Language Resources and Evaluation (2024) 58:1–6
https://doi.org/10.1007/s10579-024-09729-9

EDITORIAL

Computational approaches to Portuguese: introduction to the special issue

Diana Santos[1] **· Thiago Alexandre Salgueiro Pardo**[2]

Accepted: 15 February 2024 / Published online: 6 March 2024

1 The computational processing of Portuguese

Approximately 250 million people speak the Portuguese language. It is one of the most spoken languages on the planet, being the official language in 10 regions of the world: Brazil, Portugal, Cape Verde, Angola, Guinea-Bissau, Mozambique, São Tomé and Príncipe, Equatorial Guinea, East Timor and Macau. The provision of computational infrastructure and services to support speakers of Portuguese is therefore of obvious importance.

Computational processing of Portuguese began in the 1990s and has increased over the past three decades. Over this time, work on Portuguese has involved several scientific paradigms in the area of Natural Language Processing (NLP), such as symbolic approaches, statistical modeling and, more recently, neural learning, producing valuable theoretical contributions to the field as a whole as well as many resources, tools and applications for Portuguese.

Although an international conference on the computational processing of Portuguese (PROPOR, for PROcessing of PORtuguese) has been held annually since 1993 (among many other relevant publication venues[1]), we believe that at this time,

[1] For the interested reader, other either more local venues or with a less specific focus include STIL (Symposium in Information and Human Language Technology), JDP (*Jornada de Descrição do Português*), ELC (*Encontro de Linguística de Corpus*), SLATE (*Symposium on Languages, Applications and Technologies*), the *LinguaMÁTICA* journal (for the processing of Iberian languages) and AI-related conferences, such as BRACIS (Brazilian Conference on Intelligent Systems), ENIAC (*Encontro Nacional de Inteligência Artificial e Computacional*) and EPIA (Portuguese Conference on Artificial Intelligence), among others.

Diana Santos and Thiago Alexandre Salgueiro Pardo have contributed equally to this work.

✉ Diana Santos
d.s.m.santos@ilos.uio.no

Thiago Alexandre Salgueiro Pardo
taspardo@icmc.usp.br

[1] Linguateca and University of Oslo, Oslo, Norway

[2] Núcleo Interinstitucional de Linguística Computacional (NILC), University of São Paulo, São Paulo, Brazil

it is pertinent to provide an overview of Portuguese language processing, thereby marking a milestone in its development and, importantly, providing increased visibility for work in the area.

This special issue, published on the occasion of PROPOR's 30th year anniversary, is intended as a celebration of NLP applied to the Portuguese language over the past three decades. Over this period, a range of methodologies and activities have been applied to Portuguese, including language modeling, construction and annotation of reference and specialized corpora, methods of evaluating and improving computational linguistic resources, speech processing, metrics for textual complexity measurement, and tasks involving computational creativity. In addition to showcasing for work on the Portuguese language, this issue also aims to be an inspiration for new research initiatives for the computational processing of Portuguese, as well as a potential stimulus for international collaboration to address the challenges of Portuguese language processing.

We took the first steps to create this special issue in mid-2020. Besides an open call for papers, we contacted well known researchers in the area, especially researchers from Portugal and Brazil who are currently the main players in the area of Portuguese language processing. The call for papers generated 24 submissions, and from among these, 11 were selected to appear in this special issue. The papers describe a range of activities, from adaptation of various established models to language-specific solutions along with important long-term initiatives for language technology in Portuguese. A brief overview of the papers is provided below.

2 The contents of this issue

The topics discussed in this issue are diverse, ranging from evaluation of language models and lexical resources, annotation issues and linguistic studies to developing teaching resources and text complexity metrics in order to increase digital inclusion, as well as techniques for speech processing, computational creativity and early detection of mental health problems in social media.

