Tetun Universal Dependencies Treebank

Gabriel de Jesus

INESC TEC / Faculty of Engineering of the University of Porto (FEUP)
Porto, Portugal

1 Summary

This document details the objectives for applying to the 1st UNIDIVE Summer School, scheduled to take place in Chișinău, Moldova, from July 8–12, 2024. It describes the focus language of this application, the dataset we have released, the benefits of participating in this training, and the potential future applications of the skills and knowledge acquired.

Keywords: Tetun, Low-resource language, Corpus annotation.

2 Introduction

My name is Gabriel de Jesus, a PhD candidate enrolled in the Doctoral Program in Informatics Engineering (PRODEI) at the Faculty of Engineering, University of Porto (FEUP), Porto, Portugal. I am affiliated with both the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) and FEUP. My submission for the 1st UNIDIVE Summer School is centered on the development of the Tetun Universal Dependencies Treebank. This initiative aims to contribute for the advancement of computational tools and technologies for Tetun, an official language of Timor-Leste, aligning with my broader research interests.

Tetun, alternatively written as Tetum or Tétum, is an Austronesian language spoken in Timor-Leste, a Southeast Asian island country. It was a dialect used as both church and trade language during the colonial era until Timor-Leste restored its independence and became a new sovereign state on May 20, 2002. In 2002, the Government of Timor-Leste designated Tetun as one of the country's official languages alongside Portuguese [5], leading to its widespread usage in public life. Tetun is a low-resource language spoken by 78.78% of a 1.18 million populations [1].

The Government of Timor-Leste has established a standard orthography for Tetun through the *Instituto Nacional de Linguística*, referred to as Tetun INL. This Tetun standardization has adopted as the official Tetun being used in the education system, official publications, and media [3]. Tetun INL is based on the Latin alphabet, distributed in 5 vowels: *a, e, i, o, u*, and 21 consonants: *b, d, f, g, h, j, k, l, ll, m, n, ñ, p, r, rr, s, t, u, v, x, z* [4]. The letters *C, Q, W*, and *Y* are not used in Tetun INL, except for proper names and international symbols. The accented vowels *á, é, í, ó, ú*, are also used, and the apostrophe (') denotes a glottal stop. Additionally, the hyphen is also introduced to indicate mono-semantic compound words. Despite its official status, Tetun lacks digital content and annotated data.

I am the first Timorese researcher who take initiative to develop Tetun, with a focus on Text Information Retrieval [1]. My research endeavor is to create effective search solutions tailored for Tetun. A significant achievement of my work so far includes the development of a Tetun tokenizer[1], along with the release of a Tetun text dataset comprising 33.6k documents [2].

3 Advantages of Participation

The Tetun Universal Dependencies Treebank project represents an initiative for advancement in the computational processing of Tetun, which is crucial for developing more sophisticated language technologies for Tetun speakers. By building a treebank for Tetun, researchers and technologists can create tools that enhance machine translation, text analysis, and natural language processing applications. This is particularly pivotal for underrepresented languages like Tetun, as it helps in preserving the language, improving accessibility, and ensuring that technological advancements reach a wider segment of Timor-Leste's population.

Participation in the UNIDIVE Summer School would allow a project like the Tetun Universal Dependencies Treebank to gain international exposure and expert critique, facilitating further development and refinement. It would also provide an opportunity to network with other researchers and technologists who are working on similar projects. This collaborative environment could lead to new ideas and methodologies that could be applied to the Tetun treebank, potentially accelerating its development and increasing its accuracy and applicability. Additionally, such a project aligns well with the interdisciplinary approach of UNIDIVE, showcasing how linguistic research directly contributes to technological innovation and cultural preservation.

[1] `https://pypi.org/project/tetun-tokenizer/`

4 Future Applications

The knowledge and experience gained from the UNIDIVE Summer School will be instrumental in expanding the scope of the project for Tetun in this application. The training will provide insights into cutting-edge methods in computational linguistics, which are essential for advancing the development of language processing tools and technologies for Tetun. Exposure to a variety of interdisciplinary approaches and technologies, such as advanced annotation tools and techniques, will enable me to effectively manage in the Tetun dataset annotation processes for constructing a Tetun universal dependencies treebank. This will directly contribute to the advancement of language technologies for Tetun.

Furthermore, the collaborative environment and networking opportunities at the summer school will open doors to potential partnerships and collaborations with other researchers and institutions. These connections can lead to joint projects, further research funding, and shared knowledge that will benefit the ongoing development of computational resources for Tetun. By applying the advanced skills and innovative methodologies learned at UNIDIVE, I believe that I can contribute to the broader field of computational linguistics for low-resourced languages. This will ultimately facilitate better technological support for Tetun speakers and help preserve and promote the language on a global platform.

References

[1] Gabriel de Jesus. Text Information Retrieval in Tetun. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *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, 2023. doi: 10.1007/978-3-031-28241-6_48. URL https://doi.org/10.1007/978-3-031-28241-6_48.

[2] Gabriel de Jesus and Sérgio Nunes. Labadain-30k+: A monolingual tetun document-level audited dataset [data set]. INESC TEC. https://doi.org/10.25747/YDWR-N696, 2024.

[3] Democratic Republic of Timor-Leste DL 01/2004, Government Decree-Law No. 1/2004 of 14 April. The standard orthography of the tetun language. http://mj.gov.tl/jornal/lawsTL/RDTL-Law/RDTL-Gov-Decrees/Gov-Decree-2004-01.pdf, last accessed on February 21, 2024., 2004.

[4] National Institute of Linguistics INL. The standard orthography of the tetum language: 115 years in the making. https://archive.org/details/the-standard-orthography-of-the-tetum-language/mode/2up, last accessed on February 21, 2024., 2004.

[5] Pedro Carlos Bacelar de Vasconcelos, Andreia Sofia Pinto Oliveira, Ricardo Sousa da Cunha, Andreia Rute da Silva Baptista, Alexandre Corte-Real de Araújo, Benedita McCrorie Graça Moura, Bernardo Almeida, Cláudio Ximenes, Fernando Conde Monteiro, Henrique Curado, et al. Constituição anotada da república democrática de timor-leste, 2011. URL http://hdl.handle.net/10400.22/4008.