TU Delft
CSE3000
Research Project

It sounds like Greek to me

On the classifiability of languages using their phonemes

Author: Johannes IJpma

Supervisors: Stavros Makrodimitris, Arman Naseri Jahfari, Tom Viering

Responsible professor: Marco Loog

Problem description

Classifiers can learn to which language a word belongs, but this has not yet been tried for their phonetic transcriptions

This research:

1) Compares performance of both representations on various machine learning models

2) Investigates which language features are most relevant when classifying pronunciations

Background

IPA

International Phonetic Alphabet (IPA) is an alphabetic system for phonetic notation.

Ough	
Enough	ɪˈnʌf
Plough	plaʊ
Although	ɔːlˈðəʊ

Ipa-dict dataset

Provides dictionaries of words and their phonemic pronunciation for 24 languages

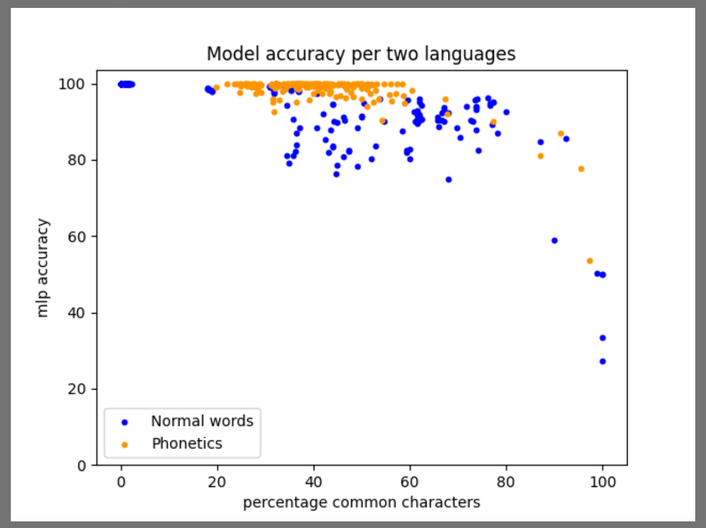
Word	IPA	Language
آباد	/aːbaːd/	Arabic
آباض	/aːbaːdˤ/	Arabic

Results

Do machine learning models perform better on regular words or the phonetic transcription?

Model Accuracy	Logistic Regression	MLP
Regular text	0.78	0.8
IPA	0.96	0.96

Does the amount of common characters in the alphabet impact the performance of the machine learning models?



How relevant is the order of the characters in classifying the language?

Model Accuracy	MLP (Character occurrence matrix)
Regular text	0.70
IPA	0.95

Conclusions

The phonetic representation outperforms the normal text. With the few language pairs where this is not the case this can be explained by that fact that, in almost all cases these languages have no common characters in their alphabets.

There does not seem to be a direct relation between the characters common to two alphabets. The accuracy varies between 75-100 percent for the normal text and between 85 and 100 for the IPA transcriptions.

The order in which the characters occur does not seems to be the important to the classifier. The accuracy loss is only 5% for the regular words and 1% for the phonetic data

Discussion

The MLP trained on the amount of each character (so no information as to in which order these characters occur) performs unreasonably well, with only a 1-5% loss compared to the one-hot encoded data.