltava de [1] de medicamento lá // si / mas na [1] na mi [1] alimentação / não nos chega //

porque / indígena não [/1] não mora nas cidades / não mora na montanha // sim // então / lá não não chega médico / não [/1] &n não é possible / que [1] que médico chegue lá //

With the transcription, there will be available the header’s file, which compile possible sociolinguistic information, comments about the transcription, and conventionalized forms. In some moments of 2, we find Portuguese and Spanish lexical combinations (“si” and “possible”). It was considered important to represent those words in the way they were spoken with the appropriate comments in the header, as follows. The layout was inspired in the C-ORAL-BRASIL’s corpus as well.

@Title: documentary_VAR

@File: VAR

@Participants: VAR, John Vargas (male, unknown, unknown, Warao immigrant, participant, Venezuela)

@Date: unknown

@Place: Belo Horizonte (MG)

@Situation: documentary made by "Jornal o Tempo" about the Warao immigration @Topic: the life in Venezuela and the reasons why his family came to Brazil

@Source: YouTube

@Length: 39"

@Words: 64

@Transcriber: Átila Vital

@Comments: The audio has a music in a very low volume from the documentary

1) Forms originated by contact: at 10", VAR speaks "bobó", instead of "vovó" (grandmother). At 36", VAR speaks "possible", instead of "possível" (possible).

2) External noises: in some moments, there are sounds of children playing.

During the audio compilation, we will value high-quality recordings. That makes possible Phonetic and Prosodic investigations. The example 3 shows an utterance with glottalization and particular morphosyntax.

Example 3 (documentary_AAA)

AAA: ficar no Brasil / é muito mais bem
//

The figure 3 shows the waveform and the spectrogram of 3. The high acoustic quality is rare to be found in emergent language descriptions.

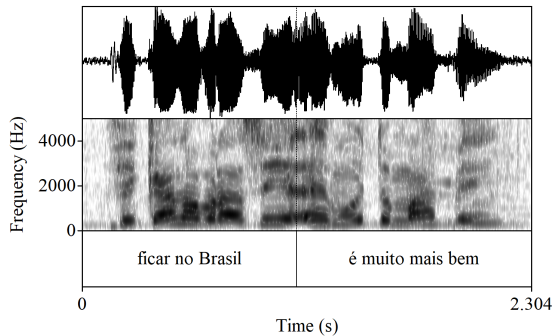


Figure 3: Waveform and spectrogram of example 3.

When building the dataset, all audio files and their respective text-sound alignments will be made available, in addition to the speakers' metadata txt files.

5 Potential linguistic phenomena

Preliminary descriptions of the Warao's signs have been made in previous works (Buzato and Vital, 2023; Buzato, 2023). Even with the objection that writing is not the primary form of emergence of a linguistic system, the proposals initially outlined aim to reflect on the structural particularities of the language used by immigrants.

In addition to the characteristics of the written representation - which seems to resemble speech (Figure 1) - and the constant borrowings from Spanish and Portuguese, some occurrences catch our attention. Firstly, there is a recurring confusion between the use of the adjective related to Venezuela (Venezuelan) and the name of the country itself. It is not uncommon to find data like "eu sou da venezuelano" ("I am from Venezuelan") or "eu (sou) venezuela" ("I (am) Venezuela"). Until now, there is no sufficiently structured data to verify the co-occurrence of these structures with prepositions or specific syntactic positions.

Another syntactically important phenomenon to be punctuated is the use of the copula. Romero-Figueroa (1997) points out that the Warao language allows the optional use of the verb copula in the expression of properties of nominal entities. The reuse of Warao's syntactic structure is what may

explain the absence of a linking verb in Figure 1, given the juxtaposition between the pronoun "eu" ("I") and the adjective "venezuelano" ("Venezuelan"). On the contrary, reflections of the confusion between the linguistic structures of the Spanish language and the Portuguese language are also found on the signs. An example is the use of an accusative pronoun postposed to the verb, as in "ajuda me" ("help me"), a less frequent form in Brazilian Portuguese, which favors the preposition "me ajuda" ("help me").

These are just some initial structural notes from the studies carried out with the signs. In the case of the audio files, we hope to confirm the data coming from the signs and describe even more contact phenomena.

6 Conclusions and future and the dataset

This is preliminary work towards the construction and availability of a dataset of an emerging contact language. Our initial objective is to contain around 60 transcribed and annotated signs, and 20 recordings of spontaneous speech, totalling approximately 1,500 words. All of them will be transcribed, segmented and aligned.

The linguistic description through immigrant signs is not common to be found in literature. Still, together with the collection of spontaneous speech data from Warao refugees, the data that will be accumulated and publicized could open the way for new methodologies in the study of emerging languages. We believe that, in the case of languages that emerge during migration crises, signs and writings asking for help may represent the only registers of the emerging forms. Both the development of methodologies and databases are welcome at this time.

At an opportune moment, when we have the first spoken and written data transcribed, annotated, and reviewed, the multimodal dataset will be freely available for web consultation.

Our goal is to document other emerging contact languages through the above protocols, using spoken and written data, mainly in low-resourced varieties in the global south. Furthermore, already extinct contact varieties, such as pidgin or mixed languages, can be transcribed and annotated using the same protocols, thus providing the creation of a set of multilingual datasets of emerging contact languages.

References

- Evangelia Adamou. 2016. *A corpus-driven approach to language contact: Endangered languages in a comparative perspective*, volume 12. Walter de Gruyter GmbH & Co KG.
- Rosario Álvarez Blanco and Ernesto Xosé González Seoane. 2020. *Calen barbas, falen cartas: A escrita en galego na Idade Moderna*. Consello da Cultura Galega.
- Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58.
- Dalmo Buzato. 2023. Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 509–519.
- Dalmo Buzato and Átila Vital. 2023. O contato linguístico em placas de refugiados venezuelanos em Belo Horizonte e região metropolitana: observações preliminares. In *Anais do Congresso Nacional Universidade, EAD e Software Livre*, volume 1.
- Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM: integrated reference corpora for spoken romance languages*. John Benjamins Publishing.
- Evandro L T P Cunha. 2020. A web como ferramenta de suporte à preservação e à revitalização linguística. *Cadernos de Linguística*, 1(3):01–14.
- Ana González Ledesma, Guillermo De la Madrid, Manuel Alcántara Plá, R De la Torre, and Antonio Moreno-Sandoval. 2004. Orality and difficulties in the transcription of spoken corpora. In *Proceedings of the Workshop on Compiling and Processing Spoken Language Corpora, LREC*.
- John Holm. 2000. *An introduction to pidgins and creoles*. Cambridge University Press.
- Shlomo Izre'el, Tommaso Raso, Alessandro Panunzi, and Heliana Mello. 2020. In search of basic units of spoken language. In *Search of Basic Units of Spoken Language*, pages 1–452.
- Maarten Janssen. 2016. Teitok: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043.
- Isabelle Léglièze and Sophie Alby. 2016. Plurilingual corpora and polylinguaging, where corpus linguistics meets contact linguistics. *Sociolinguistic studies*, 10(3):357–381.
- Yaron Matras. 2020. *Language contact*. Cambridge University Press.
- Heliana Mello. 2014. What Corpus Linguistics can offer Contact Linguistics: the c-oral-brasil corpus experience. *PAPIA: Revista Brasileira de Estudos do Contato Linguístico*, pages 407–427.
- Rodrigo Mesquita. 2020. Diaria o fixo: fotografias sociolinguísticas de Boa Vista–Roraima e as novas perspectivas para as pesquisas do contato linguístico na fronteira. In A. Cruz and F. Aleixo, editors, *Roraima entre línguas: contatos linguísticos no universo da tríplice fronteira do extremo-norte brasileiro*. Editora da UFRR.
- Salikoko S Mufwene. 2003. Language endangerment: What have pride and prestige got to do with it. *When languages collide: Perspectives on language conflict, language competition, and language coexistence*, pages 324–346.
- Naomi Nagy. 2011. A multilingual corpus to explore variation in language contact situations. *RILA : Rassegna Italiana di Linguistica Applicata*, pages 65–84.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.
- Aldina Quintana. 2020. CoDiAJe—the Annotated Diachronic Corpus of Judeo-spanish. *Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques*, (9):209–236.
- Tommaso Raso and Heliana Mello. 2012. *O Corpus C-ORAL-BRASIL*. Editora UFMG, Belo Horizonte.
- Andrés Romero-Figueroa. 1997. *A Reference Grammar of Warao*. Lincom Europa, München.
- Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futral, Benjamin Muller, Pedro Ortiz Suarez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1139–1150.
- Pedro Moutinho Costa Soneghetti. 2017. Parecer Técnico acerca da situação dos indígenas das da etnia Warao na cidade de Manaus, provenientes da região do delta do Orinoco, na Venezuela. Technical report, Procuradoria Geral da República/AM.
- UNHCR. 2021. Os Warao no Brasil - Contribuições da antropologia para a proteção de indígenas refugiados e migrantes. Technical report, Brasília.

Psychoanalytic Studies in the Digital Humanities: Employing Topic Modeling with an LLM to Decode Dreams During the Brazilian Pandemic

João Pedro Campos

Universidade Federal
de Minas Gerais
Graduate Program in Electrical Engineering
jpafcampos@gmail.com

Natalia Resende

Trinity College Dublin
School of Languages,
Literatures and Cultural Studies
resenden@tcd.ie

Ricardo de Souza

Universidade Federal
de Minas Gerais/CNPq
Faculty of Languages, Literature,
and Linguistics
ricsouza.ufmg@gmail.com

Gilson Iannini

Universidade Federal
de Minas Gerais
Department of Psychology
gilsoniannini@yahoo.com.br

Abstract

This paper reports on an interdisciplinary project involving the compilation a corpus of dream reports collected over the period between March of 2020 and March of 2021. The corpus encompasses narratives of images and scenes dreamed during the initial months of intense anxiety of the SARS-Cov-2 pandemic. The pandemic dreams corpus originated in a practicum in psychoanalytic clinic for Psychology majors at University sersidade Federal de Minas Gerais. In response to social distance security requirements, the critical data was compiled through online media. We aim to discuss the possibilities opened by topic modeling as a way to gather insights on a valuable corpus, bridging the gap between different areas to create a digital humanities trans-disciplinary endeavor.

1 Introduction

We report partial findings of an interdisciplinary project involving the compilation and analysis of a corpus of dreams collected over the period between March of 2020 and March of 2021. The corpus encompasses narratives of images and scenes dreamed during the initial months of the SARS-Cov-2 pandemic. The pandemic dreams corpus originated in a practicum in psychoanalytic clinic for Psychology majors at the Universidade Federal de Minas Gerais (Brazil), which moved online in response to social distance security requirements.

This particular historical moment established the cultural background in which a set of socially-shared representations might appear as dream imagery, therefore allowing for common references to recur in the highly individual mental activity of dreaming.

This study leverages the topic modeling capabilities of the GPT-3.5 model for the analysis of dreams during the SARS-Cov-2 pandemic. The cutting-edge NLP technology of GPT-3.5, through its topic modelling capabilities, provides a unique opportunity to explore the unconscious mind's response to unprecedented global events, as manifested in dreams. Our research is guided by three pivotal questions: Firstly, we aim to assess the degree of alignment between the GPT-3.5 model's interpretations through topics and the participants' own feelings about the pandemic, as indicated in their responses and the interpretations of their dreams. This comparison seeks to understand the extent to which an AI generated topic is aligned with subjects' dream interpretation and self-reported emotional states.

Secondly, we intend to explore what deeper insights generative AI can uncover about latent feelings embedded within dream narratives. This involves probing beyond the surface level of dream reports to reveal subtler, perhaps unconsciously held emotions and thoughts. Here we will test whether the topic modeling capabilities of GPT-3.5 are particularly suited for this task. Finally, we

are interested in investigating whether the topics generated by the GPT-3.5 model reflect feelings or thoughts that are not explicitly reported by the dreamers. This aspect of the study could reveal the potential of AI in offering alternative perspectives or unearthing hidden dimensions of psychological experiences. Such findings would not only contribute to our understanding of dream analysis in the context of the pandemic but also demonstrate the utility and potential of AI in psychoanalytic research. By addressing these research questions, we aim to bridge the gap between traditional psychoanalytic dream interpretation and modern AI-driven linguistic analysis, offering new insights into the human unconscious under the strain of a global crisis.

2 Theoretical background

For psychoanalysis, dreams are invaluable manifestations of mental life. Such manifestation has a central role in the clinical approach of diverse psychoanalytic schools (Freud, 1997; Khan, 1962; Quinodoz, 2005; Ogden, 2018), and this clinical relevance is equally recognized in research that attempts to bridge the neuroscientific hypothesis that dreaming activity reflects information integrative processing by the sleeping brain and the psychoanalytic viewpoint of dreams as the royal road to the unconscious mind (Fonagy et al., 2018; Zhang and Guo, 2018). Dreams have also been a cornerstone of the development of psychoanalytic theory. The psychological processes of condensation and displacement described as the fundamentals of dreamwork in Sigmund Freud's masterpiece *The Interpretation of Dreams* (Freud, 1997), which are acknowledged as the primary processes of unconscious functioning, have expanded into key concepts of an overarching theory of mental activity. These concepts account for an array of phenomena that includes slips, jokes, free associations guiding psychoanalysis patients' discourses, and neurotic symptoms (Freud, 1916). All such phenomena play roles both in psychoanalytic therapy and in the Freudian conception of the psychic apparatus.

Dreams are an inherently individual experience. The Freudian approach to dream interpretation departs completely from any sort of listed interpretive keys linking fixed meanings to oneiric images. Instead, this approach focuses on the emergent meanings that derive from analysands' free associations as they tell their dreams in the psychoanalytic

setting. This individuated and subjectivity-bound method for interpretation notwithstanding, dream contents can often be nested in shared cultural references. This was fully acknowledged in Freud's presentation of dream theory as a discussion of symbolism in dreamwork (Freud, 1916). The symbolic images and their hypothesized usual meanings discussed in Freud's work are clearly derived from folklore, therefore being construed as borrowings from materials such as traditional chants and widespread sayings. In other words, the symbolic imagery and their somewhat predictable meanings that may be recurrent in dreams by more than one dreamer come from collectively shared discourses belonging to Freud's milieu, images and associated meanings that were passed from individual to individual through language and common parlance.

It should be noted that symbols as relevant components of dreams have long been an anathema for practicing psychoanalysts. Discussion of symbolic meanings is often framed as precisely the type of interpretive key that is stereotypically construed as a dream dictionary, with equations of images (for example, an umbrella) with meanings (for example, evoked sexuality because of analogy with the male erected sexual organ). This type of cliché interpretation of fixed meanings runs counter the core of psychoanalytic dream theory, as well as being at odds with the practice of dream analyses in psychoanalytic clinical work. In both theory and clinical practice, it is the dreamer's associations that spring from the reported dream contents, along with the very fact that that dreamer has chosen to report that dream in the discourse he or she directs to that analyst at that point in the progress of his or her analysis, that are of paramount value for the construal of significance of a dream in a given subject's psychological experience. This has led Freud's chapter-long discussion of dream symbolism (Freud, 1916) to be an often-overlooked theme in present-day psychoanalytic thinking.

Speech and language are inseparable from psychoanalysis as either a theory of the human mental experience and as a clinical method (Forrester, 1980; Arrivé, 1992; Dunker and Kupermann, 2016; Bonfiglio, 2023). Language and linguistic theory are instrumental in Jacques Lacan's contributions to psychoanalysis, especially regarding his critical assessment of post-Freudian trends in psychoanalytic thinking and practices (Lacan et al., 2020). Lacan's proposal that the unconscious is structured as language actually sets psychoanalysis free from

certain commitments to localist understandings of the Freudian psychic apparatus, as expressed by the imaginary conception of the unconscious as some sort of cellar where undesirable or emotionally painful memories are forcefully kept. Within this perspective, the unconscious is an effect of language and its polysemy and ambiguity, of its imposing, inescapable structure but also of its paradoxical inadequacy as a fully transparent means of expression. Language is the sole material of psychoanalysis, as clearly stated by Lacan: “Whether it wishes to be an agent of healing, training, or sounding the depths, psychoanalysis has but one medium: the patient’s speech.” (Lacan et al., 2020). It is perfectly natural that the grammar of dreams is linguistically bound, as Lacan demonstrates by equating displacement with metonymy and condensation with metaphor.

Cultural and linguistic grounding of dream contents opens the possibility of a rekindled interest in symbolic meanings for psychoanalytic dream theory. We assume that such a renewed interest may depend on studies of large databanks of texts reporting dreams. We further assume that AI generated pattern analysis of texts of dream reports may provide a cognitive model for the recognition of socioculturally motivated references in dreams. The present study aims at probing precisely this hypothesis.

3 Related work

A notable trend in recent studies is the exploration of the pandemic’s psychological repercussions, not only on conscious behavior and mental health (e.g. (Borghi et al., 2021; Vindegaard and Benros, 2020), but also on the unconscious processing of these circumstances, as reflected in dreams. In this realm, Natural Language Processing (NLP) emerges as a pivotal tool. By applying NLP techniques to dream narratives, researchers are uncovering the intricate ways in which the pandemic influences our unconscious minds. This approach aligns with the interdisciplinary nature of the present study, marrying psychoanalytic insights with topic modeling and generative AI. For instance, (Mota et al., 2020) employed various NLP techniques, such as the analysis of emotional word proportions within dream reports, verbosity, the presence of anger or sadness-related terms, and semantic similarities to words like "contamination" and "cleanness." Their study aims to provide insights into the semantic

and emotional features of dreams both before and during the pandemic, allowing for a comparison of their dissimilarities. Their findings revealed that pandemic dreams exhibit a higher proportion of words associated with anger and sadness than other words, as well as greater average semantic similarities to terms like "contamination" and "cleanness."

(Šćepanović et al., 2022) investigated whether the content of dreams during the pandemic is consistent with the dreamers’ waking experiences. To this end, they used a recurrent neural network designed to extract mentions of any medical conditions and health-related phrases from free-form text. This method is applied to two datasets collected during the pandemic: 2888 dream reports (reflecting dreaming life experiences) and 57 million tweets (representing waking life experiences) that mentioned the pandemic. The common health expressions found in both sets predominantly revolved around typical COVID-19 symptoms, such as cough, fever, and anxiety. This observation suggested that dreams indeed mirrored people’s real-world encounters.

More recently, (Barrett, 2023) employed a deep learning algorithm to distinguish between discussions about COVID-19 in waking life conversations and dreams reported during the pandemic. These studies collectively highlight the efficacy of NLP techniques in aiding researchers and mental health professionals in comprehending the unconscious processes that unfold during crises like the COVID-19 pandemic. Moreover, they serve as valuable tools for testing hypotheses, including Freud’s day-residue hypothesis, which posits that elements experienced during the preceding days can be identified through careful scrutiny of dreams.

However, to the best of our knowledge, this paper represents a pioneering effort to investigate the insights that topic modeling using generative AI can provide for understanding the underlying emotions at play during the pandemic by analyzing dream reports. Below we report our methodological approach.

4 Methodology

4.1 Collecting dreams during COVID-19 Pandemic

The data was collected by groups of researchers from the universities UFMG, USP and UFRGS. The collection resulted in a database with 1158 dream narratives. The collection was carried out

through a written form, with open-ended questions with no character limit for the dream-related fields, and closed questions regarding each participant's experience with the pandemic in general, as well as basic demographic information (e.g., race, gender, age, occupation, etc.)

The form contained 4 open-ended questions about the dream, described by the following requests: (1) *Report your dream. Try to tell what you remember. Write freely;* (2) *Do you remember anything you thought, saw, heard, read, and/or experienced on the day(s) before the night of the dream that may be related to the dream and that you think is important to report?;* (3) *Try to jot down what is going through your mind right now, even if it is unrelated to the dream;* (4) *How do you understand, interpret, or explain this dream?*

The theoretical foundation for each of these questions is as follows: given the deterministic nature of the unconscious and the dream as a less indirect access to the unconscious, the meaning of the dream is contained not only in the dreamer's own report (1) but also depends on what Freud called "day residues" (Freud, 1997) (2) or memories and post-dream free associations. Additionally, we added a question guided by the ethics of psychoanalysis, according to which the dreamer is also the interpreter of their own dream (3). In addition to these long-text open-ended questions (narrative or descriptive), each participant was asked to describe their thoughts or feelings about the pandemic in 3 to 5 words (question: *Write down 3-5 words that describe your thoughts or feelings related to the pandemic*). The idea behind this request was to detect conscious thoughts or feelings of the subjects, in order to verify their correlation or not with the content of the dreams. Finally, various questions about the dreamer's life context were added, covering aspects such as whether the city they live in was in isolation or quarantine measures; to what extent the pandemic has affected the dreamer's routine; whether the participant works in essential services related to the pandemic; whether the participant or someone close was infected by the Sars-COV-2 virus; if there was any loss of relatives and friends, and so on. Additionally, in the following year, questions about vaccination were added. Among the participants, their gender is distributed as follows: 893 identify as female, 233 as male and 33 did not answer. The ages of participants range from 12 up to 73 years old, with a mean value of 33. Most participants (42%) are between 20 and 30 years

old.

We exemplify our corpus by reproducing two narratives below (translated by the authors):

1) *I dreamed that I was walking towards my parents' house. It was night, but I could see the street and the cars passing by clearly. I was naked, walking, talking to someone on the phone. This person was asking me for something, and I kept responding in a repetitive manner; "I'm naked in the street, do you know what it's like to be naked in the street? I need to go home, I can't help you." The person on the other end of the line laughed and kept asking for help. I hung up. I kept walking, and then the street that was leading me to my parents' house changed, leading me to my childhood home, also my parents'. I walked, afraid of being recognized and seen naked, until a man managed to locate me. With his gaze, he bothered me, and so, in the dream, I wished that he wouldn't touch me. That's when the dream turned dark, the streetlights turned off, everything turned pitch black, but I was still there, naked, yet calm, because of the impossibility of the man's gaze.* 2) *I was in the center of Belo Horizonte, near my grandmother, who is very active and always crosses the area several times a day. Then I realized I wasn't wearing a mask and no one else was either. Somehow I managed to walk back to Santa Luzia, where my friends and maternal family live. I ended up going towards a quarry with a very large and clean lake that doesn't exist there, where many people were having fun. There were arches and arrows all around the lake, but more on the shore. I entered the lake and my mother was there. I haven't seen her since all of this started and I left home. She asked me, "Why didn't you come back home? Do you want me to die?"*

4.2 Data Preprocessing

As mentioned in the previous section, participants were requested to freely write their dream reports and answers to the questions using an online form. As a result, these reports may contain non-standard words, including abbreviations, acronyms, slang, and unconventional orthography and punctuation. To address these potential challenges, we utilized the Python library ENELVO (Bertaglia and Nunes, 2016), designed for normalizing noisy words present in user-generated Portuguese content. This software utilizes data from a word embedding model to identify the appropriate standard Portuguese word to replace the noisy ones. Additionally, all numerical values written as digits were

omitted. For the analysis detailed in this paper, we opted to retain punctuation and stop words, enabling a more comprehensive analysis of the entire dream report by the AI tool.

This exact pipeline was applied to all columns in our collected dataset, meaning that the dream reports, interpretations and other responses were all submitted to the same transformations.

4.3 Topic Modeling Using Generative Pre-trained Transformers

This study employs a Large Language Model (LLM) for the task of topic modeling, specifically using the GPT-3.5 (text-davinci-003) model accessed via the OpenAI Python library which facilitated interaction with the GPT-3.5 model. A critical parameter in our methodology is the 'temperature' setting, which governs the model's generative creativity and the diversity of the resultant text. A high temperature parameter, such as 1.5, results in text that is varied and inventive. Conversely, a lower temperature (e.g. 0.5) yields text that is more predictable and concentrated. Our aim was to generate deterministically relevant topics; thus, the temperature was set at 0.5 to mitigate the risk of deriving extraneous topics.

We processed 1158 dream narratives through the model, instructing it to identify a single, distinct topic for each dream. The prompt was phrased as follows:

"I am presenting you with a dream description from the COVID-19 pandemic. Identify, in one word, the broader category or the high-level topic of this dream. Please respond in Brazilian Portuguese."

Using this method, the GPT-3.5 model generated 324 unique topics. The most frequent topic, found in 204 out of 1158 dream narratives, was 'fear' (Portuguese: "medo"). Owing to its prevalence, we focused our analysis on the narratives categorized under "fear". This focus allowed us to explore if the model's topic of 'fear' accurately reflects the participants' emotional responses to their dreams when requested to answer the question *Write down from 3 a 5 words that better describe your thoughts or feelings related to the pandemic*. Furthermore, we extended our analysis by prompting the model to assign topics to the participants' interpretations of their dreams when they answered the question *How do you understand, interpret, or explain this dream?*. This step aimed to determine if the model consistently identified 'fear' or related topics in the

participants' own interpretations, thereby ascertaining whether the topic of 'fear' emerged from the model's interpretation of unconscious sentiments rather than mere keyword detection.

This comparative analysis was crucial in determining if the 'fear' topic assigned by the model represented an underlying emotion not explicitly expressed by participants when describing their dream-related feelings. Additionally, it allows us to investigate the model's capability to uncover latent topics implicit within the dreams that were not explicitly identified by participants.

With this methodology, we aim to answer the following research questions:

- What is the degree of agreement and disagreement of the GPT model regarding participants' feelings about the pandemic ?
- What is the degree of agreement and disagreement of the GPT model regarding participants' interpretation of their dreams?
- What insights can the GPT reveal about latent feelings in dreams?
- Is the model capable of uncovering feelings or thoughts not explicitly reported by dreamers?

5 Results

As described in our methodology, GPT-3.5 model was employed to relate each dream report to a topic. 324 unique topics were assigned. Table 1 reports the five most frequent topics.

Topic	# of occurrences
Fear	204
Nightmare	30
Anxiety	27
Pandemic	23
Travel	20

Table 1: Top five topics assigned by GPT-3.5

In order to investigate the degree of agreement and disagreement of the GPT's topic regarding participants' feelings we analyzed participants' answers to the question *Write from 3 to 5 words that better describe your thoughts or feelings related to the pandemic*, i.e., the key words that the participants employed to describe how they feel about the pandemics, results show that the word *fear*

("medo") is the most frequent. Table 2 presents the five top frequent words in column (W).

Portuguese	English	# of occurrences
Medo	Fear	432
Ansiedade	Anxiety	257
Tristeza	Sadness	176
Angústia	Anguish	171
Incerteza	Uncertainty	90

Table 2: Most frequent words used by participants to describe their thoughts and feelings related to the pandemic

We restrain our analysis to the prevalent topic, by filtering our dataset to keep only the dreams under the topic 'fear'. Our objective was to explore the extent to which the model's most frequently assigned topic, which encompasses an emotional description ('fear'), aligns with participants' reported feelings. Specifically, we aimed to gauge the level of agreement between the model's identified topic, centered on the topic fear, and the participants' self-reported experiences expressed through keywords. Additionally, we investigated whether the narratives labeled as 'fear' by the model were explicitly linked to the word 'fear' within the dream narratives or whether they represented implicit inferences made by the model.