This special issue features three surveys, providing a historical overview of their authors' work in the area. "Human-Inspired Computational Models for European Portuguese: A review" (by António Teixeira and Samuel Silva) reports on approximately 30 years of work in speech technologies for European Portuguese. The authors discuss several specific issues with this variety of Portuguese, such as nasality. They can as well boast one of the two full human-inspired text-to-speech (TTS) systems in the world. This work has advanced, and continues to advance, the state of the art of computational methods not only for European Portuguese, but also other languages. "Automatic Generation of Creative Text in Portuguese: an Overview" (by Hugo Gonçalo Oliveira) reviews creative text generation in Portuguese over the past 16 years and describes an impressive array of different systems that generate creative text. In addition to the historical perspective, where several different lexical resources are described and utilized, the paper also evaluates current deep learning generative models for the Portuguese language and makes a plea for the integration of different approaches in the future. "NILC-Metrix: assessing the complexity

of written and spoken language in Brazilian Portuguese" (by Sydney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Sigle Hartmann and Sandra Aluísio) documents a set of 200 varied metrics and, based on these metrics, a system to assess the readability of texts in Brazilian Portuguese. This is a meta-project with a history of 13 years (the authors mention at least six different separate projects that can be considered to be a part of it) and the paper presents three different applications of the framework and their evaluation, namely: a description of complexity in subtitles for children movies and texts for elementary school; a predictor for textual complexity for original and simplified texts in a particular corpus; and a system that predicts the school grade of a teenager speaker from his/her oral narratives.

Reflecting recent trends in the field, two papers focus on language modeling applied to the task of sentiment analysis for tweets[2] in Portuguese: "A survey and study impact of tweet sentiment analysis via transfer learning in low resource scenarios" (by Manoel Veríssimo dos Santos Neto, Nádia Félix F. da Silva and Anderson da Silva Soares) emphasizes the importance of transfer learning in low-resource scenarios and discusses the associated computations costs. The authors compare the impact of different language models on the task in terms of their potentialities and limitations. "Sentiment analysis in Portuguese tweets: an evaluation of diverse word representation models" (by Daniela Vianna, Fernando Carneiro, Jonnathan Carvalho, Alexandre Plastino and Aline Paes) highlights the challenges of the informal linguistic style of tweets and explores different word embedding representations, demonstrating the relevance of language-specific solutions.

In line with the above initiatives, the paper "Assessing linguistic generalisation in language models: a dataset for Brazilian Portuguese" (by Rodrigo Wilkens, Leonardo Zilio and Aline Villavicencio) argues that, although much recent effort has been devoted to creating large-scale language models, such models lack transparency and interpretability that affects not only their applicability in downstream tasks but also the comparability of different architectures or of the same model trained using different corpora or hyper-parameters. The authors propose a set of intrinsic evaluation tasks that inspect the linguistic information encoded in some language models, including the evaluation of multiword expression and grammatical structuring issues (e.g., use of impersonal verbs, subject agreement, verb agreement, nominal agreement and use of passive and connectors). In a similar vein, the paper "Evaluation of the Brazilian Portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015)" (by Flavio Carvalho, Fabio Paschoal Junior, Eduardo Ogasawara, Lilian Ferrari and Gustavo Guedes) investigates the last version of the Linguistic Inquiry Word Count (LIWC) psycholinguistic dictionary for Brazilian Portuguese and how it correlates with the previous existing version and the corresponding dictionary for English, by examining its performance in text classification tasks.

On the topic of corpora for Portuguese, "Brazilian Portuguese Corpora for Teaching and Translation: the CoMET Project" (by Stella Tagnin) reports on an important and long-standing project for Brazilian Portuguese. The papers first overviews

[2] The works were carried out before Twitter changed its name to "X".

several available Brazilian Portuguese corpora and then focuses on the CoMET Project, which includes three corpora: a comparable Portuguese-English technical corpus (CorTec), a literary Portuguese-English parallel (translation) corpus (CorTrad) and a multilingual learner corpus, (CoMAprend), all available for online queries with specific tools. "SetembroBR: a social media corpus for depression and anxiety disorder prediction" (by Wesley Ramos dos Santos, Rafael Lage de Oliveira and Ivandré Paraboni), describe a new corpus with data from social media in Portuguese to address a more specific and challenging task: the prediction of depression and anxiety disorder. Advancing the state of the art in this kind of task, the paper argues for inclusion of parameters beyond the text itself and present four baseline methods for prediction, showing the task's usefulness.

