The project will focus on predicting morphological information automatically from a database of historical texts in Mapudungun. Specifically, the goal is to employ deep learning to perform lemmatization, the task of predicting a word base form based on its surface form. Our baseline model will be an adaptation of an existing recurrent encoder-decoder model with attention, LEMATUS (\cite{B18}). The aim is to estimate how accurate the baseline would be on predicting automatic labels for the non-annotated data. The secondary goal is to explore how the baseline model performance is affected by pre-processing the data set (e.g. normalizing spelling variations across texts) and to what extent it can be improved. Third, I will employ a different implementation of the attention mechanism which was showed to achieve state of the art performance on low resource language settings, namely hard attention (\cite{AharoniG16}).

The model performance will be evaluated on a portion of the labelled data (test-set), to obtain the model estimation accuracy.