Our results show that, among the 204 corresponding dream narratives whose assigned topic was fear, we observe that only 26% contain the word fear explicitly, and only 33% of participants' interpretations in key words employed this word. Consequently, there was agreement between the model and participants in only 33% of the dataset.

Next, we investigated whether the reports and interpretations contained words in the same semantic field of 'fear'. To this end, we created the following list of Portuguese words: *angústia, pavor, terror, temor, amedrontado, assustado, apavorado, aterrorizado, amedrontado*, which can be approximately translated as *anguish, dread, terror, fear, intimidated, scared, horrified, terrified, frightened*. Results showed again that only 26% of reports contained any of these words. Consequently, it appears that the frequency of the 'fear' topic is not strongly associated with the semantic field of the word "fear."

However, it's worth noting that when considering the entire set of words participants used to

describe their pandemic-related thoughts and feelings, 'fear' remained the most frequent word (see 2. This observation led us to hypothesize that the higher frequency of the 'fear' topic could be influenced by the prevalence of the word 'fear' in the overall dataset of dream narratives to which the model was exposed. To further explore this hypothesis, we calculated the frequency of nouns in the entire dataset of dream narratives. 3 shows the results.

Nouns	# of occurrences
Dream	1122
House	966
People	767

Table 3: The most frequent nouns within the corpus of dreams narratives

The results revealed that the three most frequent nouns were "sonho" ('dream') with 1122 occurrences, followed by "casa" ('house') with 966 occurrences, and "pessoas" ('people') with 767 occurrences. Remarkably, the word fear ranked 51st in frequency, with only 185 occurrences. Consequently, it is clear that the contribution of the word 'fear' to the model's frequent assignment of the 'fear' topic is indeed marginal. The model's topic assignment is likely influenced by a range of factors beyond simple word frequency, necessitating further investigation to gain a deeper understanding of its behavior in this context. Interestingly, this finding suggests that the GPT model is capable of inferring the presence of the emotion 'fear' based on the situations described in participants' dream narratives, rather than solely relying on explicit keywords. Further, this finding suggests that the model can identify feelings that are not explicitly mentioned, which prompted us to explore whether the model could align with participants' own interpretations of their dreams as fear.

The rationale behind this additional analysis is the following: if the model consistently assigns the 'fear' topic to participants' interpretations of their dreams, it could potentially serve as a valuable tool for psychologists and psychoanalysts seeking to better understand emotions and thoughts that may not be explicitly articulated in dream narratives or when participants are queried about their feelings and thoughts related to their dreams or situations.

To investigate this, we tasked GPT with assigning topics to participants' interpretations of their

dreams. Once again, the 'fear' topic emerged as the most frequently assigned topic, occurring 24 times out of 110 unique topics. When we assessed the level of agreement between the 'fear' topic and participants' interpretation narratives, we observed that 18 participants used the word 'fear' in their dream interpretations, resulting in a substantial agreement rate of 75% between the model's assigned topic and the participants' own interpretations. In addition, we also found that the noun 'fear' was the most frequent noun within the entire set of participants' interpretation narratives.

6 Discussion and conclusions

The results of our analysis offer insights into the relationship between participants' self-reported feelings related to the COVID-19 pandemic and the topics assigned to their dream narratives by the GPT-3.5 model. Our study sought to unravel the complex interplay between the model's topic assignments and the emotional content expressed by participants in both explicit and, especially, implicit ways.

Initially, we observed that the word "fear" stood out as the most frequently employed term when participants described their thoughts or feelings in response to the pandemic. This finding highlights the prominence of fear as a prevailing emotional response during this period, aligning with previous research that has documented how heightened anxiety and apprehension during times of crisis tend to appear symbolically in oneiric activity (e.g.: [Beradt, 2022](#)).

However, our primary focus was on the 'fear' topic assigned by the model to dream narratives. Despite the word "fear" being a common theme in participants' pandemic-related vocabulary, we noted that the frequency of this word within the entire dataset of dream narratives was marginal, ranking 51st in terms of frequency of occurrence. This raises intriguing questions about the model's behavior in assigning topics. The GPT-3.5 model appears to go beyond simple word frequency when identifying the 'fear' topic within dream narratives, suggesting its capability to infer the presence of symbolic 'fear' based on the contextual dream situations narrated by participants. Furthermore, when we delved into the alignment between the model's 'fear' topic and participants' own interpretations of their dreams as fear, we discovered a notable agreement rate of 75%. This suggests that the model's

topic assignment is not only effective at identifying implicit emotions but also tends to converge with participants' subjective understanding of their own dream experiences.

This capability of the model to detect implicit emotions within dream narratives is a noteworthy finding. It implies that the model can identify feelings that are not directly mentioned, therefore being capable of abstracting an effective meaning that is pervasive in a corpus of dream narratives belonging to a given sociocultural context but that is only expressed by discourse as diverse as the highly idiosyncratic and subjective experience of dreaming when translated into text by dreamers. This capability might become a valuable tool for psychoanalysis theorists, and perhaps even to practicing psychoanalysts and clinical psychologists, as it may be a tool to re-address the issue of symbols in Freudian dream interpretation theory. Our observations reveal that emotional colors associated with a given shared context – in other words, a given zeitgeist – might be identified beyond the detection of key words even by an AI digital tool such as the one employed in our study. This finding is relevant for the psychoanalytic theory of dreams because it sheds new light on the concept of symbolic meanings, bypassing the simplistic and definitely non-psychoanalytic view of symbols as merely a collection of dream dictionary entries with fixed meanings, and rather pointing towards an understanding of oneiric symbols and recurrent meanings that emerge out of dreamers lived experience in a given time and place in social history. In this respect, we understand the present study to be an inviting example of the introduction of AI-assisted digital humanities in the theoretical debates of psychoanalysis.

Our study thus highlights the potential utility of GPT-3.5 and similar language models in psychoanalytic, psychological and even neurocognitive dream theory. These models seem to have the capacity to unearth underlying psychological representations, providing a nuanced perspective that may complement traditional self-reporting methods in large scale studies or studies that seek to foster generalized psychic architectures and processes in human dreaming.

However, it is important to acknowledge that there can be critical limitations in our approach. While the model's performance is promising, it may not capture the full richness and subtlety of human emotions. Additionally, further research is

needed to understand the model's behavior with dream narratives from different cultural and linguistic contexts, especially from periods of time not so clearly marked by a collective crisis such as the SARS-Cov-2 pandemic of 2020. It should also be very important to further analyze the model in order to find its possible shortcomings and potential biases. These are issues to be dealt with in further steps of our research.

Acknowledgements

Participation of the PhD candidate João Campos in this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for the Improvement of Higher Education Personnel) – Brasil (CAPES) – Finance Code 001.

Dr. de Souza's participation in this study was partially funded by grant 314681/2021-3, given by the Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development).

References

- Michel Arrivé. 1992. Linguistics and psychoanalysis. *Linguistics and Psychoanalysis*, pages 1–194.
- Deirdre Barrett. 2023. *Dreams and Nightmares During the COVID-19 Pandemic*, pages 295–308. Springer Nature Singapore, Singapore.
- Charlotte Beradt. 2022. *Sonhos no terceiro Reich*. Fósforo.
- Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2016. Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.
- Thomas Paul Bonfiglio. 2023. *Linguistics and Psychoanalysis: A New Perspective on Language Processing and Evolution*. Taylor & Francis.
- Lidia Borghi, Federica Bonazza, Giulia Lamiani, Alessandro Musetti, Tommaso Manari, Maria Filosa, Maria C Quattropiani, Vittorio Lenzo, Maria Francesca Freda, Daniela Lemmo, Emanuela Saita, Roberto Cattivelli, Gianluca Castelnovo, Elena Vegni, and Christian Franceschini. 2021. Dreaming during lockdown: a quali-quantitative analysis of the Italian population dreams during the first COVID-19 pandemic wave. *Res. Psychother. Psychopathol. Process Outcome*, 24(2):547.
- Christian Ingo Lenz Dunker and Daniel Kupermann. 2016. *Por que Lacan?* Zagodoni.
- Peter Fonagy, Horst Kachele, Marianne Leuzinger-Bohleber, and David Taylor. 2018. *The significance of dreams: bridging clinical and extraclinical research in psychoanalysis*. Routledge.
- John Forrester. 1980. *Language and the Origins of Psychoanalysis*. Springer.
- Sigmund Freud. 1916. *Introductory Lectures on Psycho-Analysis*, volume 15. The Standard Edition of the Complete Psychological Works of Sigmund Freud. Available at <https://pep-web.org/browse/document/SE.015.0000A?page=PR0004>.
- Sigmund Freud. 1997. *The Interpretation of Dreams*. Wordsworth Editions. Original work published in 1900.
- M Masud R Khan. 1962. Dream psychology and the evolution of the psycho-analytic situation. *The International Journal of Psycho-Analysis*, 43:21.
- Jacques Lacan, Alan Sheridan, and Malcolm Bowie. 2020. The function and field of speech and language in psychoanalysis. In *Écrits: A selection*, pages 33–125. Routledge.
- Natália Bezerra Mota, Janaina Weissheimer, Marina Ribeiro, Mizziara de Paiva, Juliana Avilla-Souza, Gabriela Simabucuru, Monica Frias Chaves, Lucas Cecchi, Jaime Cirne, Guillermo Cecchi, Cilene Rodrigues, Mauro Copelli, and Sidarta Ribeiro. 2020. [Dreaming during the covid-19 pandemic: Computational assessment of dream reports reveals mental suffering related to fear of contagion](#). *PLOS ONE*, 15(11):1–19.
- Thomas Ogden. 2018. *Conversations at the frontier of dreaming*. Routledge.
- Jean-Michel Quinodoz. 2005. *Dreams That Turn Over a Page: Paradoxical Dreams in Psychoanalysis*. Routledge.
- Sanja Šćepanović, Luca Maria Aiello, Deirdre Barrett, and Daniele Quercia. 2022. Epidemic dreams: dreaming about health during the COVID-19 pandemic. *R. Soc. Open Sci.*, 9(1):211080.
- Nina Vindegaard and Michael Eriksen Benros. 2020. COVID-19 pandemic and mental health consequences: Systematic review of the current evidence. 89:531–542.
- W. Zhang and B. Guo. 2018. [Freud's dream interpretation: A different perspective based on the self-organization theory of dreaming](#). *Frontiers in Psychology*, 9:1553.

Decoding Sentiments about Migration in Portuguese Political Manifestos (2011, 2015, 2019)

Erik Bran Marino

Universidade de Évora
erik.marino@uevora.pt

Renata Vieira

Universidade de Évora
renatav@uevora.pt

Jesus Manuel Benitez Baleato

Universidade de Santiago de Compostela
jesusmanuel.benitez.baleato@usc.gal

Ana Sofia Ribeiro

Universidade de Évora
asvribeiro@uevora.pt

Katarina Laken

Fondazione Bruno Kessler
alaken@fbk.eu

Abstract

This research conducts a mixed-method analysis of Portuguese political manifestos from 2011, 2015, and 2019, focusing on immigration discourse. We employed Natural Language Processing (NLP) sentiment analysis, with a Multilingual BERT model, alongside qualitative examination of key statements. Findings indicate a generally positive sentiment towards migration among mainstream parties, consistently with Portugal's socio-economic context. However, the 2019 election highlighted a shift to polarized views, especially from the emerging extreme-right party Chega, mirroring wider European trends. This study underscores the interplay between political rhetoric, socio-economic realities, and immigration policy, showcasing the applicability and limitations of NLP in political sentiment analysis.

1 Introduction

Political parties significantly influence public discourse and policies through their manifestos, which may reveal their stance and strategies on key issues, such as migration. Portugal, with its diverse political landscape, offers a rich context for examining migration discourse. The manifestos of its varied political parties, ranging across the political spectrum, provide insights into evolving migration stances, shaped by Portugal's historical, cultural, and socio-economic dynamics. To explore this dynamic interaction further, this study undertakes a systematic analysis of sentiment trends around migration in the manifestos of Portuguese political parties¹ across three legislative election years: 2011, 2015, and 2019. This study analyzes the semantics and sentiments in political manifestos to trace the evolution of migration-related rhetoric across the political spectrum over time. It aims to

¹The complete list of the parties in English, along with their abbreviation, their original name and election rate can be found in the appendix.

deepen understanding of political communication strategies regarding migration.

According to Pordata², Portugal, with a population of 10,56 millions (M) in 2011, 10,38M in 2015, and 10,35M in 2019, has experienced demographic changes in the last years, particularly in terms of migration and population aging. The number of permanent immigrants in Portugal was 18,820 in 2011, 36,849 in 2015, and significantly increased to 95,382 in 2019. At the same time, the aging population of Portugal is evident, with people over 60 years old constituting a substantial portion of the population: 2,62 M in 2011 (24,81%), 2,79 M in 2015 (26,88%), and 3,01 M in 2019 (29,08%). This trend highlights Portugal's demographic challenge of an increasingly older population. Moreover, Portugal's unemployment rates during these electoral years painted a picture against which parties' discourse evolved. In 2011, the unemployment rate stood at 13.4%, reflecting the aftermath of the global financial crisis. By 2015, it had marginally decreased to 12.9%, and by 2019, it significantly dropped to 6.6%, indicating a gradual recovery and stabilization of the economy. The increase in migration, coupled with an aging population and shifting unemployment rates, provides a complex socio-economic context in which political parties operate and articulate their stances on migration. In 2011, under a socialist Prime Minister, Portugal faced a critical financial situation, leading to troika's intervention with austerity measures (Gonzalez, 2014). The public opinion blamed the socialist party for the economical bad situation of the country (Fernandes, 2011). This set a complex context for political narratives on migration (Pereira and Wemans, 2015). By 2015, the country grappled with the aftermath of economic challenges, influencing political stances during a

²Pordata is the Contemporary Portugal Database that provides authoritative and verified statistics on Portugal and Europe, reachable at pordata.pt.

period of recovery (Hutter et al., 2018; Glatzer, 2022). Contrastingly, the 2019 election took place in a more stable economic climate, affecting parties' approaches to issues like migration (Giuliani, 2022).

In light of Portugal's shifting demographics, economic changes, and varying political climates, it could be insightful to analyze how political manifestos' views on migration have transformed. By exploring changes in the sentiments and rhetoric of Portuguese political parties towards migration, we aim to reveal changes in their priorities, strategies, and ideologies. This analysis is key for understanding political responses to migration's challenges and opportunities, and its broader impact on society and policy. Given Portugal's demographic changes and economic and social transitions, this research focuses on how political narratives and sentiment around migration have shifted, providing insights for policymakers, academics, and the public on the political and practical implications of migration policies. To conclude the introduction, let's focus on the research question: how do the sentiments and rhetoric expressed in Portuguese political parties' manifestos towards migration evolve across different legislative election years, and what does this reveal about the changing political landscape in Portugal regarding migration?

2 Related work

This section assesses existing literature on the use of NLP in analyzing political discourse, with a particular emphasis on Portuguese political manifestos and migration issues. Despite extensive research in these domains, there is a notable gap in sentiment analysis of migration topics within Portuguese political manifestos. Prominent in this field are Orellana and Bisgin (2023), who employ NLP for content analysis of political manifestos. This work reveals the capabilities of NLP in discerning political changes and attitudes. Similarly, Cochrane et al. (2022) illustrate the application of computational methods to understand emotional content in political discourses, underscoring the importance of emotion in political texts. Studies such as (Jalali et al., 2012) explore Portuguese political manifestos but do not specifically address migration issues. Migration remains to a certain extent an unexplored area in the analysis of Portuguese political discourse. Regarding migration in political manifestos, Lisi and Borghetto (2018) examine

populist claims in Portuguese politics, offering insights into the framing of migration. Additionally, Gattinara and Morales (2017) delve into the securitization of immigration in Western Europe, linking public opinion and political parties' approaches to immigration. Furthermore, Haselmayer and Jenny (2017) present a unique methodology for analyzing sentiments in political communication. Their procedure involves creating a negative sentiment dictionary tailored to a specific language and domain through crowdcoding. Despite these contributions, there is a clear gap: no study has specifically applied sentiment analysis on migration in Portuguese political parties' manifestos. This lack is significant, considering the role of migration in global and national politics. Our study addresses this gap by applying sentiment analysis to the migration discourse in Portuguese political manifestos.

3 Methodology

To study the changing sentiment about migration in Portuguese political discourse, we employed a combination of NLP and data analysis techniques, both quantitative and qualitative. The study adhered to the following steps.

3.1 Identifying Migration-Related Terms

The idea is to snowballing sampling migration related terms to gather a comprehensive list. Once retrieved the list we will inquire with regular expressions the documents in order to extract only the sentences with migration related terms. We began by compiling a list of migration-related terms using the Word2Vec CBOW-300 word embedding model from *Repositório de Word Embeddings do NILC*. This model struck a balance between efficiency and richness, making it suitable for our computational limitations. The starting list we qualitatively compile comprehends the following migration related words: 'imigrante', 'refugiado', 'asilado', 'fronteira', 'integração', 'trabalhadores estrangeiros', 'políticas migratórias', 'direitos humanos', 'tráfico de pessoas'. For each term, our script finds and lists the most similar words according to the model's embeddings using cosine similarity, which calculates the cosine of the angle between the two word vectors. The smaller the angle, the greater the similarity. In other terms, a cosine similarity closer to 1 means a greater semantic similarity of the terms. This process brought us to find a list of the 10 most similar words for all the unigrams. However, for

the bigrams, this method did not work: the model was not providing any similar word. Therefore, we chose to focus on single words. Once we had a list of comprehensive migration semantic field, we checked it qualitatively to see the correctness and pertinence of the terms. We then decided to increase the list, by adding also the possible derivatives of the words. For example, for the word “migração”, we added: “migrar”, “migratório”, “migrantes”³. This choice was made because the manifestos texts were not normalized, nor the stop words and sentence punctuation were removed. This happened because the BERT model works better with raw natural language texts (Devlin et al., 2019), and also because the extraction of the sentences was impossible if the punctuation for sentence termination were removed, as the model did not know where the sentence stopped.

3.2 Extracting The Portuguese Manifestos and the Relevant Sentences from them

We accessed the manifesto data from the Manifesto Project API, a valuable resource for accessing political manifestos, to acquire manifestos from the Portuguese elections of 2011, 2015, and 2019. The Manifesto Project Database (MPD) is a comprehensive collection of political manifestos and election performance data, curated by the Manifesto Research on Political Representation (MARPOR) project. This resource is hosted on the website of the Social Science Research Center Berlin in Germany. It is renowned for its foundation in quantitative content analysis of election programs from over 50 countries, encompassing all democratic elections since 1945. The Manifesto Project stands out as one of the most widely utilized and influential datasets in the field of political science, earning recognition with the prestigious Lijphart/Przeworski/Verba Data Set Award from the American Political Science Association in 2003 for its outstanding contributions to the discipline. This resource is reachable via the Manifesto Project’s free API, using a specified URL call. This URL contains an API key and a list of keys representing specific texts and annotations to be fetched. We received the data all together, and we separated them per party. We applied this same process for each Portuguese election. We, therefore, made a program that checks the sentences from the extracted manifestos containing at least one of the identified

³The final list of words can be found in the Appendix.

migration-related terms, using regular expressions to extract the specific sentences. This is only a string-matching procedure. Finally, a list of migration related terms was made for each manifesto.

3.3 Sentiment Analysis with Multilingual BERT

Once the specific sentences containing migration terms were extracted, to get the sentiment score expressed in the extracted sentences, we employed a multilingual BERT (Bidirectional Encoder Representations from Transformers) model⁴. This BERT model, has been trained for sentiment classification in texts across 12 languages. It allows the user to analyze the sentiment of a text and get predictions for whether the text is positive, neutral, or negative. It has been distilled from a teacher model using annotated data and can be used to analyze sentiment. This model was chosen for its ability to handle multilingual text, with the idea to further expand, in future studies, this same methodology to Spanish and Italian manifestos, allowing a comparative analysis of sentiment across the nations. We applied the BERT model to each extracted sentence to generate a sentiment score for each sentence, indicating whether it conveyed a positive, neutral, or negative sentiment. This score tells the probability that such a sentence is conveying positive, negative or neutral sentiment and it is expressed in a scale from 0 to 1.

3.4 Data Normalization and Aggregation

To ensure consistency and facilitate analysis, we normalized the sentiment scores by converting the BERT model’s probabilities into a binary scale of +1 (positive), 0 (neutral), and -1 (negative). This standardization enabled us to aggregate the sentiment scores at the party and election year levels, allowing us to visualize and compare sentiment trends across different parties and over time. We then made an average sentiment score per document, by averaging the sentiment score from the extracted sentences in each manifesto. This gave us the sentiment score for migration for each party in each election taken into account.

3.5 Sentiment Trends Visualization

Finally, we utilized data visualization techniques to represent the sentiment trends observed across different parties and election years. This involved

⁴xyuan/distilbert-base-multilingual-cased-sentiments-student

creating three distinct graphs, one for each election year, to describe the evolution of sentiment towards migration over time, by showing the average sentiment score per manifesto. Subsequently, other three graphs representing the number of positive or negative sentences in the dataset, were made. The neutral sentences were excluded from the graph, as they found out to be almost not present: in the most of manifestos the number is zero and in few cases there is only one neutral sentence or by maximum two. By following these methodological steps, we analyzed the sentiment expressed towards migration in Portuguese political manifestos from 2011 to 2019. The results are discussed as follows.

3.6 LLMs and Chatbots integration

Large Language models (LLMs) were employed to assist the researchers in some phases of the research, namely code writing (using GPT-4 and Copilot) and document writing refining (GPT-4). In all the cases, all the results were double checked to make sure of the quality and consistency.

4 Results

This section presents a synthesis of our findings, examining both the sentiment scores and the frequency of migration-related discourse. Overall, the sentiment towards immigration in political manifestos has shown a positive trend, with notable consistency in the central parties' pro-immigration rhetoric. This trend contrasts with the far-right's negative discourse, highlighting the beginning of a polarized political landscape on migration issues. An examination of Social Democratic Party (PSD) and Socialist Party (PS) manifestos reveals an emphasis on integrating immigrants into the labor force, resonating with Portugal's low unemployment rates and demographic need for workers. This pragmatic approach suggests a recognition of the economic benefits of immigration, and this attitude can be seen consistently in the three election years. The only party having negative migration sentiment is Enough (CH), which only appears in 2019 election.

4.1 2011 Election Year

All the parties show a positive sentiment in their manifestos towards migration. The Ecologist Party 'The Greens' (PEV) exhibited the highest positive sentiment (1.00) towards migration, but with a limited mention count (2 sentences), suggesting that they have not focused particularly on migration

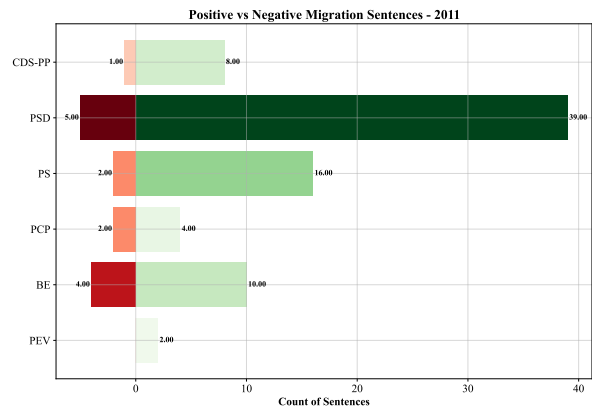


Figure 1: Count of positive and negative sentences per party in 2011 election. The longer the bar, the larger number of sentences. In red the negative sentences; in green the positive ones. The larger the number, the redder or greener the bar color becomes, depending on the polarity. This logic is applied consistently to the other graphs.

topic in their agenda. The PSD and PS showed similarly high positive sentiments (0.77 and 0.74), indicating favorable views towards migration. PSD is the party that focused the most on migration, having a total of 44 sentences related. The Left Bloc (BE) and Portuguese Communist Party (PCP) presented moderate positive sentiments (0.43 and 0.33).

4.2 2015 Election Year

Also in this election all the parties' sentiment toward migration is positive. Portugal Ahead (PàF)'s and PS' manifestos are the ones that mostly mention migration sentences, and also the ones whose average sentiment is the most positive. The PSD/CDS coalition is a recurrent conservative political and electoral partnership in Portugal established by the Social Democratic Party (PPD/PSD) and the People's Party (CDS-PP). In 2015 election they presented at the election with the name of Portugal Ahead (PàF). People-Animals-Nature (PAN) and The Ecologist Party 'The Greens' (PEV) are the parties whose sentiment is less positive. At the same time, they are also the parties that less mention migration. PEV recorded a decreased positive sentiment (0.33), aligning with an increase in migration-related mentions, suggesting a nuanced shift in their migration discourse. Nonetheless, all the parties involved in this election show a positive average sentiment towards migration.

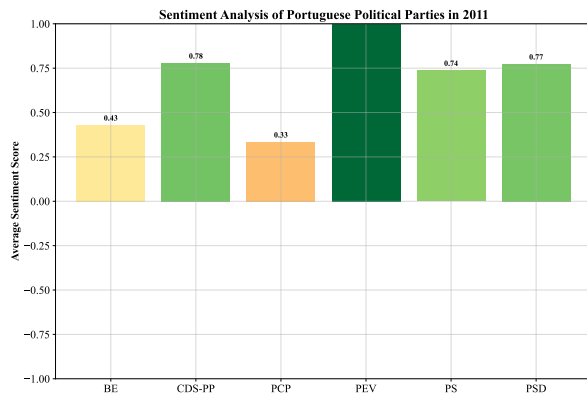


Figure 2: Distribution of average sentiment score per party in 2011 election. The longer the bar, the bigger the average score. The bar gets red if the average sentiment score is negative (< 0); it becomes green if the sentiment is positive (> 0). The stronger the score, the redder or greener the bar color becomes, depending on the polarity. This logic is applied consistently to the other graphs.

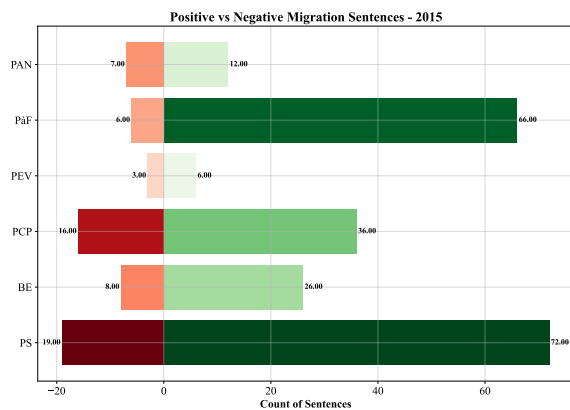


Figure 3: Count of positive and negative sentences per party in 2015 election.

