TweetTaglish: A Dataset for Investigating Tagalog-English Code-Switching

Megan Herrera, Ankit Aich, Natalie Parde
University of Illinois Chicago
{mherre42, aaich2, parde}@uic.edu

Introduction

- Tagalog is an Austronesian language and former official language of the Philippines spoken by over 23 million people in the world.
  - Tagalog is a low-resource language in NLP and traditional NLP techniques perform poorly on multilingual text
- Code-switching (CS) is a linguistic phenomenon where speakers mix multiple languages in a single utterance
- This study contributes a large, first-of-its-kind dataset containing annotated instances of Taglish code-switching and proof of validity using traditional machine learning models.

Dataset Collection

- Dictionary-based language ID algorithm using English and Tagalog dictionaries
  - If word is in neither or both dictionaries, it is labelled as “Other”
- Search terms identified by Flores (2020) Tagalog CS study

Search Term	Linguistic Purpose
magko-	Marks present and future tense
di ba	“I mean”
talaga	“really”
ano yung	“what is”
para sa	“for”
parang	“for”

- English, Tagalog, Other split calculated
- Comparison of author annotations with our language ID algorithm (κ = 0.7)

Not yet so may balak talaga lagyan haha

English=0.375, Tagalog=0.375, Other=0.25

Anuena grammys apakatagal refresh ako nang refresh di ba pwedeng isahang announce nalang

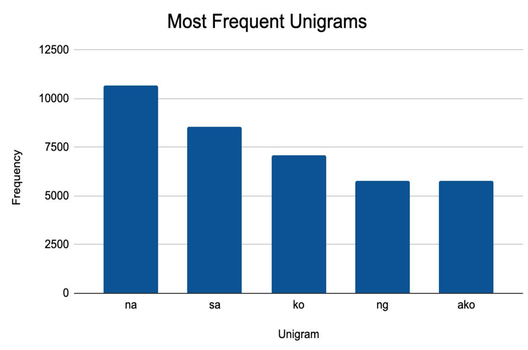
English=0.231, Tagalog=0.538, Other=0.077

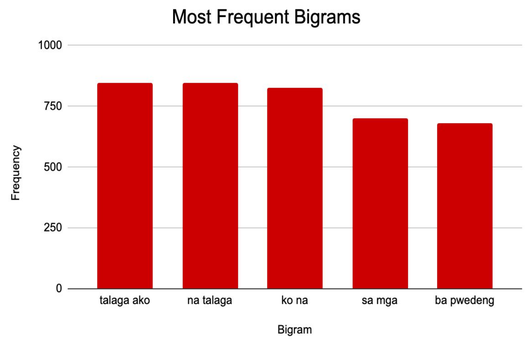
2 concerts na po ang namiss ko sana di ko na mamiss this time yung concert nila

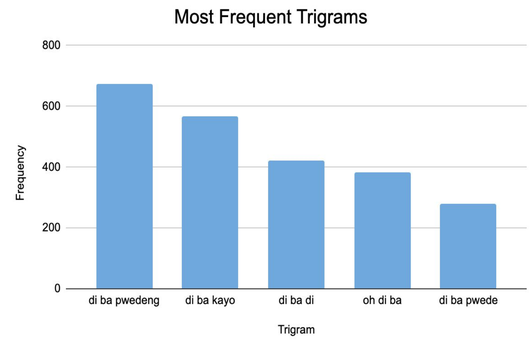
English=0.235, Tagalog=0.769, Other=0.176

Analysis & Proof of Concept

- Top unigrams, bigrams and trigrams computationally confirm that the chosen search terms from Flores (2020) are effective for gathering Tagalog CS data
- Frequent use of Tagalog personal pronouns suggest that CS is more casual and personal







Experiments & Results

- 80/20 train test split
- Performance measures
  - R^2
  - RMSE (Root Mean Squared Error)
- Higher R^2, lower RMSE scores indicate strong performance
- MLP (multilayer perceptron) performed the highest on the data
  - Best at classifying English words

	English		Tagalog		Other	
Model	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Mean	-0.000	0.190	-6.320	0.200	-0.000	0.151
Linear	0.861	0.071	0.700	0.109	0.375	0.119
SVR	0.898	0.061	0.854	0.076	0.699	0.083
SGD	0.848	0.074	0.688	0.112	0.333	0.123
Ridge	0.861	0.071	0.700	0.109	0.374	0.119
MLP	**0.909**	**0.057**	**0.883**	**0.068**	**0.797**	**0.068**

Acknowledgements

This work was supported in part by a startup grant from the University of Illinois at Chicago. We thank Dr. Jill Hallett from the UIC Linguistics Department and the anonymous reviewers for their helpful comments.