

# Malayalam Parser for Dataset Creation

The “Malayalam Parser for Dataset Creation” project aims to address the scarcity of annotated datasets in the Malayalam language for Natural Language Processing (NLP) applications. The primary objective is to develop a robust Malayalam parser capable of analyzing the syntactic and semantic structures of Malayalam sentences. The creation of this parser involves several key steps, including data collection from diverse sources, preprocessing to ensure data quality, and manual annotation of a representative subset of the data with grammatical and syntactic information. The parser development process encompasses the selection of an appropriate parsing approach, whether rule-based, statistical, or machine learning-based. The model is trained using the annotated Malayalam dataset, focusing on capturing the unique linguistic nuances of the Malayalam language. Evaluation metrics are employed to assess the parser’s performance on a separate test set, guiding iterative refinement and enhancement. The resulting Malayalam parser serves as a valuable tool for the analysis of grammatical structures in new Malayalam text data. Its application contributes to the creation of high-quality Malayalam datasets, crucial for advancing NLP research and applications in the Malayalam language. This project encourages collaboration with linguists, researchers, and the Malayalam-speaking community to ensure linguistic accuracy and relevance in the development of the parser. The “Malayalam Parser for Dataset Creation” project aligns with the broader goal of promoting linguistic diversity in NLP, addressing the challenges posed by the scarcity of resources for underrepresented languages. Through the development of this parser, the project aims to facilitate further research and innovation in Malayalam NLP, opening avenues for the exploration of various language-related tasks and applications.

## **Team Members:**

1. Gautham C Sudheer
2. Fathima Jennath
3. Godwin Gino
4. Mohammed Basil