4.3 2019 Election Year

In this election, we can observe a shift of sentiment towards migration. New entrants like Free (L) demonstrated a strong positive sentiment (0.84), indicating a possible progressive stance on migration. Established parties like PSD and PS maintained consistent positive sentiments, signaling a stable and favorable approach over time. A notable change was observed in PEV, showing a stark negative sentiment (-1.00) diverging significantly from previous years. This observation, nonetheless, is based on a single reference. It is, thus, not significant. This is the mentioned sentence:

1. «as situações de conflito não diminuem e aumentam os refugiados e povos deslocados dos

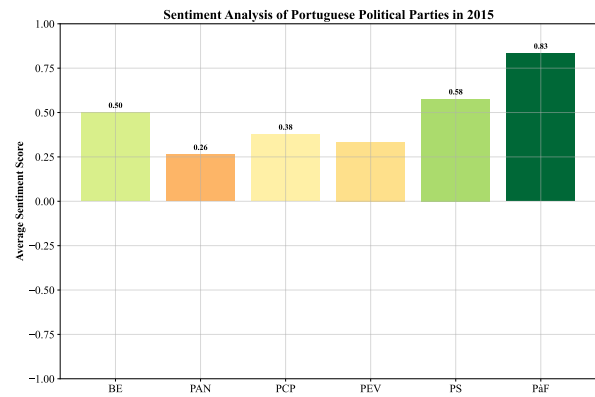


Figure 4: Distribution of average sentiment score per party in 2015 election.

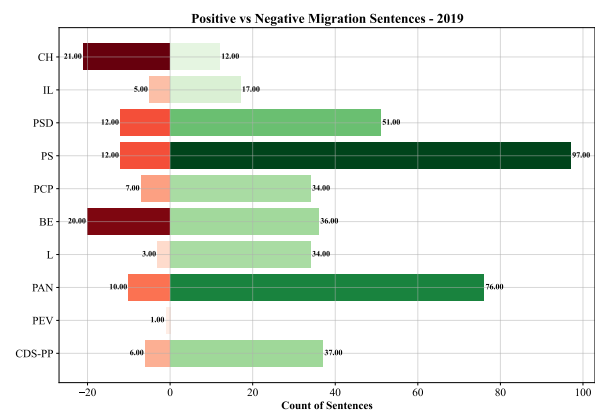


Figure 5: Count of positive and negative sentences per party in 2019 election.

seus territórios»⁵

The sentence contains words and phrases associated with negative sentiments, such as "conflito" (conflict), "não diminuem" (do not decrease), and "refugiados e povos deslocados" (refugees and displaced peoples). These terms typically carry negative connotations, signaling distress, struggle, or undesirable situations. The overall context of the sentence, which talks about increasing conflict and displacement of people, is inherently negative. BERT models are trained to understand the context and not just individual words, thus the overall negative theme of the sentence would contribute to its classification as negative (Xu et al., 2020). Also, there are no words or phrases in the sentence that introduce a positive aspect or counterbalance to the negative themes. Sentiment analysis tools often look for a

⁵The translation is «the situations of conflict do not decrease, and refugees and displaced peoples from their territories increase».

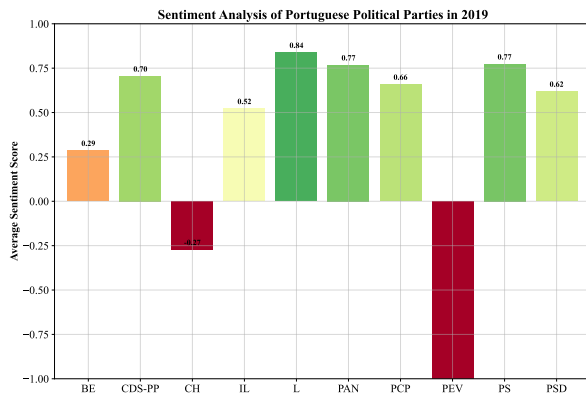


Figure 6: Distribution of average sentiment score per party in 2019 election.

mix of positive and negative cues to determine the overall sentiment, and in this case, the absence of positive cues reinforces the negative classification (Sharif et al., 2016). Nonetheless, as this is the only migration sentence in the whole PEV manifesto, it is not enough data to claim the negative positioning of PEV towards migration. Also in 2019 election the general sentiment trend towards migration is stably positive, with all the parties claiming positive sentences, with the previous exception of PEV and - interesting but not surprising - the far-right party Chega (CH). Founded in 2019 by André Ventura, it is known for its national conservative and right-wing populist stance (Mendes, 2021). The party identifies with nationalist, conservative, and personalist principles. CH advocates strongly against illegal immigration, proposing the deportation of non-working immigrants or those with criminal records. It stands against multiculturalism and the integration of sharia law into the Portuguese legal system (Mendes and Dennison, 2020).

5 Discussion

From a methodological point of view, the findings align with studies, such as (Orellana and Bisgin, 2023) and (Cochrane et al., 2022), which emphasized the utility of sentiment analysis in capturing political stances in manifestos. Nevertheless, the sentiment score in itself is an incomplete information, as we will discuss later. Our findings show a general positive sentiment towards migration across most parties, which interestingly aligns with Portugal's economic context. Given the country's reliance on foreign labor, especially in sectors requiring non-qualified work, the predominantly

positive discourse from central right and left parties appears consistent with its economic necessity, as well as demographic need, considering that the population is aging and decreasing. A more detailed examination of PSD and PS manifestos, as we will observe later on, reveals an emphasis on integrating immigrants into the labor force, resonating with Portugal's low unemployment rates and demographic need for workers. This pragmatic approach suggests a recognition of the economic benefits of immigration. However, especially in 2019 we can observe a dynamic evolution of attitudes towards migration and populist ideological rhetoric. Lisi and Borghetto (2018) found, studying the Portuguese political manifestos from 1995 to 2015 that populist language has primarily been employed by radical left-wing parties as a means to challenge and disrupt a status quo that has been established for over a decade. In 2019 elections the populist discourse arrives also from the extreme-right. These changes are reflective not only of the political and social shifting in Portugal but also of the broader trends in Western politics regarding migration and reflect the complexity of migration as a political issue, subject to various influences, including public sentiment, party ideology, and the socio-economic climate (Kwilinski et al., 2022). The migration policies of European nations have undergone a substantial transformation in the last years. Since the onset of the migration crisis in 2015, there has been a noticeable reinforcement and tightening of measures governing the status of migrants within the territories of these states. During this same period, many European countries experienced a renewed development of radical right parties (Kulaga, 2021). The rise of populist and right-wing parties across Europe and the US, often with anti-immigration stances, together with the 2015 migration crisis, might have influenced the discourse of Portuguese parties, especially those leaning towards the right. The study's findings, as discussed by Bestvater and Monroe (2023), highlight the difference between expressed sentiment and actual political stance. Sentiment scores in political manifestos serve as indicators of rhetorical strategies rather than direct reflections of policy actions (Lutz, 2019). This understanding is crucial, as a party's positive or negative language towards migration doesn't always align with their policy actions. Sentiment analysis, therefore, is valuable for decoding political rhetoric but needs to be paired with qualitative analysis for a complete understand-

ing of a party's position on migration.

Interestingly, recent trends in extreme right-wing politics, exemplified by leaders like Giorgia Meloni, show a closer alignment between public rhetoric and actual beliefs (IIPost, 2023; Toth, 2020). Unlike the usual political divergence between public statements and actions, extreme right movements like Chega (CH) exhibit a more consistent ideological stance, as their public and private expressions tend to mirror each other closely (Biscaia and Salgado, 2022). This suggests that, in such movements, the sentiment expressed might indeed reflect true policy intentions. Consequently, negative sentiments in right-wing manifestos likely indicate genuine anti-immigration views, while positive sentiments in other political spectra need further analysis to ascertain their true intent towards migrant inclusion.

5.1 Qualitative Analysis

It is important to acknowledge that sentiment classification has its limits. For instance, the sentence «together we can bring all migrants back to Africa» also conveys positive sentiment, according to the BERT model used. Nonetheless, it is hard to claim that this sentence is intrinsically supporting migrants' integration policies. It is, thus, needed a qualitative analysis, to further explore more deeply the actual stance over migration. Nevertheless, the sentiment classifier is still informative and useful, when used with a grain of salt: if applied to a whole corpus, it can suggest a more general attitude toward a specific topic. Therefore, we chose to focus qualitatively on the most eloquent sentences extracted from the manifestos. In selecting the following sentences from the political manifestos, our focus is on identifying statements that are qualitatively both relevant and eloquent in capturing the essence of the main parties' stance on immigration.

From the PSD in 2011, a key policy proposal is expressed as:

2. "criar o programa qualificação +, com o objectivo de promover o acesso ao mercado de trabalho de jovens com elevadas qualificações que, nas actuais condições, são fortes candidatos à emigração"⁶.

⁶This translates to: "Create the Qualification + program, with the objective of promoting access to the job market for highly qualified young people who, under current conditions, are strong candidates for emigration."

The sentence's sentiment has 90,17% probability of being positive, 5,84% of probability of being negative and 3,97% of being neutral, according to our model. This initiative reflects the party's focus on preventing the emigration of Portugal's skilled youth, highlighting a strategy to retain local talent by enhancing their job market opportunities.

In 2015, the PCP highlighted the issue of immigrant exploitation and rights:

3. "o combate ao trabalho clandestino, às redes que exploram imigrantes e a legalização do seu trabalho, assegurando a igualdade de tratamento e o respeito pelos direitos laborais e sociais"⁷.

The sentence's sentiment has 55,35% probability of being positive, 39,56% of probability of being negative and 5,08% of being neutral, according to our model. This statement underlines the party's commitment to protecting immigrant workers from exploitation and advocating for their legal and equal integration into the workforce, emphasizing the importance of upholding labor and social rights for all workers.

Moving to 2019 election, we observe a more pronounced focus on immigration issues. The Left Block (BE) in 2019 articulated the challenges faced by immigrants:

4. "para os e as imigrantes que aqui chegam com a sua força de trabalho e a determinação de conquistar uma vida digna, estende-se o tempo exasperante de espera por um atendimento no sef, a permanência interminável em condição irregular e a inerente exposição à violação de todos os direitos fundamentais"⁸.

The sentence's sentiment has 26,69% probability of being positive, 67,71% of probability of being negative and 5,59% of being neutral, according to our model. This statement emphasizes the bureaucratic and legal challenges immigrants

⁷Translated as: "The fight against clandestine work, networks that exploit immigrants, and the legalization of their work, ensuring equal treatment and respect for labor and social rights."

⁸Translated as: "For the immigrants who arrive here with their labor force and determination to conquer a dignified life, they face an exasperating wait time for services at SEF, endless permanence in an irregular condition, and inherent exposure to the violation of all fundamental rights."

face, reflecting a perspective of empathy and concern for their rights and dignity.

Also in 2019, People-Animals-Nature (PAN) commented on the economic aspect of immigration:

5. "no mercado de trabalho, a contribuição dos imigrantes é essencial para o aumento da mão-de-obra, principalmente nos países com populações envelhecidas, como é o caso de Portugal"⁹.

The sentence's sentiment has 83,45% probability of being positive, 11,30% of probability of being negative and 5,24% of being neutral, according to our model. This statement acknowledges the vital economic contribution of immigrants to the labor market, highlighting their role in addressing demographic challenges in countries like Portugal.

It is interesting to contrast these statements with 2019 CH's manifestos. For example:

6. "mas integração não é, nem pode ser, diluição de todas as nações europeias, e de todos os seus cidadãos, numa solução aquosa e indistinta de europeus padronizados e todos iguais".¹⁰

The sentence's sentiment has 15,12% probability of being positive, 55,16% of probability of being negative and 29,71% of being neutral, according to our model. This statement appears to reflect a concern about the loss of distinct national identities and cultural characteristics in the process of integrating diverse populations, particularly in the context of migration in Europe. The use of metaphorical language like "aqueous and indistinct solution" suggests a fear of homogenization or a loss of uniqueness that nations and their citizens might experience in the face of integration. This statement, while focusing on preserving national identity amidst integration, subtly echoes aspects of the "Great Replacement Theory." This theory, often associated with far-right ideologies, claims a deliberate replacement of European populations with immigrants, leading to cultural

⁹This translates to: "In the labor market, the contribution of immigrants is essential for the increase in the workforce, especially in countries with aging populations, such as Portugal."

¹⁰The quote translates to: «but integration is not, nor can it be, the dilution of all European nations, and all their citizens, into an aqueous and indistinct solution of standardized and identical Europeans.»

dilution. The language used in CH's manifesto, particularly about maintaining cultural uniqueness, is reminiscent of this theory's narrative (Ekman, 2022). The concern expressed in the quote about losing national distinctiveness is common among certain political groups, who argue that preserving cultural and national identities is crucial. The fear of dilution of national identity may be rooted in concerns about rapid demographic changes, economic pressures, or social cohesion (Lithman, 2010). In fact, Portugal, similarly to other Western democracies, has been having for the last decades (precisely, from 1984) a fertility rate lower than two kids per woman, meaning a progressive decreasing and aging of the population¹¹. This trend, if summed with the loss of purchasing power after the 2008 crisis, and the progressive increasing of immigrants in the country, can provide a framework for understanding the progressive anti-migration rhetoric. CH's rhetoric often simplifies the complex nature of integration and migration, presenting it as a binary choice between the loss of national identity and the preservation of cultural homogeneity (Byshok, 2020). Another relevant sentence from Chega's manifesto is:

7. "são duas ideologias em confronto: aquela que propõe e que tenta implantar um mundo sem fronteiras habitado por uma massa indiferenciada de indivíduos sem raízes, sem família, sem comunidades próximas e sem nação, o consumidor ideal porque completamente desprovido de defesas".¹²

The sentence's sentiment has 42,00% probability of being positive, 46,31% of probability of being negative and 11,67% of being neutral, according to our model. This sentence speaks to a perceived conflict between globalism and nationalism, a core tension in contemporary debates on migration. The description of a "massa indiferenciada" points to a fear of a loss of individual and national identities, which is a poignant narrative in the political discourse surrounding migration. The phrase "consumidor ideal" (ideal consumer) in Chega's manifesto carries a significant ideological implication. It

¹¹According to *macrotrends.net*

¹²The text translates to English as: "There are two ideologies in conflict: one that proposes and tries to implement a world without borders inhabited by an undifferentiated mass of individuals without roots, without family, without close communities, and without nation, the ideal consumer because completely devoid of defenses".

suggests a critique of globalism, portraying it as a system that seeks to create a homogenized population, devoid of distinct cultural or national identities. This population, described as the "ideal consumer," is seen as easily influenced and lacking in defenses against globalist agendas. The term evokes a somewhat conspiratorial tone, implying a deliberate effort by proponents of globalism to erode individual and communal identities for easier manipulation and control. This rhetoric reflects a common theme in nationalist discourses, where globalization is often depicted as a threat to national sovereignty and cultural uniqueness.

6 Limitations and Future Research Directions

Our study's approach, leveraging NLP and sentiment analysis, could offer insightful perspectives on the sentiment towards migration in Portuguese political manifestos. However, it faces inherent limitations in capturing the full spectrum of nuances within political texts. The application of sentiment analysis, particularly through the BERT model trained primarily on social media content, might not fully align with the linguistic and rhetorical intricacies of political manifestos. Additionally, the uniform treatment of emigration and immigration in our analysis may not adequately reflect the complex and multifaceted nature of these issues. Future research should aim to bridge these gaps by investigating the interrelations between political rhetoric, policy actions, and public opinion on migration. Plus, distinguish between emigration and immigration perspectives could offer deeper insights into political parties' stances and strategies.

7 Conclusions

This study has systematically analyzed the sentiment towards migration expressed in Portuguese political manifestos across three electoral cycles. Our findings reveal a generally positive sentiment towards migration among the mainstream political parties. Notably, the emergence and rise of the CH party introduce a divergent narrative, marked by a distinctly negative sentiment towards migration. This development is particularly salient in the context of the upcoming elections, where the political discourse on migration and its implications for policy and social cohesion will be pivotal.

The ascent of CH is better understood inside the broader European trend of growing right-wing

populism, characterized by a sceptical stance on immigration. This shift poses critical questions for Portugal's political landscape and its future migration policies. The findings of this study suggest that while Portugal has historically embraced a pro-immigration stance, the political rhetoric around migration is becoming increasingly polarized, mirroring wider European trends.

This study contributes to the broader discourse on migration, offering insights into the evolving political landscape in Portugal and its implications for future elections. In conclusion, the rise of the CH party and its implications for Portugal's migration discourse and policies point to a critical juncture. The upcoming 2024 elections will not only shape the country's political landscape but also determine the trajectory of its migration policies in the face of global and European challenges. It is imperative for political parties, policymakers, and civil society to engage in informed and constructive dialogue on migration, ensuring that Portugal remains a society that values diversity and inclusivity, while addressing the legitimate concerns and challenges that migration presents.

Acknowledgements

We extend our deepest thanks to the HYBRIDS project, a Marie Skłodowska-Curie Doctoral Network funded by the European Union (EU) and UK Research and Innovation (UKRI). This project's support has been fundamental to our research. We are especially grateful to the Manifesto Project for providing the data that formed the basis of our analysis. Special acknowledgment goes to Davide Bassi for his expert guidance. We also express our gratitude to the research centers CIDEHUS, CiTIUS, and FBK for their resources and support, which have been instrumental in the success of our work.

References

- Samuel E Bestvater and Burt L Monroe. 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 31(2):235–256.
- Afonso Biscaia and Susana Salgado. 2022. [Placing portuguese right-wing populism into context](#). *Contemporary Politics, Communication, and the Impact on Democracy*.
- S. O. Byshok. 2020. [Migration and recent aspects of right-wing populist discourse in europe](#). *RUDN Journal of Political Science*.
- Christopher Cochrane, Ludovic Rheault, Jean-François Godbout, Tanya Whyte, Michael W-C Wong, and Sophie Borwein. 2022. The automatic analysis of emotion in political speech based on transcripts. *Political Communication*, 39(1):98–121.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *ArXiv*, abs/1810.04805.
- Mattias Ekman. 2022. [The great replacement: Strategic mainstreaming of far-right conspiracy claims](#). *Convergence: The International Journal of Research into New Media Technologies*, 28:1127 – 1143.
- Jorge M. Fernandes. 2011. [The 2011 portuguese election: Looking for a way out](#). *West European Politics*, 34:1296 – 1303.
- Pietro Castelli Gattinara and Laura Morales. 2017. The politicization and securitization of migration in western europe: Public opinion, political parties and the immigration issue. *Handbook on migration and security*, pages 273–295.
- M. Giuliani. 2022. [Voting between two global crises. a nuts3-level analysis of retrospective voting in four south-european countries](#). *Italian Political Science Review/Rivista Italiana di Scienza Politica*.
- M. Glatzer. 2022. [Portugal’s social and labour market policy: The crisis, the troika and beyond](#). *Portuguese Studies*, 34:104 – 118.
- Pilar Gonzalez. 2014. [Gender issues of the recent crisis in portugal](#). *Revue De L’ofce*, 133:241–275.
- Martin Haselmayer and Marcelo Jenny. 2017. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & quantity*, 51:2623–2646.
- Swen Hutter, Hanspeter Kriesi, and Guillem Vidal. 2018. [Old versus new politics](#). *Party Politics*, 24:10 – 22.
- IIPost. 2023. [Scherzo telefonico a giorgia meloni](#). <https://www.ilpost.it/2023/11/01/scherzo-telefonico-meloni/>.
- Carlos Jalali, Patrícia Silva, and Sandra Silva. 2012. Givers and takers: Parties, state resources and civil society in portugal. *Party Politics*, 18(1):61–80.
- Maxim Kulaga. 2021. [Consequences of the radicalization of migration policy in western europe: Socio-economic aspect](#). *DEMIS. Demographic Research*.
- Aleksy Kwilinski, Oleksii Lyulyov, Tetyana Pimonenko, Henryk Dzwigol, Rafis Abazov, and Denys Pudryk. 2022. International migration drivers: Economic, environmental, social, and political effects. *Sustainability*, 14(11):6413.
- Marco Lisi and Enrico Borghetto. 2018. Populism, blame shifting and the crisis: discourse strategies in portuguese political parties. *South European Society and Politics*, 23(4):405–427.
- Y. Lithman. 2010. [The holistic ambition: Social cohesion and the culturalization of citizenship](#). *Ethnicities*, 10:488 – 502.
- Philip Lutz. 2019. [Reassessing the gap-hypothesis: Tough talk and weak action in migration policy?](#) *Party Politics*, 27:174 – 186.
- Mariana S. Mendes. 2021. [‘enough’ of what? an analysis of chega’s populist radical right agenda](#). *South European Society and Politics*, 26:329 – 353.
- Mariana S. Mendes and J. Dennison. 2020. [Explaining the emergence of the radical right in spain and portugal: salience, stigma and supply](#). *West European Politics*, 44:752 – 775.
- Salomon Orellana and Halil Bisgin. 2023. Using natural language processing to analyze political party manifestos from new zealand. *Information*, 14(3):152.
- Paulo T. Pereira and Laura Wemans. 2015. [Portugal and the global financial crisis: short-sighted politics, deteriorating public finances and the bailout imperative](#). *Research Papers in Economics*.
- Wareesa Sharif, N. Samsudin, M. M. Deris, and Rashid Naseem. 2016. [Effect of negation in sentiment analysis](#). *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 718–723.
- Tamas Toth. 2020. Target the enemy: explicit and implicit populism in the rhetoric of the hungarian right. *Journal of Contemporary European Studies*, 28(3):366–386.
- Hu Xu, Lei Shu, Philip S. Yu, and Bing Liu. 2020. [Understanding pre-trained bert for aspect-based sentiment analysis](#). pages 244–250.

Appendix

Date	Party Name	Original Name	Abbrev.	% vote
2011.06	Ecologist Party 'The Greens'	Partido Ecologista 'Os Verdes'	PEV	1,029
2011.06	Left Bloc	Bloco de Esquerda	BE	5,17
2011.06	Portuguese Communist Party	Partido Comunista Português	PCP	7,201
2011.06	Socialist Party	Partido Socialista	PS	28,05
2011.06	Social Democratic Party	Partido Social Democrata	PSD	38,66
2011.06	Social Democratic Center-Popular Party	Centro Democrático e Social – Partido Popular	CDS-PP	11,71
2015.10	Portugal Ahead	Portugal à Frente	PàF	36,86
2015.10	Ecologist Party 'The Greens'	Partido Ecologista 'Os Verdes'	PEV	1,008
2015.10	People-Animals-Nature	Pessoas-Animais-Natureza	PAN	1,39
2015.10	Left Bloc	Bloco de Esquerda	BE	10,19
2015.10	Portuguese Communist Party	Partido Comunista Português	PCP	7,558
2015.10	Socialist Party	Partido Socialista	PS	32,31
2019.10	Ecologist Party 'The Greens'	Partido Ecologista 'Os Verdes'	PEV	1,057
2019.10	People-Animals-Nature	Pessoas-Animais-Natureza	PAN	3,493
2019.10	Free	Livre	L	1,144
2019.10	Left Bloc	Bloco de Esquerda	BE	10,01
2019.10	Portuguese Communist Party	Partido Comunista Português	PCP	5,283
2019.10	Socialist Party	Partido Socialista	PS	38,2
2019.10	Social Democratic Party	Partido Social Democrata	PSD	29,18
2019.10	Liberal Initiative	Iniciativa Liberal	IL	1,355
2019.10	Social Democratic Center-Popular Party	Centro Democrático e Social – Partido Popular	CDS-PP	4,439
2019.10	Enough	Chega	CH	1,358

Figure 7: Table of the parties along with their abbreviation, percentage of votes archived and original names.

List of words			
acolhimento	acolher	acolhida	acolhedor
arabização	arabizar	arabizado	
asilo	asilar	asilado	asilagem
assimilação	assimilar	assimilado	assimilável
colonização	colonizar	colonizador	colonizado
colon	colônia	colonial	colonizar
deportação	deportar	deportado	
diáspora	diaspórico	diaspórica	
emigração	emigrar	emigrante	emigrado
emigrante	emigrar	emigrou	emigração
exilado	exilar	exílio	exilados
expatriado	expatriar	expatriação	expatriamento
extradição	extraditar	extraditado	
fronteira	fronteiriço	fronteiras	fronteiriça
imigração	imigrar	imigrante	imigrado
imigrantes	imigrar	imigração	imigratório
integração	integrar	integrado	integrante
judeu	judaísmo	judaico	judaizar
migrante	migrar	migração	migratório
multiculturalidade	multicultural	multiculturalismo	multiculturais
ocupação	ocupar	ocupado	ocupacional
refugiado	refugiar	refúgio	refugiados

Figure 8: Table of the total words used to extract the migration terms.

In this study, the following python libraries were utilized for data processing, analysis, and visualization: **re**, **pandas**, **matplotlib**, **seaborn**, **os**, **numpy**, **transformers**, **spacy**, **torch**, **json**, **gensim.models**, **collections**.

Analysing entity distribution in an annotated 18th century historical source

Daniel De Los Reyes¹, Renata Vieira², Fernanda Olival², Helena Freire Cameron³, Fátima Farrica²

¹Pontifical Catholic University of Rio Grande do Sul, PUCRS

²CIDEHUS - University of Évora, ³CIDEHUS - Portalegre Polytechnic University, Portugal

daniel.reyes@edu.pucrs.br, renatav@uevora.pt,

mfo@uevora.pt, helenac@ipportalegre.pt, fatimafarrica@sapo.pt

Abstract

This paper presents a distribution analysis of named entities in a historical source, an 18th century Portuguese text collection. The source has been transcribed, revised, normalised and annotated manually with the help of an annotation tool. The distribution analysis was carried out automatically with the help of an extraction parser applied to the annotated texts. The central question of this text is to analyse the meaning of this distribution.

1 Introduction

Named entity recognition (NER) for history research is becoming a trend. In (Ehrmann et al., 2023), we find a survey on named entity recognition and classification in historical documents that considers a variety of other languages. Among others equally pertinent, we may refer to recent studies for NER for historical Portuguese (Grilo et al., 2020; Aguilar et al., 2017; Zilio et al.).

This study introduces a novel exploration of a set of historical Portuguese texts referred to as the "Parish Memories", which were created during the period from 1758 to 1761. This collection holds significant cultural and historical value, comprising the responses to a survey containing 60 questions and distributed in 1758. The answers, originally handwritten by parish priests across the entire kingdom of Portugal, have been meticulously transcribed and normalised for analysis (Olival et al., 2023a).

Our earlier research (Vieira et al., 2021) conducted experiments involving three primary categories (PERSON, LOCAL, ORGANISATION). Later, we performed a *corpus*-based study to define the extension of these categories (Cameron et al., 2022), which subdivide them into more detailed classes as presented in Section 2.

This paper seeks to provide an overview and discussion of the distribution of annotated entities

within these more refined categories in the historical collection under consideration.

2 NE categories customised to historical research

The annotation process of this work endeavours to capture the intricacies conveyed in historical sources from past ages, recognising their distinctions from contemporary expressions. We started by delineating five primary categories: PERSON, PLACE, ORGANISATION, TIME, and AUTHOR WORK. The initial four categories seek to address historical queries related to Who, Where, What, and When, while the last category enables us to analyse the textual sources referenced in the *corpus*.