"Syntactic annotation for Portuguese corpora: standards, parsers, and search interfaces" (by Pablo Faria, Charlotte Galves and Catarina Magro) and "A study on methods for revising dependency treebanks: in search of gold" (by Cláudia Freitas and Elvis de Souza) look at the problem of creating and improving annotated resources using semi-automatic methods. Annotated resources are very important for linguistic studies and supervised machine learning systems, but they are cost-intensive to produce and hard to evaluate, as both papers emphasize and demonstrate. The first paper discusses historical treebanks in the Penn Treebank family in Portuguese and is concerned with presenting a complex and linguistically motivated syntactic annotation scheme, harmonized for all varieties of Portuguese, and suggests a strategy to help linguists annotate more texts with the support of different parsers. The authors argue that it may be preferable to have a rule-based parser with rules understandable by linguists instead of a (possibly more efficient and accurate) machine learning-based parser. The second paper is concerned with the problem of creating gold standards that are both linguistically correct and also friendlier to machine learning systems. The authors work on the oil and gas domain and deal with dependency treebanks. To create gold standard data, they apply a dependency parser and revise its output using several automatic methods. As in the previous paper, the authors describe their development of tools for assisting treebank revision.

3 Concluding remarks

Although Portuguese is spoken in several countries, most of the work presented in this special issue addresses Brazilian Portuguese as opposed to the variety the language prominent in Portugal itself. We can conjecture that is disparity probably stems from the fact that the researcher mass is considerably larger in Brazil[3], which also has more (related) publication venues. However, it is important to note that collaboration efforts do exist between the two countries, some of which are described in papers in this issue. This special issue is in itself an example of cross-country collaborative work.

[3] Another evidence of the researcher mass in NLP in Brazil is the release of a new large comprehensive NLP textbook by the *Brasileiras em PLN* group (Caseli & Nunes, 2023).

The papers in this special issue also make it clear that research in the area of Portuguese language processing has changed significantly over time in line with global scientific movements. Over the years, mainstream NLP research has proceeded from focus on symbolic approaches, then statistical approaches, and most recently, neural approaches; as could be expected, this special issue reflects these changes, including both more "classical" efforts (corpus construction, corpus annotation and lexical resource-based methods and evaluation) and recent trends (language modeling and sentiment analysis). Currently, AI-related approaches and machine learning (deep learning, in particular) have broadened the research field to include not only NLP researchers and corpus linguists but also general AI practitioners. The papers in this special issue reflect this trend, including papers from researchers associated with classical and pioneer research groups and projects for NLP for Portuguese (such as Linguateca (Santos, 2009)[4] and NILC (Nunes, 2010)[5]) together with papers from researchers who represent a new generation in the field.

It is highly challenging to speculate as to what the next 30 years of NLP research for Portuguese holds, especially considering how fast the fields of NLP and AI are currently evolving. However, we can hope that NLP will help to unveil the intricacies of the Portuguese language in its several varieties and, consequently, produce more NLP products (resources, tools and applications) that accommodate their similarities and differences. Portuguese language processing would also benefit from the creation and strengthening research groups focused on the topic in order to spur a greater presence in the Portuguese language processing community. More broadly, we can hope that the PROPOR conference and similarly specialized forums for presentation of work on Portuguese language processing will become larger and foster continuing progress in the field.

Acknowledgements We deeply thank all authors who sent papers to the special issue and all the reviewers who helped us select the best ones and improve them. We are grateful to Nancy Ide and Nicoletta Calzolari, the editors-in-chief, for their outstanding and constant guidance during the issue preparation. We are also indebted to Sara Goggi for her steady help in all administrative matters.

Funding Neither author received funding to write this introduction.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Caseli, H. M., & Nunes, M. G. V. (eds.). (2023). Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. BPLN. https://brasileiraspln.com/livro-pln

Nunes, M. D. G. V., Aluísio, S. M., & Pardo, T. A. S. (2010). Um panorama do núcleo interinstitucional de linguística computacional às vésperas de sua maioridade. *Linguamática, 2*(2), 13–27.

[4] https://www.linguateca.pt/

[5] https://www.nilc.icmc.usp.br

Santos, D. (2009). Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática, 1*(1), 25–59.