Owing to their intricacy and significance for the source study, the primary categories PERSON and PLACE were divided into several subcategories.

2.1 Sub-categories of Person

The society of the 18th century was characterised by the inequality of individuals before the law and numerous markers of social differentiation. Often, titles and occupational positions were integral to a person's name and identity. For the annotation to be helpful to historians, it must replicate this reality.

Therefore, the category person (PER) considers references by name (PER_NAM), occupation (PER_OCC), or social category (PER_CAT) - in that order of preference if more than one appears in the expression - and group of persons (PER_GROUP).

Examples of mentions of persons by occupation, social category, and groups of persons are:

- Arcebispo de Évora [Archbishop of Évora]
- Conde da Torre [Count of the Tower].
- Sequeiras [the Sequeira family]

There are also specific subcategories for mentions of saints (PER_SAINTE), divinities

(PER_DIV), and authors (PER_AUT).

2.2 Sub-categories of Place

Concerning places, we generalised the usual category location (LOC) to place (PLC). This category is subdivided into geopolitical entities (PLC_GPE), aquifers (PLC_AQU), mountains (PLC_MOUNT), facilities (PLC_FAC), and one extra subcategory for other locations (PLC_LOC). References to geographical points, such as rivers and mountains, are essential for geo-references.

2.3 Other categories

Regarding time expressions (TIM), we only annotated specific references to dates, for instance, the year 1755 [the year of 1755].

Organisation (ORG) includes all typologies of organisations, like, for example:

- Convento de Santo António [Santo António Monastery]
- Confraria de São Pedro [São Pedro Fraternity]

Written documents mentioned in the memories were attributed to the category AUTWORK.

3 Distribution analysis

The tool used for the manual annotation was the INCEPTION platform¹. We worked on the output files generated with the annotation information. We have different files for each parish of each municipality, this organisation allows the analyse the entities across parishes and municipalities.

The annotated subset gathers 71 parishes of Alentejo, corresponding to 17% of parishes of this region, the largest in Portugal. However, qualitatively, they belong to the most important municipalities of the region: Beja, Évora, Portalegre, and Vila Viçosa. The first three are currently the district capitals. Vila Viçosa, in the past, was the headquarters of the Duke of Bragança, the manor house that served as the birthplace of the Portuguese ruling dynasty in 1758-61.

Municipality	Parishes	Texts	NEs
Beja	29	695	1895
Évora	22	879	1836
Portalegre	14	312	855
Vila Viçosa	6	210	474
Total	71	2096	5060

Table 1: Distribution of NEs by parish

¹<https://inception-project.github.io>

Analysing named entities (NEs) extracted from historical texts of parishes in Portugal is essential to understanding the vast range of information present in the documents. In this detailed analysis, we present explanatory graphs showing the categories of named entities by the municipality, the general distribution of these categories, and the main named entities highlighted by the municipality and globally.

We created a text parser to simplify the extraction, analysis and organisation of named entities extracted from Inception’s output file. This parser can be found in the following repository². After applying the parser to the annotated texts, we explored the results of annotation and categorisation in general and also by the municipality. Table 1 shows that we analysed more than 2000 texts, covering 71 parishes across four municipalities. Also, we can see in Table 1 that we have 5060 annotated NEs as a result of the manual annotation.

3.1 Distribution of named entity by categories

Table 2: Distribution of NE categories

Category	Distribution (%)
PLC	42.66
PER	40.43
ORG	7.96
TIM	6.22
AUTWORK	2.73

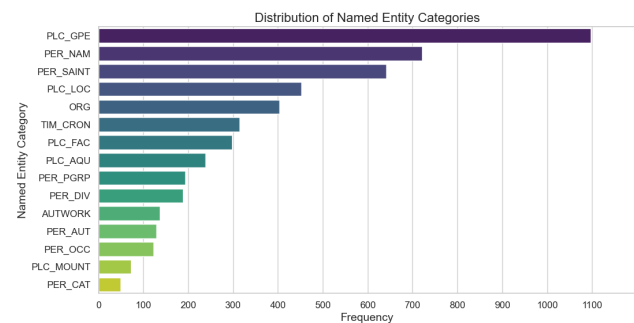


Figure 1: General distribution of NE categories

Understanding the general distribution of named entity categories across historical texts is essential to contextualising the research. Table 2 highlights the predominance of PLACE and PERSON categories compared to the others, totalling more than 80% of the named entities noted in these texts.

²<https://github.com/DanielReeyes/inception-entity-parser>

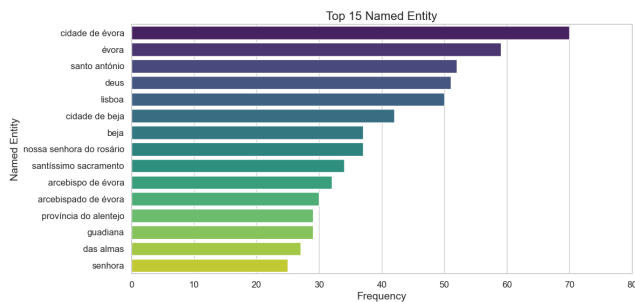


Figure 2: Top 15 NEs among all municipalities

The distribution is unbalanced, where the major categories represented in the corpus are related to geopolitical entities, person names, and saints. Persons are referenced only by category, and mountains are the least represented.

Figure 1 provides an insightful representation of the global distribution of named entities categorised into subcategories. By examining this chart, one can gain a panoramic view of the main categories present in historical texts and identify areas of focus and concentration of information. This visual representation makes comprehending the distribution of NEs and their subcategories easier, allowing for a more comprehensive analysis of the text data.

Based on the distribution of macro categories, the subcategories that appear the most are PLC_GPE (referring to geopolitical entities), PER_NAM (referring to personal names), and PER_SAINTE (referring to saint names), with 1097, 721, and 642 annotations, respectively.

To gather more specific information across all parishes, Figure 2 shows the 15 most referenced named entities across all texts from all analysed municipalities. It summarises the most significant entities found in the historical texts of all parishes, offering a comprehensive perspective on the most recurrent and important elements in analysing the named entities. Specifically, Évora, Lisbon, and Beja were the most frequently mentioned named entities in the GPE category among all entities. They were followed by named entities in the saints and holiness categories, such as Saint Anthony, God, and Our Lady of the Rosary, a widely spread devotion in Portugal after the Counter-Reformation. Lisbon also received significant mentions. In the first group, the most surprising inclusion is Beja, considering that at that time, Beja was not the episcopal capital. The question posed in the inquiry was: "How far is the parish from the episcopal capital city, and how far is it from Lisbon, the capital of

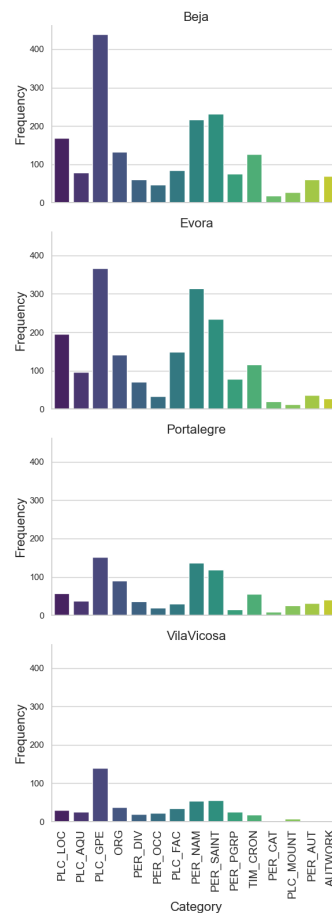


Figure 3: Distribution of NE subcategories by municipality

the Kingdom?" As a result, Lisbon was consistently cited in each parish.

3.2 Distribution of named entities by municipality

To analyse thematic variations between parishes, we analysed the distribution of named entity categories. Figure 3 displays the frequency of categories in each municipality.

The graph analysis reveals that they all follow the same pattern of subcategory distributions as the global context. All have a prevalence of PLC_GPE categories and subcategories, even when analysed scenario by scenario. The other two subcategories highlighted in the general analysis, PER_NAM and PER_SAINTE, are also present when we analyse the data by the municipality.

Identifying the main named entities in each municipality is crucial to gaining a more specific view of local particularities. Figure 4 representing the top 15 NEs by municipality provides a better understanding of the local context. This visualisa-

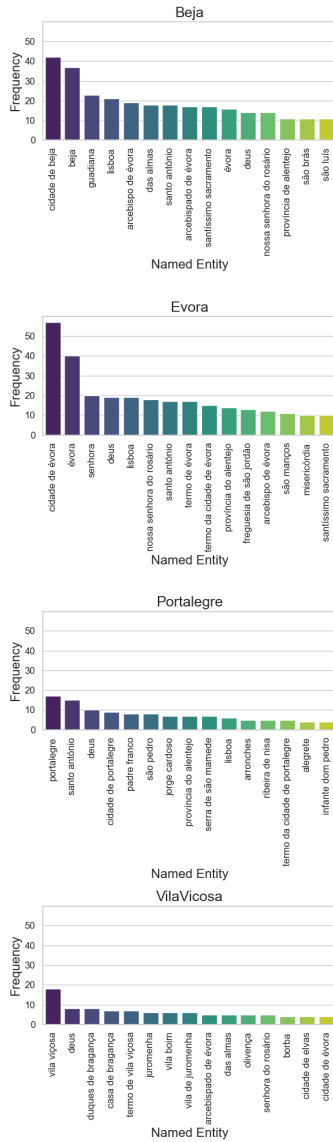


Figure 4: Top 15 NEs by municipality

tion highlights the most prominent entities in each location, contributing to a deeper understanding of local history. For instance, we can see the importance of Évora across municipalities, we see different saints mentioned across regions, Saint Anthony ("Santo António") in Beja and Évora, Saint Peter ("São Pedro") in Portalegre. These two male saints represent two of the three most venerated in Portugal. Saint Anthony, who is believed to have been born in Lisbon, has been widely invoked since the 17th century to locate lost objects, while Saint Peter serves as the guardian of the keys to heaven, the patron saint of the Church, and the papacy (Farmer, 1997). The latter also symbolises the reaffirmation of the triumphant Catholic Church after the Counter-Reformation. The Bragança family is highly mentioned in Vila Viçosa. It represented

a key element of the town's identity due to its direct ties to royalty and its contributions to the local community through the sponsorship and establishment of convents, chapels, and other communal facilities. It was consistently referenced positively, and in just one of the parishes, there was mention of the obligation for the population to pay taxes when constructing mills and other devices in the local aquifers (Olival et al., 2023b).

4 Conclusion

In this work, we presented a study on the collection of *Parish Memories*, which describes aspects of Portugal from the 18th century. In this work, Named Entities were analysed regarding their distribution in the parishes of the Alentejo region. The predominance of GPE can reinforce the idea that this survey was launched in 1758 to resume the project of a Geographical Dictionary of Portugal, initiated before the earthquake of 1755 and interrupted by this catastrophe (Olival et al., 2023a).

Exploring this data helps achieve valuable insights from historical registers about the parishes in Portugal at that time, helping to gain a richer and more contextualised understanding of local history. By studying these Named Entities through manual annotation of historical texts, we can create more robust and reliable datasets and compare between parishes. The annotation enables us to conduct experiments to develop and test methodologies such as Artificial Intelligence models for extracting named entities, making it possible to automate this type of task (Santos et al., 2024).

Acknowledgements

This work has received financial support from the Portuguese Science Foundation FCT in the context of the projects CEECIND/01997/2017 and UIDB/00057/2020 - <https://doi.org/10.54499/UIDB/00057/2020>.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Helena Freire Cameron, Fernanda Olival, Renata Vieira, and Joaquim Francisco Santos Neto. 2022. *Named entity annotation of an 18th century transcribed corpus: problems, challenges*. In *Proceedings of the*

- Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022*, volume 3128 of *CEUR Workshop Proceedings*, pages 18–25. CEUR-WS.org.
- Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. [Named entity recognition and classification in historical documents: A survey](#). *ACM Comput. Surv.*, 56(2).
- David Farmer. 1997. *The Oxford dictionary of saints, 4th ed.* Oxford University Press, Oxford, UK.
- Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. [The BDCamões collection of Portuguese literary documents: a research resource for digital humanities and language technology](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 849–854, Marseille, France. European Language Resources Association.
- Fernanda Olival, Helena Freire Cameron, Fátima Farrica, and Renata Vieira. 2023a. [As memórias paroquiais \(1758\) do atual concelho de vila viçosa](#). *Calipole: revista de Cultura*, (29):85–128.
- Fernanda Olival, Helena Freire Cameron, and Renata Vieira. 2023b. [As memórias paroquiais: Do manuscrito ao digital](#). *Atas da Jornada de Humanidades Digitais do CIDEHUS*.
- Joaquim Santos, Renata Vieira, Fernanda Olival, Helena Cameron, and Fatima Farrica. 2024. [Named entity recognition specialised for portuguese 18th century history research](#). In *Proceedings of International Conference on the Computational Processing of Portuguese (PROPOR 2024)*.
- Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. [Enriching the 1758 portuguese parish memories \(alentejo\) with named entities](#). *Journal of Open Humanities Data*, 7:20.
- Leonardo Zilio, Maria Jose Bocorny Finatto, and Renata Vieira. [Named entity recognition applied to portuguese texts from the 18th century](#). In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022)*, volume 3128.

Roda Viva boundaries: an overview of an audio-transcription corpus

Isaac Souza de Miranda Jr.

Federal University of
São Carlos / Brazil
isc_jr@live.com

Gabriela Wick-Pedro

Center for Artificial Intelligence / Brazil
gabiwick@gmail.com

Cláudia Dias de Barros

Federal Institute of Education,
Science and Technology of São Paulo / Brazil
claudias84@gmail.com

Oto Vale

Federal University of
São Carlos / Brazil
otovale@ufscar.br

Abstract

This paper highlights the initial steps and challenges in creating an audio transcription corpus focused on interviews, intended for linguists, computer scientists, historians, sociologists, political scientists, communication professionals, and digital humanities researchers. The Roda Viva corpus, derived from the renowned Brazilian TV program since 1986, currently offers only textual versions. The overarching goal is to annotate it with morphological and syntactic characteristics, along with audio transcription annotations, emphasizing conversational dynamics.

1 Introduction

This paper outlines the preliminary efforts and challenges in constructing an audio-transcription corpus within the interview genre, poised for exploration by linguists, computer scientists, historians, sociologists, political scientists, communication professionals, and researchers engaged in digital humanities.

While we currently present the initial textual versions of the corpus, our ultimate goal is to annotate it with morphological and syntactic characteristics through the Universal Dependencies methodology (De Marneffe et al., 2021). The Universal Dependencies framework is an initiative aimed at creating consistent morphosyntactic annotation for languages. The methodology encompasses the morphological characteristics annotation (POS and inflections such as gender, aspect, tense, and others) of the sentence elements as well as the relationships between them. Additionally, we aim to incorporate an annotation related to audio-transcription, highlighting significant aspects of the conversational flow and interactions between interviewees and interviewers.

The Roda Viva corpus comprises transcribed and rewritten interviews, transformed into journalistic

texts sourced from the **TV Cultura** show **Roda Viva**. Although our ultimate goal is to establish a multimodal corpus with two subcorpora (one for speech and the other for the transcription), at this moment, only the textual dataset is currently available.

The **Roda Viva** is a renowned interview show on TV Cultura, on air since 1986, and is one of the mainstays of Brazilian television. Airing on Mondays, the program hosts political figures, artists, scientists, and intellectuals, with interviews conducted by journalists or professionals from various fields. In 2007, Fapesp initiated the **Portal Roda Viva**¹ project, resulting in the **Memória Roda Viva**² portal providing complete transcriptions of 713 interviews, with 556,671 sentences, conducted between January 1986 and July 2009.

While the portal is a valuable resource, it has been the subject of academic study in only two works, Botin (2016) and Pacheco (2020). Despite being a rich resource, the **Memória Roda Viva** lacks formalization as a linguistically constructed corpus. Therefore, we intend to present the information available on the portal in a structured linguistic corpus and discuss transcription interventions identified throughout processing.

2 Related works

The previously mentioned works in the introduction, Botin (2016) and Pacheco (2020) are directly associated with the Memória Roda Viva project and primarily conduct theoretical linguistic analyses, rather than focusing on Natural Language Processing (NLP).

With the advancement of Automatic Speech Recognition (ASR), there has been a surge in the creation of multimodal corpora, predominantly cen-

¹<https://bv.fapesp.br/pt/auxilios/23029/projeto-portal-roda-viva/>

²<https://rodaviva.fapesp.br/>

tered on audio-transcription, for NLP. Recently several datasets aimed at ASR have been introduced for Brazilian Portuguese (BP), including the CE-TUC dataset (Alencar and Alcaim, 2008), comprising 145 hours of references and 1,000 sentences spoken by various speakers; Common Voice Corpus 6.1 (Ardila et al., 2020), version pt_63h_2020-12-11, containing a total of 63 hours of audio from 1,120 different speakers; Multilingual LibriSpeech (MLS) dataset (Pratap et al., 2020), providing a total of 3.7 hours of audio for BP; Multilingual TEDx Corpus (Salesky et al., 2021), featuring 164 hours and 93k sentences for BP; and, more recently, CORAA ASR (Candido Junior et al., 2023), encompassing a total of 692.13 hours of audio in BP.

Regarding works predating 2020, noteworthy projects include the NURC (Norma Urbana Culta) (Silva, 1996), which have been active for over 30 years across five different Brazilian capitals (Recife, Salvador, São Paulo, Rio de Janeiro, and Porto Alegre). These projects involve the collection and provision of audio-transcriptions representing diverse manifestations of Brazilian Portuguese.

Additionally, the C-Oral-Brasil-I corpus (Raso and Mello, 2012), associated with the C-Oral-Brasil project (Raso et al., 2015), is a collection of audio transcriptions comprising 139 texts with a cumulative duration of 21.13 hours of audio.

This type of corpus is interesting both for linguistic research and for Digital Humanities studies. One inspiration for the format was the work of (Escoufflaire et al., 2023), who presented 13 years of news from the Belgian French television news site.

3 Corpus Construction

The initial version of the corpus consists of all interviews available on the Memória Roda Viva Portal, in a total of 713 documents. Each document is an interview transcribed and rewritten in a journalistic format. This first set consists of 556,671 sentences, 9,432,547 tokens and 2,606,013 types, as displayed in Table 1.

Version	Sentences	Tokens	Types
V0.1	556,671	9,432,547	2,606,013
V0.2	542,716	8,996,276	2,420,553

Table 1: Corpus Textual Data

Utilizing the first interview dataset, we conducted searches for corresponding editions on the

YouTube channel³ and the show’s website⁴. As a result, we identified a total of 364 interviews with an accompanying video version, as depicted in Table 1:

Interviews	Output
With video	364
Without video	349
Total	713

Table 2: Roda Viva Corpus Interviews with available video

Language	Quantity
PT-BR	308
PT-BR/English	22
PT-BR/Spanish	16
PT-BR/PT-EU	8
PT-BR/French	6
PT-BR/Italian	1

Table 3: Languages distribution of the Interviews with Available Videos

Language	Duration
All languages	522h 06min 46s
PT-BR	446h 18min 49s

Table 4: Languages distribution of the Interviews with Available Videos

Subsequently, we seek to confirm which of these interviews were conducted in Brazilian Portuguese, taking into account Portuguese interviewees who reside in Brazil, such as Maria da Conceição Tavares in the maria_da_conceicao_tavares_1995 interview available on corpus. As illustrated in Table 2, 308 of 364 interviews were conducted in Brazilian Portuguese.

The 308 interviews in Brazilian Portuguese represent an amount of 466 hours of video, as highlighted in Table 3 and 4.

3.1 Current Versions

The textual corpus is available in two versions, as shown in Table 1. The first version, Version 0.1, comprises texts acquired directly from the Roda Viva portal. These texts have undergone a cleaning process, during which elements such as links

³<https://www.youtube.com/rodaviva>

⁴<https://cultura.uol.com.br/programas/rodaviva/>

and icons found on the original pages of each interview were eliminated. The primary goal was to meticulously preserve the textual content of the interviews, ensuring the comprehensive retention of their original form.

The second version of the corpus arose due to the need to remove the interventions made by the transcribers in the original texts. These interventions, which will be elaborated upon in the subsequent section, involved the insertion of textual information not present in the interview videos. These additions, marked as comments within square brackets ([]), were excluded in Version 0.2 of the corpus. While the disparity between version 0.2 and 0.1 is 2.5% in terms of the number of sentences and 4.6% for tokens, these interventions could introduce noise during textual processing.

Consequently, the Version 0.2 represents the transcribed texts without the introduced interventions by the transcribers.

Both versions are accessible in two formats: a compilation of CSV files (with one interview per file) and a JSON file encompassing all interviews. In CSV files there are five columns: DATA (the date of the interview in the format DD/MM/YYYY), ENTREVISTA, (the name of the interviewee), ORDDEM (the order of speech in the interview), LOCUTOR (the name of the speaker) and FALA (the textualisation of the speaker's speech). These resources are available on our GitHub page⁵.

4 Transcription Intervention

In the original files, each interview contains certain textual interventions, always enclosed within square brackets. These interventions result from the retextualization process undertaken during transcriptions. They manifest in various forms, with some being predictable and frequent, such as the completion of words omitted during speech:

- (1) Eu acho que **[ele]** é o melhor do mundo como chargista
*I think **[he]** is the best cartoonist in the world*
- (2) Vocês não acreditam, **[mas esse assunto]** não me preocupa agora
*You don't believe it, **[but this issue]** doesn't concern me now*

Or pertaining to the conversational flow, addressing

⁵<https://github.com/<ANONYMIZED>/Roda-Viva>

the interaction among interview participants:

- (3) Se você..., na eleição..., poderia fazer... **[falando junto com o Markun e concordando com ele]**
*If you..., in the elections..., you could... **[speaking together with Markun and agreeing with him]***
- (4) ele fez... **[imita a pessoa respirando fundo]** Eu disse: matei o velho! **[Risos]**
*he did... **[imitates someone taking a deep breath]** I said: I killed the old man! **[Laughs]***

And also occurring as abbreviations and acronyms:

- (5) PIB **[produto Interno Bruto]**
*PIB **[Gross Domestic Product]***
- (6) PT **[Partido dos Trabalhadores]**
*PT **[Workers Party Brazil]***

Occurrences of other interventions are less frequent and predictable. These instances typically involve explanations about a subject discussed during the speech, often assuming an encyclopedic character. The following examples have one occurrence in the corpus:

- (7) O Paulinho **[Paulinho da Viola, cantor e compositor]** gravou o quê?
*Paulinho **[Paulinho da Viola, singer and composer]** recorded what?*
- (8) Quer dizer, saber como é que isso transformou os países da “cortina de ferro” **[expressão criada em 1946 pelo primeiro-ministro britânico, Sir Winston Churchill, para designar a política de isolamento adotada pela União Soviética e seus estados-satélites após a Segunda Guerra Mundial. Foi uma expressão usada no Ocidente para designar a fronteira imaginária que dividiu a Europa em duas áreas de distintas: os países socialistas e os países capitalistas]**?
*I mean, knowing how this change “Iron Curtain” countries **[an expression created in 1946 by former British Prime Minister Sir Winston Churchill to designate the isolation policy adopted by the Soviet Union and its satellite states after The Second World War. It was an expression used in the West to designate the imaginary bor-***

der that divided Europe into two distinct areas: socialist countries and capitalist countries]?

Although these interventions constitute a minor portion of the corpus, there are 35,663 unique occurrences of them, and they may contain pertinent information about what was said, as exemplified in (1) and (2), or about significant interactions among interviewers and interviewees, as demonstrated in (3) and (4).

As one of the upcoming steps, we aim to establish a taxonomy for these unique interventions and implement an annotation process to classify and, when necessary, differentiate them from the text.

5 Conclusion and future steps

In this initial overview of the corpus, it has become evident that the interventions conducted by the transcribers, despite constituting a smaller portion of the corpus, are highly relevant to the ultimate goal of annotation – covering both morphological and syntactic aspects as well as conversational elements. A dedicated annotation distinguishing elements such as word completion, conversational flow, abbreviations/acronyms, and topic explanations is crucial for ensuring a comprehensive and accurate version of the corpus.

Following the annotation of transcription interventions, the subsequent steps entail the automatic annotation of the corpus using the Universal Dependencies guidelines⁶ – a framework for the uniform annotation of grammar, encompassing parts of speech, morphological features, and syntactic dependencies, across various human languages – through the parser being developed by the POeTiSA⁷ (Portuguese processing – Towards Syntactic Analysis and parsing) project. This annotation will undergo review and validation by linguists.

The corpus is currently undergoing a review of annotations conducted by the POS annotator Porttagger (Silva et al., 2023), which specializes in Brazilian Portuguese, developed by the previously mentioned research group.

Additionally, we aim to conduct annotations focusing on semantic-discursive aspects, specifically emphasizing translation interventions that convey irony and sarcasm (Pedro, 2018). In conjunction with morpho-syntactic annotation, we plan to iden-

tify negative triggers and their respective scope elements.

Therefore, upcoming versions of the Roda Viva corpus will incorporate the described annotation, along with the availability of audio and videos corresponding to each interview.

Acknowledgements

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 #19/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

Isaac Souza de Miranda Jr. was supported by the São Paulo Research Foundation (FAPESP grant #23/01892-4).

References

- V. F. S. Alencar and A. Alcain. 2008. *Lsf and lpc - derived features for large vocabulary distributed continuous speech recognition in brazilian portuguese*. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1237–1241.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. *Common voice: A massively-multilingual speech corpus*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Livia Maria Botin. 2016. *Ciência e tecnologia em debate: uma análise das entrevistas do programa Roda Viva, da TV Cultura*. Phd thesis, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, Brasil.
- Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, et al. 2023. *Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese*. *Language Resources and Evaluation*, 57(3):1139–1171.
- Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal*

⁶<https://universaldependencies.org/>

⁷<https://sites.google.com/icmc.usp.br/poetisa>

- dependencies. *Computational linguistics*, 47(2):255–308.
- Louis Escoufflaire, Jérémie Bogaert, Antonin Descampe, and Cédric Fairon. 2023. The RTBF corpus: a dataset of 750,000 belgian french news articles published between 2008 and 2021.
- Priscilla Hoelz Pacheco. 2020. *A construção "acontece que" no português brasileiro contemporâneo : Uma análise baseada no uso*. Master thesis, Universidade Federal Fluminense, Niterói, Brasil.
- Gabriela Wick Pedro. 2018. *ComentCorpus: identificação e pistas linguísticas para detecção de ironia no português do Brasil*. Master thesis, Universidade Federal de São Carlos, São Carlos, Brasil.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. [MLS: A Large-Scale Multilingual Dataset for Speech Research](#). In *Proc. Interspeech 2020*, pages 2757–2761.
- Tommaso Raso and Heliana Mello. 2012. The c-oral-brasil i: Reference corpus for informal spoken brazilian portuguese. In *Computational Processing of the Portuguese Language*, pages 362–367, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Tommaso Raso, Heliana Mello, Maryualê Mittmann, et al. 2015. O projeto c-oral-brasil. *CHIMERA: Revista de Corpus de Linguas Romances y Estudios Lingüísticos*, 1:31–67.
- Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*.
- Emanuel Huber da Silva, Thiago Alexandre Salgueiro Pardo, and Norton Trevisan Roman. 2023. Etiquetação morfosintática multigênero para o português do brasil segundo o modelo "universal dependencies". *Anais*.
- Luiz Antônio da Silva. 1996. [Projeto nunc: Histórico](#). *Linha D'Água*, (10):83–90.

Demos

GiDi: A Virtual Assistant for Screening Protocols at Home

Andrés Piñeiro-Martín^{1,2}, Carmen García-Mateo¹, Laura Docío-Fernández¹,
María del Carmen López-Pérez², and Ignacio Novo-Veleiro³

¹GTM research group, AtlanTTic Research Center, University of Vigo, Vigo, Spain

²Balidea Consulting & Programming S.L., Santiago de Compostela, Spain

³Home Hospitalization Unit, University Hospital of Santiago de Compostela, Spain

Abstract

Home hospitalisation is emerging as a key pillar in the evolution of medical care, providing effective and safe hospital care for those patients for whom hospitalisation at home is the best option. However, its applicability is limited by its reliance on healthcare professionals physically travelling to patients' homes. This paper presents GiDi, a virtual assistant designed for screening patients with acute heart failure during home hospitalisation, providing healthcare professionals with information about the patient's condition, and allowing them to prioritise and focus on patients who really need attention. GiDi, fluent in Galician and Spanish, overcomes the challenge of bilingual environments. Developed with state-of-the-art open-source technology, it adheres to stringent healthcare data governance. This work and demonstration showcases our experience in integrating GiDi's components and AI modules, in close collaboration with medical professionals, and presents a robust industrial solution tailored to the Galician-Spanish context.

1 Introduction

Home hospitalisation (Hospitalización a Domicilio, HADO (Hermida-Porto et al., 2015)) is a key factor in the evolution of healthcare, particularly within the framework of the Servicio Gallego de Salud (SERGAS). As healthcare adapts to meet the increasing needs of an ageing population with chronic conditions, HADO embodies a patient-centred approach, providing compassionate care in the familiarity of the patient's home. However, there are practical limitations to its scalability - the need for healthcare professionals to physically visit patients at home, while crucial for personalised care, presents logistical challenges. To overcome these challenges and unlock the full potential of home hospitalisation, a key opportunity lies in exploring how emerging technologies, such

as virtual assistants, can enhance and streamline patient care.

GiDi¹, a virtual assistant designed for monitoring acute heart failure (Farmakis et al., 2015) patients at home, addresses the practical challenges of HADO while improving the efficiency of healthcare professionals. By providing real-time patient information and prioritisation capabilities, GiDi enables healthcare teams to focus where needed, ultimately improving the quality and responsiveness of home care. In this context, GiDi's role goes beyond the use of technology to redefine what is possible within the home hospitalisation model.

Developed using an Ethics by Design methodology, GiDi is the result of collaboration with medical professionals from the IDIS (Instituto de Investigación Sanitaria de Santiago de Compostela) Foundation, as well as stakeholders and end-users. It incorporates state-of-the-art language technology to create a functional pipeline in Galician and Spanish, demonstrating the feasibility of building industrial solutions in bilingual and low-resource language environments using open source technology and robust data governance. This paper outlines the design of GiDi, provides detailed descriptions of its modules and explains its demonstration.

2 GiDi Description

2.1 Origin, design and pilot

The GiDi project (June 2022 - October 2024) is the result of a collaboration between Balidea S.L.², the GTM Department of the University of Vigo³ and the IDIS Foundation⁴, and is part of the Eurostars-3 programme, co-funded by the CDTI and supported by the Horizon Europe Research and Innovation Framework Programme of the European Union.

¹GiDi takes its name from the phonetic transcription of the initials of "GrandDaughter".

²<https://balidea.com/>

³<http://gtm.uvigo.es/en/>

⁴<https://www.idisantiago.es/>

The assistant is designed to adapt a daily screening protocol for acute heart failure, based on an alert system designed by experts at the IDIS Foundation. This system, developed according to the Ethics by Design methodology (Piñeiro-Martín et al., 2022), involves GiDi asking patients questions by voice about basic measurements (sleep, weight, urine, oxygen saturation, blood pressure, etc.) that trigger alerts. Using natural language understanding, GiDi then communicates the screening results to healthcare professionals, providing a comprehensive and technologically advanced approach to healthcare screening.

The project is in its final phase of testing and development. In the coming months (June 2024), a pilot phase will start to evaluate the screening and the prototype. The pilot will be conducted with end users, in collaboration with the HADO team and with nursing home companies as stakeholders.

2.2 ASR multilingual system

GiDi is designed to comprehend spoken Galician and Spanish, and to achieve this, advanced multilingual models and multilingual strategies have been employed within the Automatic Speech Recognition (ASR) module. Specifically, we have performed multilingual fine-tuning of the new wav2vec bert 2.0 speech encoder (Barrault et al., 2023) and of the popular pre-trained XLS-R model (Babu et al., 2022) using balanced data from the Librispeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020) and audios from the FalAI dataset (Piñeiro-Martín et al., 2023; Piñeiro-Martín et al., 2024).

2.3 Virtual assistant text-based solution

To develop our text-based conversational solution, we used Rasa (Bocklisch et al., 2017), a collection of open-source Python libraries designed for developing conversational software. Rasa allows the development of Natural Language Understanding (NLU), Dialogue Management (DM) and Natural Language Generation (NLG) modules in both Galician and Spanish. This is achieved through bilingual/language-agnostic pipelines and language models, such as LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2022), incorporating state-of-the-art Natural Language Processing (NLP) techniques, including the use of Transformers architectures (Bunk et al., 2020).

To create the decision tree and implement the screening protocol, we use a pipeline that integrates

static rules and a model that learns from examples of conversations (Vlasov et al., 2019). Finally, due to the nature of the assistant's responses, they are based on templates and are slightly modified depending on the context of the conversation.

2.4 Bilingual text-to-speech

Given the bilingual nature of the assistant and the user's ability to switch between Galician and Spanish during the conversation, GiDi must deliver speech with consistent performance and characteristics in both languages. It is essential that GiDi adapts its voice to the language chosen by the user.

Due to the limited availability of suitable options, it was decided to use the synthetic voice available in Galician and to adapt the messages in Spanish to be compatible with this voice. For Galician, a language with relatively low resources, the only open-source text-to-speech (TTS) system available was COTOVÍA's voice (Banga et al., 2008; Díaz et al.), which unfortunately fell short of contemporary standards in terms of quality. However, a recent breakthrough has been achieved with the release of the first neural synthetic voice model for Galician through the Proxecto NÓS (de Dios-Flores et al., 2022). This newly developed voice model is now seamlessly integrated into GiDi.

2.5 Demonstration Plan

To showcase the capabilities of GiDi, we have developed a comprehensive demonstration plan that highlights its key features. Users can access the virtual assistant through any web browser⁵, enabling them to explore and experience the following functionalities in Galician and Spanish:

- **Screening Protocol Simulation:** Provide the user with the ability to initiate and simulate the screening protocol, report the measures and generate the appropriate alarms. Access to the assistant's decision tree and screening protocol is provided for this simulation. In addition, the patient can follow the protocol by scheduling and continuing, pausing and resuming, proactively skipping and initiating, correcting and requesting measurements.
- **Language Handling:** GiDi supports bilingual conversations in Galician and Spanish. GiDi is able to recognise the language of the

⁵The test assistant will be accessible via the following link: <https://gidibot.balidea.com/>

speaker and switch accordingly. To achieve this, we have developed a multilingual communication strategy where the assistant only assumes the language until it clearly recognises the user’s language or the user indicates which language they want to communicate in. This strategy involves the use of a multilingual ASR model, ASR models for each language once a language has been unambiguously recognised, and a text-based language identification (LID) model based on FastText (Bojanowski et al., 2017).

- **Advanced Intent Classification through Language Understanding:** Improve intent classification by using sentence embeddings from language agnostic models such as LaBSE and by searching for keywords in the user’s message.
- **Contextual Understanding:** Analyzing both user messages using self-attention mechanisms over the sequence of dialogue turns and the bot’s memory (slots) for nuanced contextual comprehension.
- **Enhanced Named Entity Recognition (NER):** Identifying and transcribing spoken numbers and measurements to ensure protocol adherence. For this purpose, we fine-tune the BERT model (Devlin et al., 2019) to perform NER.
- **Contextual Response Replay:** Users can effortlessly revisit the context of the conversation by requesting a repetition of the last response, ensuring clarity and maintaining a smooth interactive experience.

The demonstration will allow access to a test prototype to explore the functionalities presented above, but full access to screening will be limited as this is not the objective of this demonstration, and user identification functionalities and integration with the management platforms developed for this project will be disabled. Figure 1 shows an example of a conversation in the demo chat widget.

As part of the demonstration, a guide will be provided to help identify how to test the main functionalities of the assistant, indicating the measures included in the screening protocol, or examples of how to communicate with the assistant.

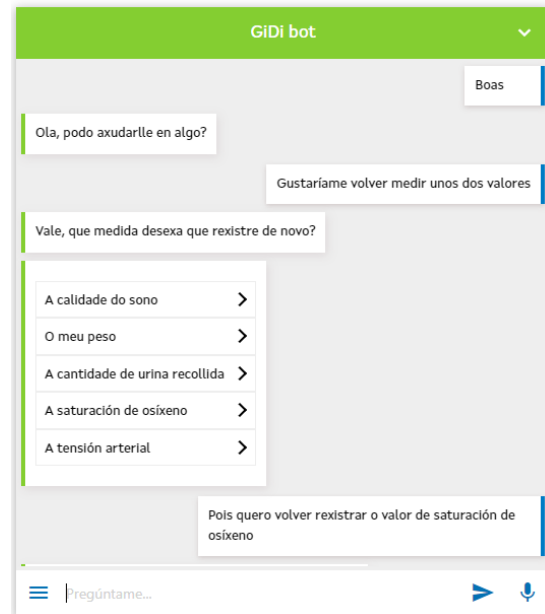


Figure 1: Conversation example in the GiDi chat widget.

3 Conclusions

In summary, GiDi, our bilingual virtual assistant for home hospitalisation, demonstrates the successful integration of open source technologies to build a robust pipeline. Developed using an Ethics by Design approach, GiDi not only skilfully navigates bilingual conversations, but also streamlines health screening protocols for acute heart failure patients. This work highlights the feasibility and effectiveness of building end-to-end solutions using open source tools, and offers a look at the potential transformative impact of virtual assistants in home healthcare.

Acknowledgements

This research has been supported by the Galician Innovation Agency through the program “Doutoramento Industrial” and by the Xunta de Galicia through the grants: “Centro singular de investigación de Galicia accreditation 2019-2022” and GPC ED431B 2021/24.

References

- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. [XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale](#). In *Proc. Interspeech 2022*, pages 2278–2282.
- Eduardo R Banga, Francisco Méndez, Francisco Campillo, Gonzalo Iglesias, and Laura Docío. 2008. Descripción del sintetizador de voz Cotovía para la evaluación Albayzin TTS.
- Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.
- Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.
- Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*.
- Iria de Dios-Flores, Carmen Magarinos, Adina Ioana Vladu, John E Ortega, José Ramon Pichel Campos, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín Diz, Manuel González González, et al. 2022. The nós project: Opening routes for the galician language in the field of language technologies. In *Proceedings of the workshop towards digital language equality within the 13th language resources and evaluation conference*, pages 52–61.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Francisco L Campillo Díaz, Francisco J Méndez Pazó, and Eduardo Rodríguez Banga. Estado actual y líneas futuras del sistema de conversión texto-voz gallego-castellano Cotovía.
- Dimitrios Farmakis, John Parisis, John Lekakis, and Gerasimos Filippatos. 2015. Acute heart failure: epidemiology, risk factors, and prevention. *Revista Española de Cardiología (English Edition)*, 68(3):245–248.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavzhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Leticia Hermida-Porto, LM Dopico-Santamariña, Fernando Lamelo-Alfonsín, B Aldamiz-Echevarría Iraurgui, MA Silva-César, and Luciano Vidán-Martínez. 2015. Hospitalización a domicilio en hospitales públicos gallegos. *Galicia Clínica*, 76(1):7–12.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.
- Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, María del Carmen López-Pérez, and José Gandarela-Rodríguez. 2024. FaIAI: A Dataset for End-to-end Spoken Language Understanding in a Low-Resource Scenario. In *Proceedings of the LREC-COLING 2024 Conference*. LREC-COLING.
- Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, and María del Carmen López-Pérez. 2022. [Ethics Guidelines for the Development of Virtual Assistants for e-Health](#). In *Proc. IberSPEECH 2022*, pages 121–125.
- Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, and María del Carmen López-Pérez. 2023. FSTP Project Report - BEST Assistants - Building E2E Spoken-Language Understanding Systems for Virtual Assistants in Low-Resources Scenarios.
- Vladimir Vlasov, Johannes EM Mosig, and Alan Nichol. 2019. Dialogue transformers. *arXiv preprint arXiv:1910.00486*.

FazGame: A Game Based Platform that Uses Artificial Intelligence to Help Students to Improve Brazilian Portuguese Writing Skills

Jéssica S. Santos and Gabriel Coelho and Sidney Melo and
Oniram Atila and Carla Zeltzer

FazGame, Rio de Janeiro, Brazil

{jessica.soares,gabriel.coelho,sidney.melo}@fazgame.com.br

Abstract

This article describes the FazGame platform, which is a tool developed to assist students' learning process of Brazilian Portuguese as a native language through the creation of narrative and interactive games. The textual content of students' games, such as dialogues between characters and game messages, is evaluated by an Artificial Intelligence based module that uses Natural Language Processing techniques, language models, and classifiers to identify problems such as misspelled words, lack of punctuation and word capitalization, verbal and nominal agreement errors, and cohesion and coherence problems. Based on this analysis, pedagogical interventions are displayed to assist students in the process of improving writing skills. In addition, reports are provided to help teachers monitor students' progress and also to highlight the main difficulties of a given student or class.

1 Introduction

Students in elementary school may face several challenges in the learning process of Brazilian Portuguese (BP) as their native language. These difficulties can stem from a combination of factors, including the complexity of the language, individual learning styles, and the teaching methods employed. Among some common challenges, we can cite:

1. **Spelling and grammar:** Spelling and grammatical errors are common during the language learning process, particularly when students need to learn specific rules. For example, the use of verb conjugations, punctuation, capitalization and accentuation rules, and other grammatical aspects such as agreement can pose a challenge. Also, students may know the words but not know how to write them correctly according to the grammar. It can be common, for example, when it comes to

terms that have a phonetic pattern with the same pronunciation but are written differently. One example of this occurs with the following terms in BP: *voce*, *voçe*, *vossê*, *você*. Even though they sound the same, the last case is the only one that corresponds to the correct word, which denotes the pronoun *you* in English.

2. **Cohesion and Coherence:** Expressing thoughts clearly and coherently in writing can be challenging, especially in long texts. Students may face difficulties in organizing ideas, constructing sentences and paragraphs, and developing a cohesive and coherent narrative.
3. **Reading Comprehension:** Limited vocabulary can be a barrier to comprehension. Students may struggle to understand the meanings of new words, affecting their focus and their ability to interpret texts.
4. **Lack of Motivation:** Motivation and interest play a crucial role in learning. Disinterested classes may not fully engage students in BP classes, impacting their understanding and performance.

In this context, the FazGame platform¹ is an educational game-based platform that is being developed to tackle the aforementioned issues. It uses Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) to automatically detect writing problems in the textual content of narrative games created by students. After identifying those problems, the AI module exhibits a set of pedagogical interventions to help the students correct the textual errors or inconsistencies. The results of the AI-based analysis are also stored in a database system in order to persist information and provide reports about the student's learning process, such as the most common category of errors in different periods of time. Such reports aim to help the

¹<https://www.fazgame.com.br>

080 teachers have a vision of the class’s problems with
 081 regard to learning textual production, and allow
 082 them to deal with the main difficulties in an inter-
 083 active and personalized way, optimizing students’
 084 time and learning. Throughout the entire develop-
 085 ment process, we count on the collaboration of a
 086 pedagogical team that plays a crucial role in vali-
 087 dating the techniques used and the data provided
 088 through interactions with students and teachers on
 089 the platform.

090 **2 FazGame platform**

091 On the FazGame platform, students learn by creat-
 092 ing narrative games. In addition to create textual
 093 elements (like dialogues between characters and
 094 game messages), users must select logical connec-
 095 tors (and/or/end of game/scene change) to set game
 096 flow according to user actions (like clicking on
 097 characters, objects, items, and so on). In this way,
 098 it is possible for the same game to result in different
 099 stories that vary according to the user’s actions, as
 100 each action can lead to a different path. For exam-
 101 ple, the student may create a question dialog that
 102 is associated with two possible answers, and each
 103 answer redirects the user to a different scene. It
 104 is possible to assign points as a reward when the
 105 correct answer is selected. Also, games can be asso-
 106 ciated with different pedagogical tracks that define,
 107 for example, the theme that must be explored by
 108 the student in the narrative game. Figure 1 illus-
 109 trates an overview of the entire scope of the project.

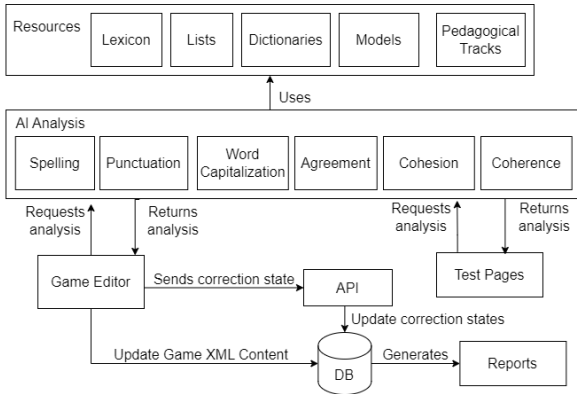


Figure 1: Representation of the entire scope of the project environment.

110 As can be seen in the *AI Analysis* group depicted
 111 in Figure 1, our methods detect problems from six
 112 primary categories (Spelling, Punctuation, Word
 113 Capitalization, Agreement, Cohesion, Coherence).
 114

115 These categories further branch into specific subcat-
 116 egories, each undergoing tailored treatments that
 117 account for linguistic and technological aspects, as
 118 described below. Initially, the problems are cate-
 119 gorized by considering general mistakes and chal-
 120 lenges associated with the structure of words and
 121 sentences when writing, particularly in the context
 122 of elementary school proficiency, as described in
 123 Section 1. For future work, we intend to refine
 124 our methods for different scenarios, depending on
 125 specific grade levels of elementary school.

126 **Textual Analysis – Spelling and Grammar:**

127 The textual analysis related to spelling and gram-
 128 mar combines a set of strategies: Levenshtein (Yu-
 129 jian and Bo, 2007) algorithm, lexicon and auxiliary
 130 lists, phonetic patterns, Large Language Models
 131 (LLMs) for analyzing suggestions probability, Part-
 132 of-speech tagging (POS-tag) and Named Entity
 133 Recognition (NER) (Marrero et al., 2013) tech-
 134 niques, and a set of predefined rules. The Leven-
 135 shtein algorithm and a lexicon resource (based on
 136 BRispell²) are used to find potential suggestions
 137 for misspelled words. In this step, a set of phonetic
 138 patterns is also considered. In cases where there
 139 are many suggestions for a given misspelled word
 140 (many lexicon words at the same Levenshtein dis-
 141 tance), we use the BERTimbau (Souza et al., 2020)
 142 model to select the most likely word to be sug-
 143 gested in the given sentence context. We also save
 144 in the database the occurrence of offensive terms³
 145 and informal language, such as slangs, presented in
 146 the text. A list containing popular foreign words is
 147 used to prevent them from being recognized as lan-
 148 guage errors. We also use a list of names combined
 149 with POS-tag and NER models to identify charac-
 150 ters’ names and avoid recognizing them as errors,
 151 although they are not listed in the lexicon. Also,
 152 this process is important to detect capitalization
 153 and punctuation errors. Punctuation error detection
 154 also uses predefined rules. One example of that is
 155 the comma vocative error, when there is no comma
 156 in a dialogue that is started by a greeting and is
 157 followed by a proper noun (e.g.: “*Oi Maria*” which
 158 corresponds to “*Hi Maria*” in English). SpaCy li-
 159 brary⁴ and LanguageTool rules (Mozgovoy, 2011)
 160 are adopted to support part of this process. Verbal
 161 and nominal agreement errors are captured by com-
 162 bining predefined rules with a dependency parser

²<https://www.ime.usp.br/~ueda/br.ispell/>
³Additionally, these terms are replaced by special charac-
 ters so that they are hidden from the game.
⁴<https://spacy.io/>

and morphological features analysis.

Textual Analysis – Cohesion and Coherence:

Conjunctions are very important elements of cohesion in textual construction as they help to chain ideas together. To look for cohesion problems, we use the BERTimbau model to identify potential incorrect uses of conjunctions according to context. Next, the total number of correct uses of conjunctions is counted and divided by the total number of words, as this metric can be helpful to evaluate textual cohesion (Leal et al., 2023). To evaluate game coherence, we intend to measure two aspects: (i) game graph analysis: development of an algorithm that runs through the game graph and searches for paths that do not lead to other scenes and, consequently, to the end of the game; and (ii) alignment of game theme with the track that it is associated when this is the case. We are investigating strategies like Topic Modeling (Kherwa and Bansal, 2019), Topic Extraction (Campos et al., 2020), Embeddings, and LLMs to check whether the game textual content fits to the track theme.

All strategies and technologies used for textual analysis are incorporated into the platform following a review and development procedure that consists of: I) checking whether the strategy provides solutions for BP, II) checking whether the primary purpose of the strategy aligns with the requirements of our textual analysis context, III) aligning which part of the strategy scope in our textual analysis pipeline, IV) validating the potential results and output possibilities from the analysis in collaboration with the pedagogical team, V) refining the output considering linguistic gaps from the strategy identified in the validation stage, VI) implementing pre- and post-processing to align the input and output required for our textual analysis scope, VII) testing the final method with existing games and new sentences representing diverse usage scenarios. Additional linguistic rules can be mapped by the pedagogical team in the fourth step, in addition to ensuring that the method accurately analyzes and provides correct data regarding the problem categories listed previously.

Pedagogical Interventions: Textual interventions can be viewed as tips to guide students in the correction of their errors, such as the example illustrated in Figure 2. Instead of automatically correcting the text, a set of sequential tips is presented to the students as an opportunity to think about their mistakes and correct them. This set of interventions is created in advance based on each error

subcategory detected in the *AI Analysis* and represents part of the output shown in the *Game Editor* (depicted in Figure 1 by the arrow named “Returns analysis” going from the *AI Analysis* to the *Game Editor*). In each case, the pedagogical team formulates a systematic series of recommendations that aims to reach the solution related to the detected error step-by-step. The suggestions encompass examples of both correct and incorrect sentences or guides on locating the spelling errors and presenting alternatives for their resolution. In general, interventions are organized into three levels that differ according to their specificity and have dynamic fields for displaying the word/sentence with the error detected and for the student to rewrite it based on the recommendations. For example, the third and last intervention can be the presentation of the most likely suggestion. Students can agree with the suggestion or ignore it.

As previously stated, the suggestions strictly concern the structural aspects of writing words and sentences that relate to the challenges of elementary school proficiency mentioned in Section 1. Refining these suggestions is within the scope of future work, considering different interactions with the students on the platform according to elementary school grades.

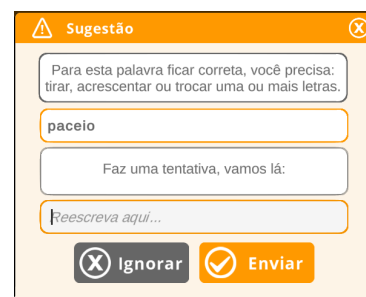


Figure 2: Spelling error - First pedagogical intervention.

Reports: Reports are generated based on the errors stored in the SQL database. An overall summary of the learning process of students and classes is exhibited, taking into account three topics: (i) errors distribution by category in a given period; (ii) most popular errors by category in a given period; (iii) total of correct words divided by the total number of words in two different periods of time to capture students/classes progress. An example of an available report is exhibited in Figure 3.

Distribuição de Erros por Categoria

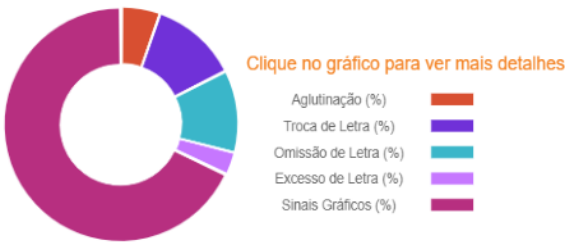


Figure 3: Error distribution by category - Pie chart

The FazGame platform integrates a Rails application with the Unity game editor, which in turn connects to NLP and AI-based modules that were implemented in Python and are made available through AWS Lambda functions. All communication is done through APIs.

3 Tool Demonstration

The FazGame platform can be accessed online through access logins. We intend to demonstrate its functionalities by giving access to the audience so that they can create narrative games and see the pedagogical interventions based on NLP and AI that are displayed to users after textual analysis. Furthermore, the audience will be able to access reports generated based on the detection of errors in BP language. We also intend to make test pages available so that users can test features related to NLP and AI modules that are not yet integrated into the game editor.

Acknowledgements

This work was partially funded by FAPESP and CNPq.

References

- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.
- Pooja Kherwa and Poonam Bansal. 2019. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24).
- Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, pages 1–38.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Maxim Mozgovoy. 2011. Dependency-based rules for grammar checking with languagetool. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 209–212. IEEE.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Indexing Portuguese NLP Resources with PT-Pump-Up

Rúben Almeida
INESC TEC
ruben.f.almeida@inesctec.pt

Ricardo Campos
INESC TEC, Uni. Beira Interior
ricardo.campos@ubi.pt

Alípio Jorge
INESC TEC, Uni. of Porto
amjorge@fc.up.pt

Sérgio Nunes
INESC TEC, Uni. of Porto
ssn@fe.up.pt

Abstract

The recent advances in natural language processing (NLP) are linked to training processes that require vast amounts of corpora. Access to this data is commonly not a trivial process due to resource dispersion and the need to maintain these infrastructures online and up-to-date. New developments in NLP are often compromised due to the scarcity of data or lack of a shared repository that works as an entry point to the community. This is especially true in low and mid-resource languages, such as Portuguese, which lack data and proper resource management infrastructures. In this work, we propose PT-Pump-Up, a set of tools that aim to reduce resource dispersion and improve the accessibility to Portuguese NLP resources. Our proposal is divided into four software components: a) a [web platform](#) to list the available resources; b) a [client-side Python package](#) to simplify the loading of Portuguese NLP resources; c) an [administrative Python package](#) to manage the platform and d) a [public GitHub repository](#) to foster future collaboration and contributions. All four components are accessible using: https://linktr.ee/pt_pump_up

1 Introduction

The topic of NLP resource management in European languages was initially introduced by [Danzin \(1992\)](#), with the first references to Portuguese resources presented ten years later in the works of [Santos \(2002\)](#). The recent advances in NLP, linked to the development of large language models, reintroduced the debate about NLP resource management due to the large volume of training data required by these architectures. Several platforms have been recently introduced, offering different approaches to addressing this problem. Our analysis identified more than 13 platforms that include, to some extent, Portuguese NLP resources (Table 1). These platforms have different geographic origins and operate independently of each other, contribut-

ing to resource dispersion. In a mid-resource language such as Portuguese ([Joshi et al., 2020](#)), this resource dispersion phenomenon exacerbates the already existing challenges linked to the reduced amount of NLP resources, negatively impacting the accessibility to these resources.

To address these challenges, we extend the surveying works of [Almeida \(2023\)](#) and propose PT-Pump-Up, a set of tools that support the development of the first centralising platform for Portuguese NLP resources. In this demonstration, we present the minimum set of valuable features to achieve this goal divided across the four software components that compose PT-Pump-Up: a) A web platform¹; b) A client Python package²; c) An administrative Python package³ and d) A public GitHub repository⁴. Additional details about this release are available in the wiki of the project⁵.

2 PT-Pump-Up

The PT-Pump-Up architecture is presented in Figure 1, which highlights not only the features already implemented but also the work in progress and future plans associated with this project. In this demonstration, we present four scenarios where PT-Pump-Up can be employed to mitigate resource dispersion and enhance synchronization across diverse platforms that support Portuguese NLP resources.

2.1 Indexing Portuguese NLP Resources

The PT-Pump-Up [administrative package](#) permits authenticated [CRUD operations](#) to manage the resources indexed in the platform. These actions can also be done using the [web interface](#), ensuring that the absence of programming skills is not a barrier to interacting with the platform. In Listing 1, we

¹<http://pt-pump-up.inesctec.pt/>

²<https://pypi.org/project/pt-pump-up/>

³<https://pypi.org/project/pt-pump-up-admin/>

⁴<https://github.com/LIAAD/PT-Pump-Up>

⁵<https://github.com/LIAAD/PT-Pump-Up/wiki>

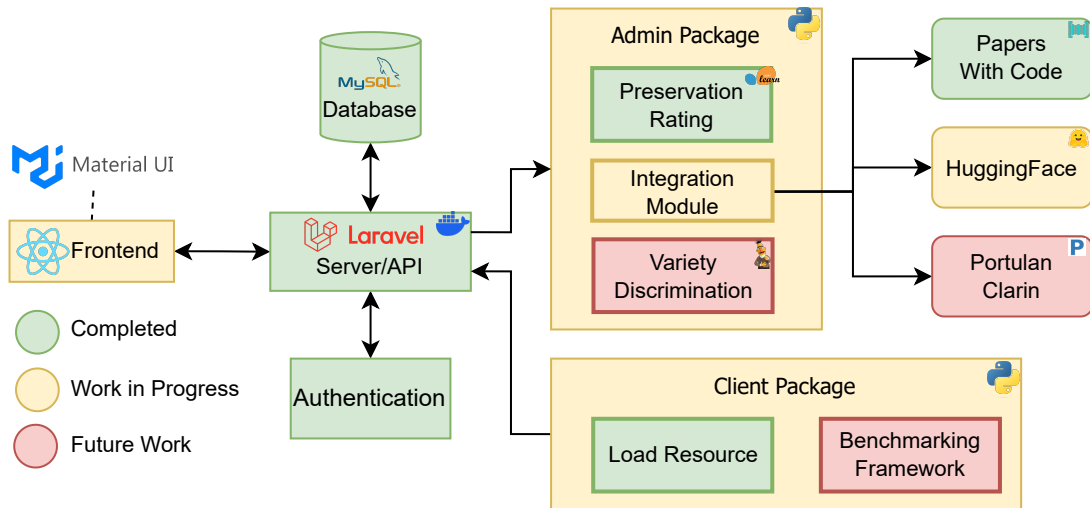


Figure 1: Architecture of PT-Pump-Up. Background colours highlight the completeness of each module.

demonstrate how a new NLP task can be included in PT-Pump-Up with a few lines of Python code.

```

from pt_pump_up_admin.PT_Pump_Up import
→ PT_Pump_Up
from pt_pump_up_admin.crud.NLPTask import
→ NLPTask

pt_pump_up = PT_Pump_Up(
    # Get token: pt-pump-up.inescotec.pt/dashboard
    bearer_token=str,
)
NLPTask(pt_pump_up=pt_pump_up).insert(
    name=str,
    acronym=str,
    papers_with_code_ids=list,
)

```

Listing 1: Inserting a NLP task to the database.

The low-code, open-source collaborative (Co-lab.), off-the-shelf approach proposed in this paper permitted the indexation of 28 datasets and 3 models from the 13 platforms listed in Table 1. Some of these platforms have multiple geographic origins (Origin), but they are mainly Portuguese and Brazilian. Additionally, some of them are no longer updated. The per-NLP Task counts of the datasets indexed are provided in Figure 2.

2.2 Fine-Grained Analysis of Resources’ Linguistic Variety

We believe the biggest limitation of current platforms relies on the lack of Portuguese varieties’ discrimination features (# PT Var.). Many of these platforms either index resources in different Portuguese varieties without detailing how many varieties are considered and how these distinctions were made (signalled in Table 1 with **A**) or, de-

spite permitting other varieties, focus mainly on European and Brazilian Portuguese (2+).

To surpass this limitation and promote the development of mono-variety Portuguese NLP resources, PT-Pump-Up uses a Portuguese variety identification model to scan each resource for its Portuguese variety upon submission. The outputs of this model are then used to provide detailed information about the Portuguese variety of that resource.

2.3 Easy Access to Portuguese NLP Resources

The PT-Pump-Up Python client permits the easy loading of Portuguese NLP resources. The resource is loaded directly if it has a copy in HuggingFace⁶; if not, it returns the metadata that describes it. In Listing 2, we demo how to use PT-Pump-Up to achieve this goal using a few lines of code.

```

from pt_pump_up.PT_Pump_Up import PTPumpUp
client = PTPumpUp()
all_ner_datasets =
→ client.all_datasets(nlp_task="Named Entity
→ Recognition")
print(all_ner_datasets.head())
# Dataset is Loaded as a HF Dataset object
dataset = client.load_dataset(english_name=str)

```

Listing 2: Load Portuguese named entity recognition dataset.

2.4 Measure Resource Preservation Needs

We propose a *resource preservation rating* to identify less accessible resources. Unlike existing platforms (Table 1) that tend to either focus exclusively on storing metadata about the resources (Meta.), or

⁶<https://huggingface.co>

Platform	Updated	Origin	# PT Var.	Colab.	Meta.	Res.
NILC: Tools and Resources	✓	BR	1	×	×	✓
Portulan Clarin (Branco et al., 2023)	✓	PT	⚠	⚠	⚠	✓
Portuguese-NLP	✓	BR	⚠	✓	✓	×
HuggingFace	✓	FR	⚠	✓	⚠	✓
PapersWithCode	✓	USA	⚠	✓	✓	×
ELRA	✓	BE	1	×	×	✓
Open Language Archives Community (Simons et al., 2003)	✓	USA	2+	×	✓	×
European Language Grid (Rehm et al., 2020)	✓	DE	⚠	⚠	✓	×
Linguateca (Santos et al., 2004)	2012	PT	⚠	×	×	✓
Organização Etica.AI	2018	BR	⚠	×	✓	×
ACL Wiki: Resources for Portuguese	2020	USA	⚠	×	✓	×
AiLab	2021	BR	⚠	✓	✓	×
PT-Pump-Up	✓	PT	✓	✓	✓	⚠

Table 1: Platforms supporting Portuguese NLP resources indexing. In dark-gray we highlight those that are no longer active.

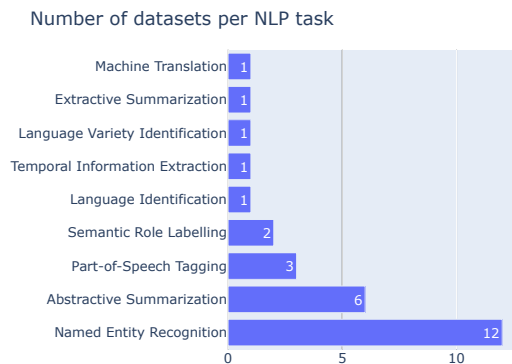


Figure 2: Per-NLP task dataset counts.

store entire copies of existing resources (Res.), in PT-Pump-Up we propose a hybrid approach that uses the *resource preservation rating* to avoid resource duplication.

For resources that exhibit high preservation ratings, only metadata is stored, whereas those with lower ratings are given priority for human intervention and the creation of a backup copy. The preservation rating can be provided during resource submission or automatically determined using a decision tree integrated into the [admin package](#).

2.5 Integrate With Papers With Code

The PT-Pump-Up integration module included in the admin package compresses the logic developed to enforce resource synchronization with other plat-

```

from pt_pump_up.rating.Preservation_Rating
↳ import PreservationRatingDataset

# Instantiates a client
client = PreservationRatingDataset()

preservation_rating = client.predict(...dataset
↳ properties)
print(preservation_rating)

```

Listing 3: Predicting preservation rating of a dataset based on its metadata

forms. In this release, we deliver the tools to support the integration with [Papers With Code](#). This module presents challenges due to the heterogeneity of systems used by the targeted applications. In Listing 4, we demonstrate how PT-Pump-Up can be used to synchronise a resource with Papers With Code using a few lines of code.

```

from pt_pump_up.papers_with_code.PapersWithCode
↳ import PapersWithCodeDataset, PapersWithCode

# Login in PapersWithCode
client = PapersWithCode(username=str,
↳ password=str)

#Create Dataset instance
dataset = PapersWithCodeDataset(...dataset
↳ properties)
#Publish Resource
client.insert(dataset)

```

Listing 4: Insert dataset metadata to Papers With Code.

3 Conclusion and Future Work

This paper details the first release of PT-Pump-Up and how its tools can be used to address the challenge of Portuguese NLP resource dispersion. In this release, we deliver the minimum set of valuable features capable of demonstrating the four software modules that compose PT-Pump-Up. This project is a work in progress, with many future work topics identified. In particular, we highlight the need to extend the integration module to other platforms and develop initiatives to promote PT-Pump-Up and motivate new elements to join the team with the ultimate goal of improving the development of Portuguese NLP solutions.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 | <https://doi.org/10.54499/LA/P/0063/2020>. The authors Ricardo Campos, Alípio Jorge and Sérgio Nunes would like to acknowledge the National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC)..

References

- Rúben Almeida. 2023. Building portuguese language resources for natural language processing tasks. MSc Thesis, Faculty of Engineering, University of Porto.
- António Branco et al. 2023. The clarin infrastructure as an interoperable language technology platform for ssh and beyond. *Language Resources and Evaluation*, pages 1–32.
- A Danzin. 1992. [Towards a european language infrastructure \(dg xiii\)](#).
- Pratik Joshi et al. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). *CoRR*, abs/2004.09095.
- Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, et al. 2020. European language grid: An overview. *arXiv preprint arXiv:2003.13551*.
- Diana Santos. 2002. Um centro de recursos para o processamento computacional do português. *DataGramaZero-Revista de Ciência da Informação*, 3(1).

Diana Santos et al. 2004. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa.

Gary Simons et al. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.

plain X – AI Supported Multilingual Video Workflow Platform

Carlos Amaral^P, Catarina Lagrifa^P, Mirko Lorenz^D,
Peggy van der Kreeft^D, Tiago Veiga^P

^P Priberam, Portugal, ^DDeutsche Welle, Germany

Abstract

plain X¹ is a web-based software tool for multilingual adaptation of video, audio, and text content. The software is a 4-in-1 tool, combining several steps in the adaptation process, i.e., transcription, translation, subtitling, and voice-over, all automatically generated, but with a high level of editorial control. Users can choose which engines are used, depending on the languages and tasks. They can range from the big tech (e.g., Azure, Google, OpenAI) to smaller players (e.g., Opentrad for translation between Galician and Portuguese). As a result, plain X enables a smooth semi-automated production of subtitles or voice-over, much faster than with older, manual workflows. The software was developed out of EU research projects and has already been rolled out for professional use. Based on the European origin, the balance between technology and human expertise is strongly considered – plain X brings Artificial Intelligence (AI) into the multilingual media production process, while keeping the human in the loop.

1 Introduction

Originally, plain X was built for media broadcasters, although its use has been extended to other sectors in need of language adaptations as well. A key driver is the growing amount of content which needs language adaptation, based on user or market needs, for enhanced accessibility and/or to comply with regulation. Feature development was initially based on the needs of Deutsche Welle (DW), a world broadcaster producing news content

in over 30 languages². The plain X platform is the result of a partnership between DW as user partner and Priberam, a Portuguese scaleup that develops AI powered products based in language technologies.

The platform simplifies the multilingual adaptation process to a large degree, enabling easy subtitling in source and any target language requirement. To ensure the best possible results the software strongly relies on a “human in the loop” approach, by providing editorial tools to combine AI and human language expertise. After being rolled out for daily use in Deutsche Welle, several other clients are already using plain X, based on a software-as-a-service (SaaS) subscription model.

2 Challenges

The concept for plain X originated from the need to produce more with less, i.e., to use automation in the production process, so media producers can increase the volume of certain target languages, distribute content in more languages, or use synthetic voice, allowing to reach more people in their own spoken tongue, including in specific African or Asian regions. Since the rollout, conversations with users show a strong and growing demand to better serve regions with more than one language (for example, Spain) or large groups of immigrants (Germany). So far, the sheer volume of work was often considered simply too high, which could change with workflow tools like plain X.

Another key element of plain X is that the platform is engine agnostic, foreseeing access to the best available engines now and in the future. As an example, DW produces content in so many

¹ <https://www.plainx.com>

² <https://corporate.dw.com/en/multimedia-content-in-30-languages/a-15703976>

languages, it is essential to cover as many languages as possible, in the best possible quality, through a combination of engines from carefully selected providers, for instance for transcription or translation. In plain X, users can freely switch between different transcription, translation and voice-over engines. The platform architecture allows for the update or inclusion of additional engines in a short time. The same happens with new features like diarization or voice cloning text-to-speech models. Automated benchmarking of the different models allows choosing the best default engine for each language.

Being engine agnostic by design has been key for the support of low-resource languages by integrating engines from smaller players like Opentrad, a translation engine between Spanish, Portuguese, Galician and Catalan or the one from Lesan, for major Ethiopian and Eritrean languages.

Tables 1 to 3 show the current engines and number of languages supported for transcription, translation and voice-over.

Engine	# Languages
Amberscript	39
Azure	77
Google	72
Selma	6
Speechmatics	50
Whisper	100
All engines	108

Table 1 - Current transcription engines

Engine	# Languages
Azure	111
Google	127
DeepL	30
Lesan	3
Meta (OSS)	100
Opentrad	4
UTran	3
All engines	165

Table 2 - Current translation engines

The number of supported languages in Table 2 are target languages. Considering all combinations of source and target languages, plain X currently handles more than 38.000 language pairs.

Engine	# Languages
Azure	74
Eleven Labs	29
Google	44
Selma	2
All engines	77

Table 3 - Current voice-over engines

plain X provides system default engines per language (pair)/task combination. These can be overridden per task, user or organization allowing full control of the models used. The system defaults are chosen according to evaluations of the available engines for a specific language or language pair. Whenever possible, these are made by native speakers. The plain X team proactively tries to get feedback from users of languages not yet used by anyone else, guiding them through the testing of the available alternatives. Whenever new models, new versions of existing models or new languages are added, the plain X team evaluates them and suggests the users of those languages to do the same. The system defaults are updated whenever needed to ensure the best quality at each moment in time for all users that just want to rely on the simplest yet powerful plain X experience.

As expected, most of the language quality issues were reported for low-resource languages such as Burmese, Somali and Tibetan. This has been a major driver for the plain X team to look for alternatives to “big” providers. That’s how translation engines like Lesan, Opentrad and UTran were spotted and integrated in plain X. These are just the first three because there are several small companies training such models, independently from the big technology providers. This is why we expect measurable progress for low-resource languages in the next 12 to 18 months.

As new clients started to use the platform, new features were added making plain X capable of dealing with subtitle templates for social media (portrait and square videos) and compliant with subtitling guidelines, for instance, from streaming platforms, through a flexible rule-based subtitle segmenter.

Integration with internal content management and publishing systems was also a key requirement to reach the highest level of user acceptance, beginning with DW.

3 Origin

plain X initially came out of the SUMMA multilingual media platform, funded by the European Commission's H2020 project as a basic prototype for controlled transcription and translation for media monitoring purposes within DW and BBC Monitoring.

This prototype was then further developed and funded through the Google Digital News Initiative projects speech.media and news.bridge.

Finally, Deutsche Welle, Germany's international broadcaster in need of such platform, and Priberam decided to turn the prototype into a scalable, fully operational multilingual platform for wider use, supporting the needs of broadcasters and other multilingual content producers. That was the birth of plain X, a platform which turns content from and into virtually any language.

From prototype to product launch, besides the development and improvement of new and existing features, the UI and scalability, legal requirements related to privacy and GDPR were also implemented in plain X, making it ready to the market.

4 Workflow

The goal-oriented workflow is easy to use, but very powerful, offering editorial users the comfort of their familiar workflow, yet encompassing advanced automated technologies to support them in the creative process.

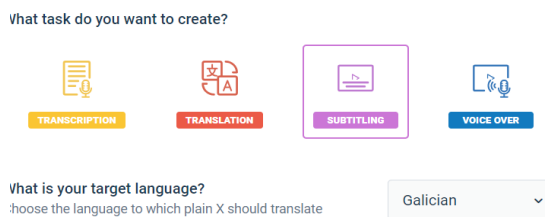


Figure 1 - Selection of goal for uploaded content

The first step is ingestion of content, be it video, audio, or text, with a growing number of supported input formats.

The next step for audiovisual content is transcription, through speech-to-text in the source language. That could be an end goal, for instance for interviews.

³ Priberam is a member of the Centre for Responsible AI (<https://centerforresponsible.ai/>)

This also allows for a primary output of automatically generated source-language subtitles, which can be used as open or closed captions.

The next step is automated translation to a selected target language, which can be post-edited, something that is transversal to the entire platform – the human in the loop, one of the Responsible AI principles³, that is, the user has always full control of the results. Again, the translation can be an end goal on its own, and used as input text for re-speaking, for example. One transcription can be translated to multiple languages.

However, it can also generate automated subtitling in the same target language. The subtitles are generated automatically taking into account not only the times given by the transcription engine but also a set of rules that can be customized. Again, the generated subtitles can be easily edited with the live preview of the subtitles in the video according to the selected template, for instance to avoid overlap of subtitles with name labels on the screen.

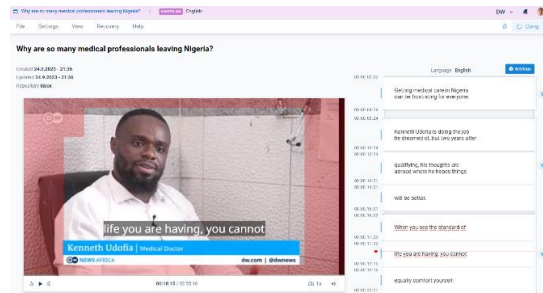


Figure 2 - Subtitling using templates

As a final step, the translation can be used for voice-over, by converting text to speech in the target language after selecting a synthetic voice. The user can also control the voice to a very high degree by applying simple commands to control the pronunciation, intonation and prosody of single words or full sentences.

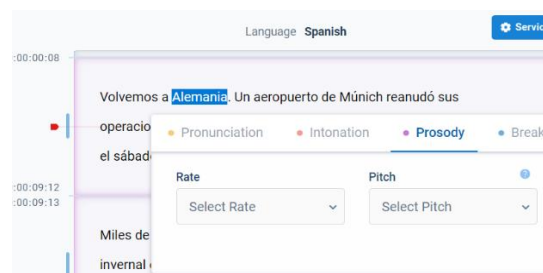


Figure 3 - Voice-over control options

Transcription and translation are the two most frequent tasks because one of the most used workflows in plain X is the transcription of audio interviews, followed or not by its translation. For video content, the combination of the sequence transcription -> translation -> subtitling is the one driving the rising usage of plain X answering a general media trend to have a much higher percentage of videos available with subtitling in a variety of languages.

Voice-over usage is still much lower because of the number of languages currently available (see Table 3) and the low quality of some of them.

The user reviewing and editing features embedded in all tasks are part of a human-centric platform that includes collaborative tools and workflows in every step, as required. Kanban-style boards clearly show the status of each task and the requests for input from a user, for instance, reviewing a translation.

Subsequently, other target languages can be added and produce equivalent content.

Estimates based on user surveys indicate time savings of 50% for transcriptions (speech to text), 30% for translations (here the human in the loop quality is the most important to reach high quality) and 70% for subtitling. These will surely rise with the quality improvement of the models.

Subtitling might be the most important task when combined with a more streamlined workflow and integration as plain X is simply faster compared to several tools.

5 Integration

It was vital to integrate this tool into the existing workflow infrastructure at Deutsche Welle and to allow for customization. This meant connecting it to input platforms for a smooth ingestion, as well as output tools for an efficient post-production and publication in the company style and branding.

Subtitle templates help to prepare the output in a particular house format. Other customizations include library management and access, setting subtitling rules, assigning roles to users, keeping track of usage and billing.

Working directly in a user environment from the start, with user input and feedback at every stage, allowed us to build a user-oriented platform to support editors in their adaptation process with the help of AI, while minimizing the feeling of insecurity and threat coming from automated processing.

The APIs used to integrate plain X with other systems can also be used for automation purposes like the fully automatic subtitling of video content from DW that is then pushed to Frankfurt Airport screens or the transcription and translation of video and audio content for media monitoring purposes. The voice-over engines in plain X can also be used to automatically generate podcasts from text content.

6 Future work

To cater for new use cases coming from current or new clients, a roadmap for the development of plain X is defined and is constantly being updated.

The “human in the loop” approach will likely gain relevance, because even at the now much higher quality of AI output, language must be treated with care and expertise.

Currently an internal benchmarking tool is used to compare the output quality of the different engines, which includes automated as well as user evaluation, and set the best rated engine as the default. We are currently working on a system where output labelled as final can also be used for benchmarking as well as training of engines and modules. As a result of the human in the loop approach and the growing number of users, the outputs from plain X will provide an increasingly reliable estimation of their quality. We will also test providing a tip to users suggesting to rerun a certain task with a different model whenever a certain threshold of mistakes is reached when editing the output of a transcription or translation task.

Some of the planned improvements already in development are the integration of quality estimation models, the ingestion of existing client terminologies and translations memories, speech Named Entity correction and the usage of LLMs to rewrite long subtitles, simplify text for more intelligible voiceover or customize the writing style of translations.

An overall policy is adherence to European privacy, data-protection and AI approaches as well as similar regulations in other geographies.

Acknowledgments

This work has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 957017, Project SELMA (<https://selma-project.eu>).

Perfil Público: Automatic Generation and Visualization of Author Profiles for Digital News Media

Nuno Guimarães
INESC TEC
University of Porto

Ricardo Campos
INESC TEC
University of Beira Interior

Alípio Jorge
INESC TEC
University of Porto

{nuno.r.guimaraes,ricardo.campos,alipio.jorge}@inesctec.pt

Abstract

Interest in the news has been declining and digital news subscriptions are still a hard sell for the average Internet user who is often used to consuming news through social media without any fees. To attract readers and engage with them, digital news outlets are forced to look for and integrate innovative solutions. In this work, we propose Perfil Público, a web platform that allows users to find news media authors based on their writing style and the topics they write about. Our solution combines a framework to generate authors' profiles automatically with a web platform that aims to facilitate the search, filter, and recommendation of digital news media authors.

1 Introduction

The massification of the Internet had an impact on the way consumers read news (Martinez-Alvarez et al., 2016). The COVID-19 pandemic only helped to accelerate the transition towards a digital-dominated media ecosystem. The declining interest in news, low newspaper sales, and only a small percentage of readers (17%) willing to pay for digital news subscriptions (Newman et al., 2023) had several impacts on the business model and format of journalism, which can only be overcome with innovative solutions and features. Several digital news media (such as the New York Times, but also Observador, Expresso, Público, and Correio da Manhã in Portugal) have adhered to subscription-only articles. Some of them, such as Jornal Público are working towards data journalism and interactive infographies to increase the number of paid subscribers. Nevertheless, a stronger commitment to the development of new solutions to captivate readers is necessary to guarantee the sustainability of the different digital news media. In that sense, we argue that one way to engage users is by allowing them to connect to authors based on their topics of interest and the author's writing style. The idea

draws parallels with book authors, where readers have their preferences based on genre and writing style.

To cope with this, we developed Perfil Público, a platform that allows readers to find digital news media authors based on their writing style and topics of interest. Towards this end, we present a framework that aims to generate each author's profile based on a time span of news articles collected from the Arquivo.pt (Gomes et al., 2013)¹. These profiles are then presented in a web platform, with search, recommendation, and filtering functionalities to promote easy navigation and captivate users' interest in the solution provided. To showcase this, we have developed a demo (<http://perfilpublico.dcc.fc.up.pt/>) on top of Público news outlet articles. We rely on Arquivo.pt for data retrieval and author profile generation, making this a scalable solution that can be easily adapted to other Portuguese news media, giving smaller and region-based news media a plug-and-play solution without the need for additional data to be stored locally.

Perfil Público methodology can be divided into two components: 1) the framework, which is responsible for the automatic generation of the authors' profiles and 2) the web platform for readers to find the authors that most suit their preferences. The framework and web platform are available in the GitHub repository².

2 Profile Generation Framework

The framework can be divided into three steps: 1) extraction of the required data 2) feature extraction at the article level and 3) authors' profile generation.

Data Extraction: The first step in creating author profiles is to collect the articles. Although,

¹<https://arquivo.pt/>

²<https://github.com/nrguimaraes/PerfilPublico>

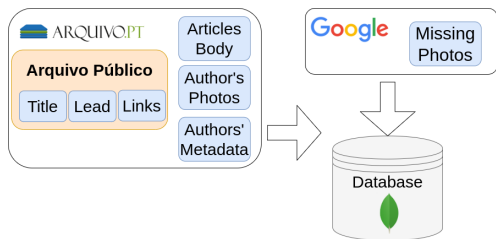


Figure 1: Data Extraction Diagram

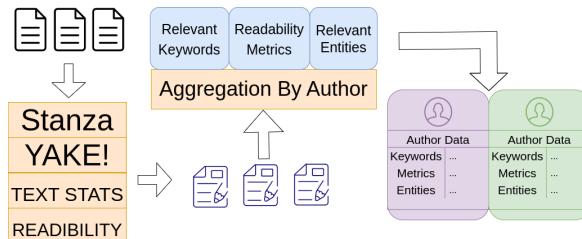


Figure 2: Feature Extraction and Profile Generation

nowadays, each digital news media has its webpage with the articles, it is not 100% guaranteed that all the articles’ history is preserved and made available to the general public. In addition, smaller or local news digital media usually do not have the resources (financial or otherwise) to maintain a large archive of past articles. Therefore, to provide a scalable solution independent of the current state of the different digital media websites, Perfil Público builds its data extraction process on top of Arquivo.pt, which provides easy access to webpages and news articles from different Portuguese digital news sources through the years.

To collect articles from Arquivo.pt we rely on Public Archive (Campos et al., 2023), a tool that facilitates the extraction of news media articles from the Portuguese web archiving infrastructure. This tool allows the extraction of the title, author, lead, and link from the archives of 5 Portuguese digital news media websites. We complement Arquivo Público with modules to extract the body of the text, as well as the author’s metadata (role and description) and photo. As a large number of authors did not have their profile photos available, we also used an unsupervised method supported by Google Images API to extract the remaining photos, using as a query the author’s name and the name of the news media. Finally, all the information retrieved was stored in a database. A diagram of the data extraction workflow is presented in Figure 1.

Feature Extraction: In this step, we first clean each article by removing possible HTML tags and non-ASCII characters. Next, we run Stanza (Qi et al., 2020) Named-Entity Recognition (NER) model for Portuguese to extract Persons, Organizations, and Locations in each article. Similarly, we applied the state-of-the-art keyword extractor Yake (Campos et al., 2020) to extract the most relevant unigram, bigram, and trigram keywords. This data is essential to get a grasp on the topics that each article addresses.

Another set of features we focus on are text statis-

tics such as the number of words and sentences, the mean of words per sentence, and the mean of syllables per word. In addition, to understand each author’s writing style, we computed four readability metrics: Flesch-Kincaid, Gunning Fog, ARI, and Coleman-Liau. These metrics were already adapted to Brazilian Portuguese (Moreno et al., 2022). However, we modified the complex word list used to better suit the European Portuguese language. Similar to (Moreno et al., 2022), we considered a complex word if is not present in the first 5000 words available in Linguateca frequent tokens resources³ after applying some filters to remove non-word entries.

Profile Generation: To automatically build each author’s profile, we first calculate the aggregated features from all the articles of an author and average the numeric ones (readability metrics and text statistics). Concerning the entities, we select the top-10 most frequent entities for each category to characterize the author. In the keywords, we used Yake score and for each different keyword, we sum all the scores. Then, we selected the top-10 most relevant keywords based on those scores for each n-gram selected. We also established 3 metrics to characterize the writing style of each author. The first, concerns the average article length of the author (measured using the number of words). The second, evaluates the readability of the author (using the average of the four readability metrics extracted). Finally, the third tries to grasp the descriptiveness of the articles by leveraging the number of entities mentioned and their diversity. The intuition is that the diversity of the entities allows a richer contextualization. For example, an article that mentions at least a location, organization, and person is closer to following the 5w1h framework used in journalism (Bleyer, 1932) to answer when, where, who, why, and how. Additional entities will further

³<https://www.linguateca.pt/aceso/tokens/formas.totalpt.txt>



Figure 3: Author profile with the features extracted



Figure 4: Articles timeline and recommendations

enrich the context of the article. These 3 metrics were computed using the features extracted from each author’s article and averaged by the collection of articles. Figure 2 presents the workflow for the feature extraction and profile generation processes.

3 Web Platform

Perfil Público web platform can be divided into three different sections.

Main Page: The main page features a search bar to search for a specific author’s name. It also presents a visual hierarchy focused on horizontal scrolling with a set of topics and the most relevant authors associated with each one of them.

Advance Filters Page: Allows users to filter authors based on the readability metrics or topics they write. It provides three range sliders to adjust the interval for each metric. In addition, it also provides the user with a search bar to find authors based on specific topics.

Author Page: Each author’s page combines the author’s metadata and profile generated. The web profile includes a profile picture, name, description, role, and the number of articles published by year. The features mentioned in Section 2 are presented in different visualizations (e.g. the author’s entities and keywords are converted in word clouds and the metrics are presented in progress bars). The web profile also integrates 1) a timeline with the

titles, publication dates, and links to each article’s preserved page in Arquivo.pt and 2) a recommendation section of authors with similar writing styles (based on the Euclidean distance of the 3 metrics proposed). Figure 3 and 4 show the features presented in each digital news media author’s profile.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 | <https://doi.org/10.54499/LA/P/0063/2020>. The authors Alípio Jorge and Ricardo Campos would like to acknowledge the National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

References

W.G. Bleyer. 1932. *Newspaper Writing and Editing*. Houghton Mifflin.

Ricardo Campos, Diogo Correia, and Adam Jatowt. 2023. *Public News Archive: A Searchable Subarchive to Portuguese Past News Articles*. In *Advances in Information Retrieval*, volume 13982, pages 211–216. Springer Nature Switzerland, Cham.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. *YAKE! Keyword extraction from single documents using multiple local features*. *Information Sciences*, 509:257–289.

Daniel Gomes, David Cruz, João Miranda, Miguel Costa, and Simão Fontes. 2013. *Search the Past with the Portuguese Web Archive*. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 321–324. Association for Computing Machinery.

Miguel Martinez-Alvarez, Udo Kruschwitz, Gabriella Kazai, Frank Hopfgartner, David Corney, Ricardo Campos, and Dyaal Albakour. 2016. *First International Workshop on Recent Trends in News Information Retrieval (NewsIR’16)*. In *Advances in Information Retrieval*, volume 9626, pages 878–882. Springer International Publishing.

Gleice Carvalho de Lima Moreno, Marco P. M. de Souza, Nelson Hein, and Adriana Kroenke Hein. 2022. *ALT: um software para análise de legibilidade de textos em língua Portuguesa*. ArXiv:2203.12135.

Nic Newman, Richard Fletcher, Kirsten Eddy, Craig T Robinson, and Rasmus Kleis Nielsen. 2023. [Reuters Institute digital news report 2023](#). Technical report, Reuters Institute for the Study of Journalism.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python Natural Language Processing Toolkit for Many Human Languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Exploring Open Information Extraction for the Portuguese language: An integrated monolithic approach in Cloud environment

Augusto Sampaio Barreto Daniela Barreiro Claro
FORMAS Research Center - Institute of Computing
Federal University of Bahia – Salvador - Bahia - Brazil
augusto.barreto@ufba.br dclaro@ufba.br

Abstract

This work addresses Open Information Extraction (Open IE) as a crucial area for structuring unstructured data, aiming to identify and represent information through triples. The Open IE approach proposes a domain-independent paradigm that extracts potential relationships between entities using generalized patterns. Despite advancements, the lack of studies in the Portuguese language emphasizes the need to explore specific techniques. This is worse regarding easy access to take advantage of triples extracted by OpenIE models. In this context, we present an integrated hub of services in the Open IE domain, allowing users to extract triples and compare their results. This hub is developed under a monolithic architecture with the Django framework, coupled with deployment on Google Cloud to reinforce the efficiency and adaptability of inclusion and removal of services. Within this framework, it would be possible for non-computer experts to use the advantages of OpenIE triples from the Portuguese languages.

1 Introduction

Open Information Extraction (Open IE) is a crucial area that aims to structure data from unstructured sources, with the main goal of identifying and representing information through triples that express relationships. The Open IE approach represents a domain-independent extraction paradigm, using generalized patterns to extract all potential relationships between entities (Etzioni et al., 2008).

Triples are essential for capturing the meaning of information present in unstructured data. Despite advances in the last decade, most of this progress has focused on the English language, with few studies dedicated to Portuguese in the last five years (Bender, 2019).

The limitation of studies in the Portuguese language emphasizes the need to explore and enhance

the application of Open IE techniques in such language. However, works in this area publish their results as a package, making it difficult to use by non-computer experts. This article aims to contribute to this gap by presenting a hub of services that integrate methods from Open IE in Portuguese, enabling a single environment for users to extract triples and compare the results.

Our framework deserves a hub of Open IE services that implements each service as an Open IE method for the Portuguese language. We call this hub of Open IE services as FORMAS Open IE Framework.

This article is structured as follows: the next section describes the background. Section 3 presents each Open IE method. Section 4 describes our hub of Open IE services. Section 5 discusses our conclusions and some envisioning work.

2 Background

The advantages of Open IE, as outlined by (Gamallo, 2012), encompass domain independence, unsupervised extraction, and greater scalability. These features highlight Open IE's adaptability, efficiency, and scalability in handling diverse subjects and extensive volumes of unstructured text.

Concerning architecture, the system's architectural choice is pivotal in shaping Open IE systems' development and maintenance. Various paradigms, including monolithic, microservices, layered, and event-driven architectures, offer distinct characteristics influencing scalability, flexibility, and modularity (Fowler, 2003; Newman, 2015a; Richardson, 2018).

A brief exploration of these architectural approaches follows:

Monolithic Architecture: Integrates all components into a single application, simplifying development and maintenance. While suitable for applications with precise requirements, it

080 may face scalability challenges as the system
081 grows (Fowler, 2003).
082 *Microservices Architecture*: Divides a system
083 into independent services, promoting scala-
084 bility and independent evolution. Facilitates
085 modular development, updates, and mainte-
086 nance (Newman, 2015a).
087 *Layered Architecture*: Organizes the system
088 into abstraction levels, promoting modularity.
089 While simplifying maintenance, inter-layer
090 dependencies may limit scalability (Fowler,
091 2003).
092 *Event-Driven Architecture*: Components com-
093 municate through asynchronous events, fa-
094 voring scalability and dynamic responsive-
095 ness. Allows distributed processing but re-
096 quires careful management of asynchronous
097 events (Richardson, 2018).

098 The choice of architecture depends on project-
099 specific needs, with monolithic architectures of-
100 fering simplicity, microservices providing scalabil-
101 ity, layered architectures ensuring modularity and
102 event-driven architectures supporting reactivity in
103 distributed systems. Each decision entails profound
104 implications for system evolution and maintenance,
105 emphasizing the importance of considering project
106 characteristics and requirements.

107 3 Services

108 Services are implemented as Open IE methods. A
109 set of OpenIE methods for the Portuguese language
110 was selected to be part of the first version of this
111 hub of services: DptOIE, PTOiE-Flair, and Chat-
112 GPT. We detailed each one as follows.

113 3.1 DptOIE

114 DptOIE (Oliveira et al., 2023) is a method de-
115 veloped for Open Information Extraction (OIE),
116 specifically designed for the Portuguese language.
117 The main objective of DptOIE is to extract valuable
118 information or "facts" from sentences by analyzing
119 their syntactic structure and dependencies. DptOIE
120 has three main phases: Pre-processing, triple ex-
121 traction and special cases.

122 The preprocessing carries out a *Tokenization*,
123 *Part-of-Speech* (POS) tagging, and *Dependency*
124 *Analysis* to inputted sentences. DptOIE identifies
125 triples (subject, relation, object) by traversing the
126 dependency tree using a Depth-First Search (DFS)
127 approach. Each triple consists of an argument
128 (Arg1), a relation, and the second argument (Arg2).

DptOIE addresses specific linguistic construc- 129
tions, such as: 130

- Coordinative Conjunctions (CC): Handles 131
conjunctions like "and" or "or" to generate 132
multiple triples from a single sentence. 133
- Subordinate Clauses:Manages adjective, ad- 134
verbial, and substantive clauses, linking them 135
to the main clause to form coherent triples. 136
- Appositives: Derives additional triples from 137
sentences with appositives, creating synthetic 138
clauses. 139

140 Consider the sentence "O diretor do hospital,
141 Júlio, vendeu sua fazenda." DptOIE extracts the
142 main triple: (O diretor do hospital; vendeu; sua
143 fazenda). Moreover, it recognizes the appositive
144 "Júlio" and generates a new triple: (O diretor do
145 hospital; é; Júlio). Additionally, it applies transi-
146 tivity to create an additional triple: (Júlio; vendeu;
147 sua fazenda).

148 3.2 PTOiE-Flair

149 The PTOiE-Flair is a new OpenIE model based on
150 deep neural networks that enables the generation of
151 triples given a sentence. PTOiE-Flair was trained
152 with two datasets, LSOI and S2, achieving SOTA
153 results for the Portuguese language.

154 Consider the sentence "Os cachorros, que são
155 mamíferos, são os melhores amigos do homem."
156 PortNOIE extracts two triples: ["Os cachor-
157 ros"/ARG0, "são"/V, "mamíferos"/ARG1, "são"/V,
158 "os melhores amigos do homem"/ARG1], that is:
159 (i) *Os cachorros; são; os melhores amigos do*
160 *homem* and (ii) *Os cachorros; são; mamíferos*.

161 3.3 ChatGPT

162 The use of ChatGPT for triple extraction repre-
163 sents an innovative approach to integrating natu-
164 ral language technologies with the extraction of
165 structured information. Considering the ChatGPT,
166 carefully formulating prompts is essential to guide
167 the model in generating structured responses.

168 Prompt example: "Provide a triple containing
169 a subject, a relation, and an object based on the
170 following statement: 'The event occurred when'."

171 We employed the ChatGPT as a service, pro-
172 viding custom prompts and receiving structured
173 responses as a triple structure. Initially, we utilized
174 the davinci-003 variant of the ChatGPT model.
175 However, in future deployments, new models such

as GPT-4 or GPT-4-turbo could be seamlessly integrated without compromising performance.

Consider the prompt: "Describe a situation in which" followed by a specific context. ChatGPT generates a structured response, such as "(a scientist; makes; a significant discovery)".

4 FORMAS OpenIE Framework

The monolithic architecture based on Django framework simplifies the development and maintenance of robust systems. According to (Newman, 2015b; Fowler, 2014), monolithic architectures provide an integrated approach, making implementing and managing functionalities easier.

Our architecture comprises two modules as depicted in Figure 1: a Frontend and a Backend. Our front end was integrated with Django, and our back end was initially developed as a service hub with three methods: Chatgpt, Ptoie-flair, and DptOIE.

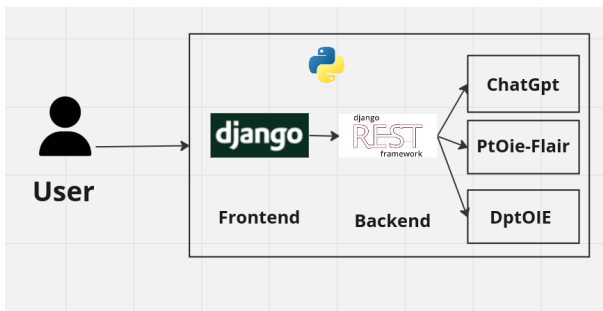


Figure 1: Architecture

Integration with Django enables the creation of a cohesive application, facilitating efficient communication between components. As mentioned by (Mokhtar, 2018), the use of monoliths is particularly advantageous when a pragmatic and efficiency-centered approach is sought, making it ideal for applications with clear and well-defined requirements.

In addition to the monolithic architecture, deploying the service on Cloud enhances resource adaptability for users. Utilizing cloud services provides dynamic scalability, allowing adjustments based on demand. According to (Kavis, 2014), Cloud services offers a reliable and flexible infrastructure, ensuring that users have access to resources tailored to their needs.

4.1 Graphical Interface

Our framework has a graphical interface depicted in Figure 2 that enables the use of visual elements to extract triples.



Figure 2: Interface

The user introduces a sentence in Portuguese, and our Framework answers with the triple extracted. The blue rectangle is omitted for anonymous review, and neither is the line under the title. Firstly, the user describes the sentence and selects which OpenIE method would prefer to extract the triples. Results are shown immediately after the button.

5 Conclusions and Future Work

The integration between the monolithic architecture using Django and deployment on the Cloud reinforces the robustness and adaptability of the Open Information Extraction services. This approach, supported by solid theoretical foundations, allows non-computer users to evaluate the extraction on three OpenIE methods.

In this work, while the primary goal was to provide a unified platform for users to access and utilize different OpenIE methods, the comparative analysis of these methods was not within the scope. Such comparisons entail diverse performance metrics and evaluation criteria, which could vary depending on the specific natural language processing tasks at hand. However, as future work, conducting comparative evaluations across various NLP tasks could offer valuable insights into the strengths and weaknesses of different extraction methods. By expanding the scope of evaluation beyond OpenIE,

242 we can further refine and optimize these methods to
243 better serve the needs of the Portuguese language
244 community. We envision enhancing the implementa-
245 tion and including more methods for the Open IE
246 Portuguese language community.

247 **Acknowledgments**

248 This material is partially based upon work sup-
249 ported by the FAPESB under grant INCITE
250 PIE0002/2022 and FAPESB TIC 0002/2015 and
251 CAPES Financial code 001.

252 **References**

- 253 Emily M. Bender. 2019. On the lack of study of non-
254 english languages in nlp. In *Proceedings of the First*
255 *ACL Workshop on Ethics in NLP*, pages 7–13.
- 256 Oren Etzioni, Michele Banko, Stephen Soderland, and
257 Daniel S. Weld. 2008. [Open information extraction](#)
258 [from the web](#). *Commun. ACM*, 51(12):68–74.
- 259 Martin Fowler. 2003. *Patterns of Enterprise Application*
260 *Architecture*. Addison-Wesley.
- 261 Martin Fowler. 2014. [Microservices](#).
- 262 Pablo Gamallo. 2012. Overview of open information
263 extraction. In *Open Information Extraction: Volume*
264 *377 of CEUR Workshop Proceedings*, pages 1–13.
- 265 Mike Kavis. 2014. *Architecting the Cloud: Design De-*
266 *isions for Cloud Computing Service Models*. John
267 Wiley & Sons.
- 268 Khaled Mokhtar. 2018. [The advantages and disadvan-](#)
269 [tages of monolithic and microservices architectures](#).
- 270 Sam Newman. 2015a. *Building Microservices: Design-*
271 *ing Fine-Grained Systems*. O’Reilly Media.
- 272 Sam Newman. 2015b. *Building Microservices: De-*
273 *signing Fine-Grained Systems*, 1st edition. O’Reilly
274 Media.
- 275 Leandro Oliveira, Daniela Barreiro Claro, and Marlo
276 Souza. 2023. [Dptoie: a portuguese open information](#)
277 [extraction based on dependency analysis](#). *Artif. Intell.*
278 *Rev.*, 56(7):7015–7046.
- 279 Chris Richardson. 2018. *Microservices Patterns: With*
280 *Examples in Java*. Manning Publications.

Blip Copilot: a smart conversational assistant

Evandro Fonseca, Tayane Soares, Dyovana Baptista,
Rogers Damas and Lucas Avanço

Blip

evandro.fonseca, tayane.soares, dyovana.baptista,
rogers, lucas.avanco
{ @blip.ai }

Abstract

This paper describes Blip Copilot plugin, an AI-based assistant that provides quick and smart suggested answers for an enriched conversational experience.

1 Introduction

Typically, customer service chats in large and mid-sized companies can face high demand, causing attendants to become overwhelmed and resulting in delays in responding to customers. Considering this overload and current advances in the development of language models in Natural Language Processing (NLP) (Brown et al., 2020), in this paper we present Blip Copilot. Blip Copilot is an assistant that optimizes customer service on Blip Desk¹. With this extension, attendants can access personalized answer suggestions provided by a language model that takes into account a knowledge base built for a specific context or domain. All copilot suggestions are generated considering the conversational context (thread of messages in a chat) and the knowledge base provided during setup. Our approach is fully supervised. It is up to attendant to decide select, discard or edit the provided suggestions. To generate more accurate responses, we use NLP techniques to process, extract and find relevant information regarding current topic conversation. It will be more described in subsequent sections.

2 Architecture

Our architecture is language and LLM model agnostic. In our experiments for the Portuguese language, we tested two models: PaLM-2(Anil et al., 2023) and GPT-3.5-turbo(Brown et al., 2020).

¹Blip Desk is a customer service tool that allows a chatbot to redirect (overflow) a user's conversation to a human attendant on different channels. <https://help.blip.ai/hc/pt-br/articles/4474416681495-Visão-geral-do-Blip-Desk>

However, GPT-3.5-turbo has presented more accurate responses than PaLM-2. When developing applications that incorporate LLM into their pipeline, it is important to consider the limitations of these models. The texts generated by these tools may be biased, and their answers will not always be assertive (Bender et al., 2021). Therefore, it is always necessary to always validate the output of applications that use LLMs. Given these limitations, Blip Copilot combines the use of Retrieval Augmented Generation (RAG) (Lewis et al., 2020) with the prompt provided to the LLM to improve the accuracy of suggested responses. In this way, suggested responses are generated considering the context of the conversation in the customer service chat and the context retrieved from the specific domain knowledge base provided during plugin configuration. Before using Copilot, it is necessary to configure the knowledge base and other specific features, such as the company name, service demands, and other information². Throughout this process, each entry of knowledge base is embedded and stored in a vector database. Once the setup process is complete, copilot is ready to use. Figure 1 shows the Copilot pipeline. When the attendant (user) calls Blip Copilot, an embedding vector of the last n^3 messages are extracted. The next step is to look for the most similar instance previously built in the setup stage. For this, we use cosine similarity(Rahutomo et al., 2012). Finally, we build a LLM call using the conversational context, the retrieved knowledge base instance, and the entire customer setup information.

3 Interface

In order to provide an easy setup, we turn available a web interface, which consists in allow specific configurations, such as: brand name, service

²see section 3 –Interface

³the window of messages is a configuration parameter

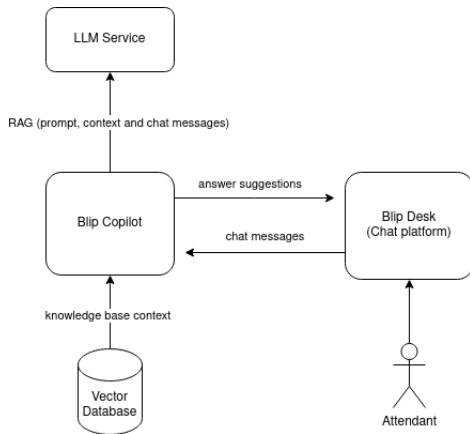


Figure 1: Copilot architecture

demands, profile, additional rules(observations) among others. Our interface is distributed in four tabs: Basic setup, knowledge database upload, greeting messages setup and advanced configurations.

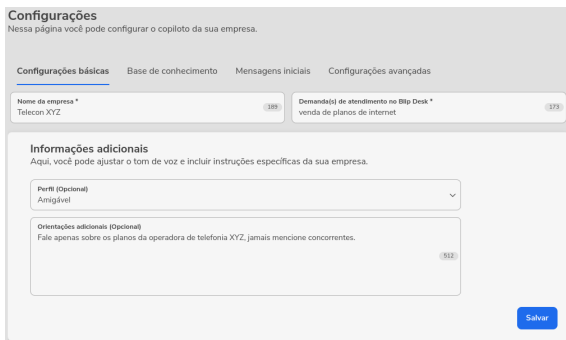


Figure 2: Basic setup

In figure 2 there specific configurations that addresses the copilot behavior. Each presented field is very important because the model follows a specific prompt which considers user data to provide the suggested answers. In the mentioned image, we can see that this copilot represents a Telecom company, acts in Sales of internet plans, should provide responses using a friendly discourse and never talk about other Telecom providers.

In Figure 3 the database upload interface is presented. The knowledge base is very useful to address information about business rules or specific products and its restrictions, for example. Regarding data structure, Blip Copilot accepts two file formats(txt and tsv), each line represents an instance. Moreover it is possible to provide some contextual information, complying with the following pattern: “Topic | Description”. Ex:

- Basic Plan | The Basic Plan is an option for

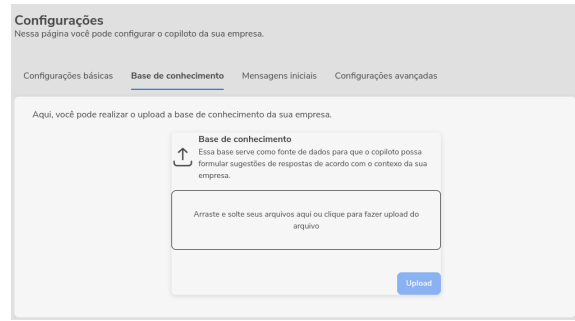


Figure 3: Knowledge base upload interface

basic internet needs. With speeds of up to 10 Mbps ...

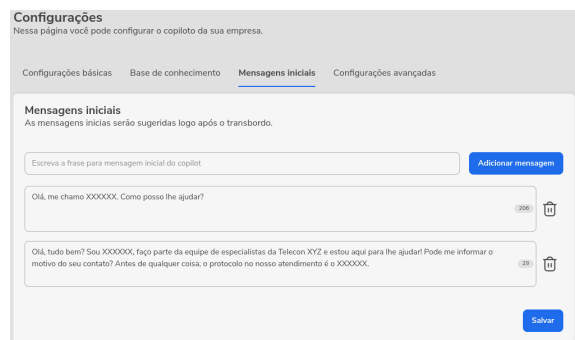


Figure 4: Greeting messages setup

In Figure 4 we present the greeting messages setup interface. The greeting messages are provided as a first suggestion. It is useful to introduce a chat.

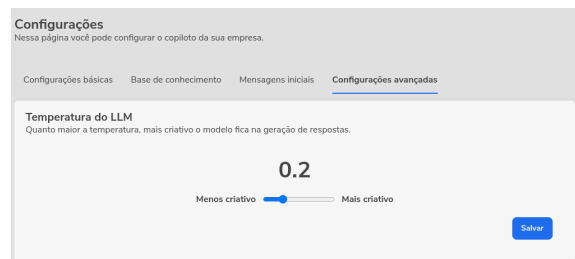


Figure 5: Advanced settings

The Figure 5 refers to model temperature. This parameter influences its "creativity" or randomness and is widely used by many LLM models. However, when we use high temperatures there is a risk of model "hallucinating", that is, high temperatures increase the randomness of the model and can generate answers with low assertiveness or that are wrong.

Finally, we show two cases and their outputs. In Figure 6, it is clear that the customer needs a second copy of his bill, and in Figure 7 the customer asks

about available internet plans. Here it is clear that copilot makes use of customer knowledge base to retrieve the correct plan names and values.

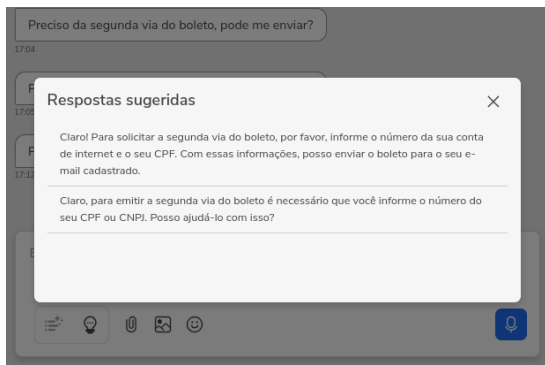


Figure 6: Suggested answers - invoice

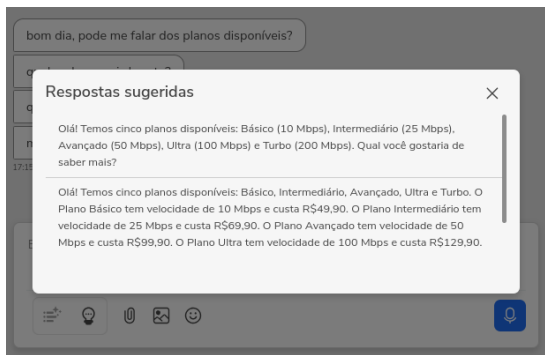


Figure 7: Suggested answers - internet plans

4 Experiments and Results

In order to show the relevance of Blip Copilot use, we have conducted an experiment involving five participants. Basically we compared each attendant with itself, considering the number of closed tickets and the average time of each service with and without Blip Copilot. For this experiments two window time were considered: September 2023 (without copilot) and November 2023 (with copilot). We skipped October because we believe there is a learning curve in refining the knowledge base to produce more accurate responses. Thus, each attendant had one month to learning how to use copilot and to obtain a better performance of its use.

Attendant	Closed Tickets Without Copilot	Closed Tickets Using Copilot	Avg. Time Without Copilot	Avg. Time With Copilot
attendant 1	34	773	01:00:12	00:15:54
attendant 2	568	628	00:36:10	00:17:28
attendant 3	875	823	00:15:00	00:08:31
attendant 4	651	782	00:22:11	00:11:41
attendant 5	612	719	00:28:40	00:14:33

Table 1: Results

As a result, it is possible to see that there is a significant improvement in average service time when Copilot is used. All attendants considerably reduced their time, and except for Attendant 3, all other participants increased the amount of their closed tickets.

5 Conclusion

In this paper, we presented Blip Copilot, a smart conversational assistant that considers the chat context, the brand/customer database, knowledge base, and custom parameters to suggest accurate responses. We also showed our Copilot architecture and how this chat assistant can improve the day-to-day of many brands/businesses, reducing the service time and maximizing the attendants efficiency. Also, the Copilot smooths the onboarding process of new attendants. As further work, we intend to add a feedback system so that Copilot can learn from user feedback and become more accurate over time. We also want to integrate our Copilot with Blip Desk mobile version⁴.

References

- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Arimitsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1.

⁴A Blip Desk version for smartphones

Galician–Portuguese Neural Machine Translation System

Sofía García González

imaxin|software / Salgueiriños de Abaixo, N 11 L6, Santiago de Compostela
sofia.garcia@imaxin.com

Abstract

This paper presents the first Galician–Portuguese (GL–PT) bilingual neural machine translation (NMT) model. Due to the lack of Galician–Portuguese parallel data, this model was trained on synthetic data converting the Spanish part from original Spanish–Portuguese corpora to Galician using the RBMT system Apertium.

1 Introduction

In recent years, neural machine translation (NMT) has become the state-of-the-art in this natural language processing (NLP) area. It has shown promising results in various language pairs. However, developing efficient translation models for low-resource languages such as Galician is challenging due to the need for large training parallel corpus (Haddow et al., 2022).

O Proxecto Nós (The Nós Project) has currently developed neural MT models for Spanish–Galician¹ and English–Galician² pairs in both directions. These models were trained converting the Portuguese part from original English–Portuguese and Spanish–Portuguese corpora to Galician. (Ortega et al., 2022). However, there is currently no NMT system for Portuguese–Galician pair, except for multilingual models where Galician is included as M2M (Fan et al., 2021) or NLLB (Costa-jussà et al., 2022). Furthermore, despite the closeness of these two languages, both the RBMT system Apertium (Forcada et al., 2011) and the port2gal³ transliterator perform poorly in both translation directions, particularly to put it into production as a company.

Therefore, this paper presents a Galician–Portuguese neural translation model tailored to the

¹https://huggingface.co/proxectonos/Nos_MT-0penNMT-es-gl

²https://huggingface.co/proxectonos/Nos_MT-0penNMT-en-gl

³<https://fegalaz.usc.es/~gamallo/port2gal.htm>

administrative domain, which imaxin|software provides to clients such as the *Xunta de Galicia* (Galician Government) with GAI0⁴ or the Galician Parliament.

2 Methodology

2.1 Training Corpora

In accordance with the de Dios-Flores et al. (2022) strategy, the process was divided into two steps. Firstly, we gathered two Spanish–Portuguese parallel macrocorpora: CCMatrix,⁵ and OpenSubtitles v2018.⁶; and a legal-domain corpus: the Spanish–Portuguese DGT v8⁷ (see Table 1 for corpus sizes). Then, using the RBMT system developed for GAI0, we created synthetic corpora translating the Spanish part into Galician, in order to obtain synthetic Portuguese–Galician parallel corpora.

Domain	Dataset	Number of Sentences
General Domain	CCMatrix	25M
	OpenSubtitles	25M
Legal Domain	DGT v2019	3.5M

Table 1: Spanish–Portuguese training corpus sizes

2.2 Architecture

Regarding the training process, we have used the Transformer architecture from OpenNMT-py⁸ open-source framework. For this initial model, we have assigned greater weight to the generic CCMatrix and OpenSubtitles corpora, with weights of 50 for both macrocorpora, while the DGT corpus had a weight of 20. The training parameters can be seen in Table 2.

⁴*Xunta de Galicia*’s MT system based on Apertium, <http://tradutorgaio.xunta.gal/TradutorPublico/traducir/index>.

⁵<https://opus.nlpl.eu/CCMatrix-v1.php>. We only used the half size of CCMatrix. Thus, we selected 25M random sentences

⁶<https://opus.nlpl.eu/OpenSubtitles-v2018.php>

⁷<https://opus.nlpl.eu/DGT-v2019.php>

⁸<https://github.com/OpenNMT/OpenNMT-py>

Parameters	Values
Model	Transformer
dropout	0.1
average_decay	0.0005
label_smoothing	0.1
optimization	adam
learning_rate	2
warmup_steps	8000
batch_size	8192

Table 2: Training Parameters

2.3 Evaluation

The corpora used to evaluate the NMT model were: Flores200-dev (Goyal et al., 2022)⁹, News Test References for MT Evaluation (NTREX) (Barrault et al., 2019)¹⁰ and a 1k corpus extracted from CCMatrix. See Table 3 for sizes¹¹.

Evaluation Dataset	Size
Flores200-dev	1k
NTREX	2k
CCMatrix-test-dataset	1k

Table 3: Portuguese-Galician Evaluation test sizes

On the other hand, we used the Sacrebleu framework¹² as recommended by Post (2018). This framework includes: BLEU (Papineni et al., 2002), chrF (Popović, 2015) and TER (Snover et al., 2006) metrics. Moreover, we also used the current state-of-the-art COMET (Rei et al., 2022)¹³.

3 Results

The following tables report the results for each evaluation dataset: Flores200-dev (Table 4), NTRIX (Table 5) and CCMatrix (Table 6). We have used *Apertium* as the baseline to compare our results.

MT Systems	BLEU	chrF	TER	COMET
Apertium	21.3	52	62.8	0.824
imaxin software model	24.2	54.3	61.2	0.769

Table 4: Flores200-dev results in gl-pt systems

⁹<https://github.com/facebookresearch/flores/tree/main/flores200>

¹⁰<https://github.com/MicrosoftTranslator/NTREX>

¹¹Because of the lack of legal-domain test datasets in this language pair, we have not been able to make a specific evaluation in this domain.

¹²<https://pypi.org/project/sacrebleu/>

¹³We have used the wmt22-comet-da model

MT Systems	BLEU	chrF	TER	COMET
Apertium	23	53.4	63.3	0.810
imaxin software model	21.6	51.9	64.6	0.745

Table 5: NTRIX results in gl-pt systems

MT Systems	BLEU	chrF	TER	COMET
Apertium	41.6	69.4	51.3	0.848
imaxin software model	32.7	69.1	52	0.888

Table 6: CCMatrix test results in gl-pt systems

4 Analysis

As shown in the tables, with the exception of the flores200-dev test (Table 4), Apertium continues to outperform our NMT model. The difference in results is particularly remarkable on the test taken from the CCMatrix corpus (Table 6), where Apertium outperforms the neural model by 10 BLEU points. However, both translation systems yield unsatisfactory results for two closely related languages. The absence of an authentic Galician-Portuguese corpus poses a challenge for developing good quality NMT models. In fact, one of the major issues with macrocorpora such as CCMatrix is that they mix variants of Portuguese from Portugal and Brazil, resulting in inconsistent language during translation. That is, they are unable to maintain the same variant throughout the translation process. On the other hand, Apertium does not present this issue, as it is a system designed to translate to and from the European variant of Portuguese. Therefore, in the future, a more in-depth analysis is necessary to determine how different varieties of Portuguese affect NMT models development.

5 Conclusions

This demo model provides a starting point for NMT between Galician and Portuguese. In the future, other strategies will be tested, such as deeper cleaning of the web-extracted corpora, distinguishing between Portuguese variants, or creating legal test corpora for this language pair, which currently does not exist and hinders accurate evaluation for this domain. The development of high-quality parallel corpora will be crucial for the future development of NMT models.

6 Demonstration

Our demonstration will be show on an **imaxin**software webpage where users will be able to translate any text from Galician to

Portuguese to test this model.

References

- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Iria de Dios-Flores, Carmen Magarinos, Adina Ioana Vladu, John E Ortega, José Ramom Pichel Campos, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín Diz, Manuel González González, et al. 2022. The nós project: Opening routes for the galician language in the field of language technologies. In *Proceedings of the workshop towards digital language equality within the 13th language resources and evaluation conference*, pages 52–61.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.
- Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. [Survey of low-resource machine translation](#). *Computational Linguistics*, 48(3):673–732.
- John E Ortega, Iria de Dios-Flores, Pablo Gamallo, and José Ramom Pichel. 2022. A neural machine translation system for galician from transliterated portuguese text. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing. CEUR Workshop Proceedings*, volume 3224, pages 92–95.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Nós-TTS: a Web User Interface for Galician Text-to-Speech

Carmen Magariños¹, Alp Öktem², Antonio Moscoso Sánchez¹, Marta Vázquez Abuín¹, Noelia García Díaz¹, Adina Ioana Vladu¹, Elisa Fernández Rei¹, María Baqueiro Vidal³

¹Instituto da Lingua Galega (ILG), Universidade de Santiago de Compostela, Spain

²Col-lectivaT, Spain

³Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain

Abstract

Speech synthesis, also known as text-to-speech (TTS), aims to generate human-like speech from text. The recent emergence of end-to-end deep learning TTS models has led to impressive natural-sounding results. Nevertheless, expanding these models to multiple languages and speakers poses challenges, especially for low- or limited-resource languages. In this context, we introduce Nós-TTS, a user-friendly web interface for Galician TTS developed under the Nós project. The proposed interface offers a choice among three distinct voices trained on diverse conditions regarding data quantity, training approach, and input modality. Although in an experimental stage, informal listening tests have shown a satisfactory performance of the models.

1 Introduction

Speech synthesis, also referred to as text-to-speech (TTS), is the automated generation of human-like speech by machines or computers (Dutoit, 1997; Taylor, 2009). More specifically, TTS techniques aim to synthesize intelligible and natural speech from input text. Over the years, different TTS approaches have been proposed (Tabet and Boughazi, 2011), the most prominent being concatenative unit-selection (Black and Campbell, 1995; Hunt and Black, 1996) and statistical parametric synthesis (Black et al., 2007; Zen et al., 2009).

In recent years, deep learning-based TTS systems have emerged as a powerful alternative to traditional synthesis (Ning et al., 2019; Tan et al., 2021). These systems use deep neural networks (DNNs) as the model backbone and have shown the ability to produce high-quality natural-sounding speech. However, these models often rely on massive single-speaker datasets (20-40 hours) for optimal performance. This high data demand poses a significant drawback, particularly for languages with limited resources, such as Galician, since ac-

quiring this data can be costly and time-consuming. While various techniques like transfer learning, multilingual training, or zero-shot learning have been applied to alleviate this problem (Casanova et al., 2022), their effectiveness still depends on factors such as the data quality, quantity, and the specific traits of the target languages.

Within the framework of the Nós project (Vladu et al., 2022; de Dios-Flores et al., 2022), we present Nós-TTS, a user-friendly web interface for Galician text-to-speech conversion. Nós-TTS allows users to input a text in Galician and synthesize the corresponding speech using one of three distinct voices: Celtia, Sabela, or Icíá. The underlying voice models are built on Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) architecture (Kim et al., 2021) and were trained using the Coqui-TTS library (Eren and The Coqui TTS Team, 2021) on different Galician TTS datasets.

The following sections briefly describe the proposed system, discuss the evaluation strategy and main findings, and summarize future work.

2 System Description

Nós-TTS¹ is a web user interface for speech synthesis in Galician. As shown in Figure 1, the proposed interface takes an input text in Galician (up to 1000 characters) and generates the corresponding synthesized speech by clicking the “Xerar voz” button. Once available, the audio will automatically start playing, and audio controls will be shown, including play, pause, volume, and position bar. Depending on the browser, the playback speed may also be adjusted. Other available features are download the synthesized audio (down arrow button) and clear the input text (“Borrar texto” button).

¹<https://tts.nos.gal/>

2.1 Voice models

In its current version, the TTS system integrates three voice models with different characteristics:

- **Celtia.** Female voice model trained from scratch on a subset of the Nos_Celtia-GL corpus (Vázquez Abuín et al., 2023). This corpus, created under the Nós project, comprises a total of 20,000 sentences recorded by a professional voice talent. Specifically, a subset of 13,000 sentences were used to train the model, which corresponds to 15.5 hours of speech. The Celtia model (Magariños, 2023) was trained directly on grapheme inputs and includes a text normalization step based on the front-end of Cotovía (Rodríguez Banga et al., 2012).
- **Sabela.** Female voice model trained from scratch on the Sabela corpus within the CRPIH UVigo-GL-Voices dataset (CRPIH and GTM, 2023). This corpus comprises 10,000 sentences recorded by a professional radio broadcaster, amounting to approximately 14 hours of speech. The Sabela model (Öktem et al., 2023) was trained on phonemes and incorporates the Cotovía front-end for text normalization and grapheme-to-phoneme conversion.
- **Icía.** Female voice model fine-tuned from the previously described Celtia model using the Icía corpus within the CRPIH UVigo-GL-Voices dataset (CRPIH and GTM, 2023). The Icía corpus comprises around 3,000 sentences, equivalent to approximately 4 hours of speech, recorded by an amateur speaker. Icía (Moscoso et al., 2023) is a phoneme-based model which integrates the front-end of Cotovía for both text normalization and phonetic transcription.

All the models are openly available in Hugging Face².

2.2 Models' Architecture

The incorporated voice models are based on VITS (Kim et al., 2021), a fully end-to-end TTS model that leverages cutting-edge deep-learning techniques like adversarial learning (Goodfellow et al., 2014), normalizing flows (Rezende and Mohamed,

²<https://huggingface.co/proxectonos>

2015), variational auto-encoders (Kingma and Welling, 2014) and transformers (Vaswani et al., 2017) to achieve results comparable to ground truth. Its architecture combines the Glow-TTS encoder (Kim et al., 2020) and HiFi-GAN vocoder (Kong et al., 2020) within the same training pipeline. By jointly learning the acoustic model and the vocoder, VITS overcomes some issues of the two-stage models. It also incorporates a stochastic duration predictor that allows synthesizing speech with different rhythms from the same input text.

2.3 Text Processing Module

The proposed TTS system includes a text processing module based on the Cotovía front-end. Depending on the selected voice, as described in Section 2.1, the text processing module performs one or both of the following functions: (1) text normalization; (2) phonetic transcription with stress marks.



Figure 1: View of the NÓS-TTS user interface.

3 Evaluation and Discussion

Traditionally, the quality of TTS systems is assessed through perceptual listening tests with human subjects. These tests commonly employ perceptual metrics, such as Mean Opinion Score (MOS) (Ling et al., 2021), to rate speech characteristics including overall quality, naturalness, or similarity to the target voice.

This form of subjective measures are the gold standard for the speech synthesis task, yet it proves to be time-consuming and demanding in terms of test preparation and listener recruitment. Consequently, models are typically initially assessed through informal listening tests, with more extensive formal evaluations reserved for final models.

While the models currently integrated into the NÓS-TTS interface are experimental, they have demonstrated competent performance in informal

listening tests, showcasing high naturalness and quality. For each voice, the models exhibiting superior performance in these informal evaluations will undergo subsequent formal evaluations involving a statistically significant number of listeners.

Nevertheless, these informal evaluations reveal insightful findings regarding the performance of the three voice models. Notably, the Celtia model stands out as the undisputed leader in terms of overall quality. Its superior results in audio quality, choice of pronunciation, precision of phonemes and naturalness of prosody, position it as the most robust and satisfactory option. This outstanding performance is directly attributed to the quality of the corpus used in its training, which was meticulously designed and developed to ensure balanced and representative textual content, voice talent with good vocal characteristics, and high-quality recordings.

Second in the ranking, the Icíá model is positioned as a solid alternative, despite the limited amount of data and the speaker being non-professional. In this case, the applied fine-tuning techniques have mitigated the data limitations, resulting in a synthetic voice with noticeably more natural prosody compared to the Sabela model.

On the other hand, the Sabela model faces significant challenges, primarily related to the lack of naturalness in prosody. This limitation is evident both in the original recordings used for training and in the generated synthetic voice. The main reason for this lack of naturalness seems to be the particular style and rigid prosody associated with typical news readings on radio and television used during the recordings. This finding underscores the importance of considering not only the quantity but also the quality and diversity of training data to achieve optimal results in speech synthesis.

Another important consideration when comparing the different models pertains to the input modality for training, namely graphemes versus phonemes. A model trained on phonemes is expected to converge more rapidly, and using phonemes as input is also anticipated to aid in disambiguating the pronunciation of specific graphemes. An example of this is the grapheme <x>, which in Galician can be pronounced as [ks] or [j] depending on the word. In this particular case, we have observed that the Celtia model, trained on graphemes, mispronounces this grapheme in some words (e.g., <x> as [j] instead of [ks] in *boxeo* and *axila*), whereas the Icíá and Sabela models, trained

on phonemes, correctly differentiate between the two pronunciations. We aim to address this minor drawback of the Celtia model by training a new model based on phonemes. This final comparison reveals the importance of having a proficient text processing module to achieve precise phonetic transcriptions.

4 Future Work

The proposed system is in a continuous improvement stage, with ongoing efforts to perfect the quality of the models and expand the voice catalog. Future work will involve testing new architectures for the existing voices and training new models with additional speakers, including male voices. As mentioned in Section 3, formal evaluations will be conducted on models achieving the best-perceived performance in informal listening tests. We also plan to improve the text processing module by implementing changes in the Cotovía front-end.

Acknowledgements

This research was funded by “The Nós project: Galician in the society and economy of Artificial Intelligence”, resulting from the agreement 2021-CP080 between the Xunta de Galicia and the University of Santiago de Compostela, and thanks to the Investigo program, within the National Recovery, Transformation and Resilience Plan, within the framework of the European Recovery Fund (NextGenerationEU).

References

- Alan W. Black and Nick Campbell. 1995. [Optimising selection of units from speech databases for concatenative synthesis](#). In *Proc. 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, pages 581–584.
- Alan W Black, Heiga Zen, and Keiichi Tokuda. 2007. [Statistical Parametric Speech Synthesis](#). In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1229–IV–1232.
- Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. [YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

- CRPIH and GTM. 2023. [CRPIH_UVigo-GL-Voices: Galician TTS dataset](#).
- Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, Jose Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín, Manuel González González, Senén Barro, and Xose Luís Regueira. 2022. The Nos' Project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the TDLE Workshop LREC2022*, pages 52–61, Marseille. European Language Resources Association (ELRA).
- Thierry Dutoit. 1997. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers.
- Gölge Eren and The Coqui TTS Team. 2021. [Coqui TTS](#).
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. [Generative adversarial nets](#). In *Advances in neural information processing systems*, pages 2672–2680.
- Andrew J. Hunt and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference*, volume 1, pages 373–376. IEEE.
- Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungho Yoon. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA. Curran Associates Inc.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.
- Diederik P. Kingma and Max Welling. 2014. [Auto-Encoding Variational Bayes](#). In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA. Curran Associates Inc.
- Z.H. Ling, X. Zhou, and S King. 2021. The blizzard challenge 2021. In *In Proceedings of the Blizzard Challenge Workshop 2021*.
- Carmen Magariños. 2023. [Nos_TTS-gl-celtia-vits-graphemes](#).
- Antonio Moscoso, Carmen Magariños, and Alberto Bugarín-Diz. 2023. [Nos_TTS-gl-icia-vits-phonemes](#).
- Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. [A review of deep learning based speech synthesis](#). *Applied Sciences*, 9(19).
- Danilo Rezende and Shakir Mohamed. 2015. [Variational inference with normalizing flows](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France. PMLR.
- Eduardo Rodríguez Banga, Carmen García-Mateo, Francisco Méndez-Pazó, Manuel González-González, and Carmen Magariños. 2012. Cotovía: an open source TTS for Galician and Spanish. In *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH*, pages 308–315.
- Yousef Tabet and Mohamed Boughazi. 2011. Speech synthesis techniques. A survey. In *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, pages 67–70. IEEE.
- Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu. 2021. [A Survey on Neural Speech Synthesis](#). *ArXiv*, abs/2106.15561.
- Paul Taylor. 2009. *Text-to-Speech Synthesis*. Cambridge University Press.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Adina Ioana Vladu, Iria de Dios-Flores, Carmen Magariños, John E. Ortega, José Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. Proxecto Nós: Artificial intelligence at the service of the Galician language. In *SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations*, A Coruña, Spain.
- Marta Vázquez Abuín, Noelia García Díaz, Adina Ioana Vladu, Carmen Magariños, Adrián Vidal Miguéns, and Elisa Fernández Rei. 2023. [Nos_Celtia-GL: Galician TTS corpus](#).
- Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. [Review: Statistical Parametric Speech Synthesis](#). *Speech Commun.*, 51(11):1039–1064.
- Alp Öktem, Carmen Magariños, and Antonio Moscoso. 2023. [Nos_TTS-gl-sabela-vits-phonemes](#).

Autopilot: a smart sales assistant

Amanda Oliveira and **João Alvarenga**
amanda.oliveira,joao.alvarenga@blip.ai
Evandro Fonseca and **William Colen**
evandro.fonseca,william.colen@blip.ai
Blip
Brazil

Abstract

In the evolving landscape of digital commerce, the integration of Artificial Intelligence (AI) into widely-used communication platforms like WhatsApp represents a significant advancement. This paper presents the 'Autopilot: a smart sales assistant', an AI-powered chatbot designed to enhance the shopping experience and boost sales through WhatsApp.

1 Introduction

In the dynamic and ever-evolving domain of digital retail, our initiative focuses on integrating Artificial Intelligence (AI) into WhatsApp, a widely used communication platform, to enhance sales dynamics. This paper outlines the development and deployment of the 'Autopilot', an AI-driven chatbot, utilizing Large Language Model (LLM) technology, specifically designed to enrich shopping experiences within WhatsApp.

Echoing the insights of Vinuesa (Mohanty et al., 2023), our approach is indicative of the broader trend in employing AI for customer service management, where AI tools like chatbots and sentiment analysis play a pivotal role in enhancing customer interactions and satisfaction. This paradigm shift towards AI-driven solutions reflects a commitment to delivering more personalized and efficient shopping experiences, hinting at a future rich with predictive analytics and advanced personalization.

The Autopilot initiates its role when a potential customer interacts with a social media advertisement, leading them to a dialogue on the company's WhatsApp platform. This system goes beyond traditional chatbot functionalities, not just by interpreting customer inquiries but also by tailoring product recommendations to user preferences in a conversational manner.

Our work details the intricate integration of an LLM with WhatsApp, highlighting the AI's versatility in understanding and addressing a wide range

of customer needs within a sales context.

2 Autopilot Architecture

In this research, we utilize the most advanced technology available in the market in terms of chatbots. Therefore, we make use of Large Language Models (LLMs) integrated into WhatsApp through the Blip¹ platform. The integration with the LLM is carried out in Python, and the behavior of the Autopilot is determined by the prompt used by the LLM, which was specially developed for this specific sales case.

We meticulously navigate through the sales journey, from the initial interaction to the completion of the transaction, emphasizing the AI's ability to smoothly transition customers to human representatives for finalizing payments, thus blending the precision of AI with a human touch.

The aim of this implementation is to generate a significant increase in customer engagement and sales conversion rates, demonstrating the effectiveness of AI-driven sales strategies. This research not only contributes to the corpus of AI applications in e-commerce but also paves the way for future explorations in enhancing AI-assisted communication platforms for business growth.

3 Demonstration

We will present a practical and dynamic demonstration of the Autopilot, an innovative system that integrates artificial intelligence into the sales process through WhatsApp. The presentation will be anchored by a detailed video that portrays the consumer's journey from the first point of contact to the completion of the purchase.

The video starts with a consumer browsing a social network, where he comes across and interacts with an attractive ad. This initial click is the

¹<https://www.blip.ai/>

starting point for a unique and personalized shopping experience. The user is then redirected to the company's WhatsApp, marking the beginning of their interaction with the Autopilot. This point of the demonstration highlights how our solution captures the customer's attention in a familiar environment and smoothly leads them to an integrated sales channel.

Once on WhatsApp, the video illustrates the conversation between the consumer and the Autopilot. The assistant, equipped with a Large Scale Language Model, analyzes the customer's needs and preferences, offering precise and personalized product recommendations. This part of the demonstration is crucial to show the system's ability to understand and effectively respond to customer inquiries, thereby improving the user experience and increasing the chances of sale.

An innovative aspect of our system is its ability to handle more complex requests, such as a discount. In the video, we will demonstrate how the Autopilot recognizes the need for human intervention and redirects the customer to an attendant, who then offers a special offer and facilitates the payment process. This transition from AI to human service is done smoothly, ensuring that the customer feels constantly supported and valued.

The completion of the purchase is facilitated by a payment link generated by the attendant, allowing the customer to complete the transaction conveniently and securely. The video highlights how the entire process, from the click on the ad to the completion of the purchase, occurs in a single channel, simplifying the customer's journey and increasing the efficiency of the sales process.

4 Conclusion

In summary, this article introduced the innovative concept of the Autopilot, integrating LLM into WhatsApp, as a promising approach in digital commerce. The perspectives presented suggest significant potential to enhance the shopping experience and optimize sales strategies. This pioneering initiative highlights the interaction between AI and customer service, paving the way for future innovations in the sector.

Acknowledgements

The authors acknowledge the financial support of Blip.

References

Aishwarya Mohanty, Jitendra Mohanty, Naveen Lingam, Sagarika Mohanty, and Ajitav Acharya. 2023. Artificial intelligence transforming customer service management: Embracing the future. *The Oriental Studies*, 12:35–47.

