Comparative reconstruction of Middle Chinese using deep learning

Andrew Yang[1], Ken Hong[2], Nash Luxsuwong[3]

[1]Department of Linguistics, [2]Symbolic Systems Program, [2]Department of Economics, Stanford University

Project Mentor: Jon Kotker

Task definition

- Comparative reconstruction establishes features of an ancestor language based on related languages.
- Middle Chinese (MC) refers to a language believed to be spoken around 600 CE near current-day Xi'an.
- Mandarin, Cantonese, and many other languages are seen as divergent developments from MC.
- Korean, Japanese, and Vietnamese vocabularies have been heavily influenced by MC pronunciation.
- Linguists have manually used these related languages to reconstruct the pronunciation of MC.
- We use ML methods to reproduce reconstructions comparable to results of prominent linguists.

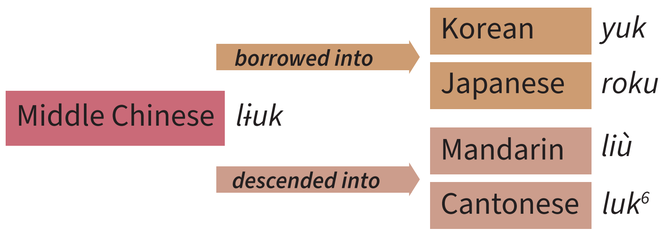


Figure 1. Similarity between the pronunciations of 六 (six) in MC and related languages

Data

- Dataset scraped from *Wiktionary* containing entries for pronunciations of 15250 distinct characters in Mandarin, Cantonese, Korean, Japanese.
- Syllables separated into syllabic units *onset, nucleus, coda,* and *tone*.

Figure 2. Demonstration of how a syllable can be analyzed as onset, nucleus, coda, tone.

Models

- For each of the *onset*, *nucleus*, *coda*, and *tone* categories, we trained a one-vs-rest logistic regression model as well as a deep neural network.
- Each model was evaluated on its ability to reproduce the MC reconstruction by Karlgren (1922).

ZERO RULE BASELINE

- A separate model using Zero Rule prediction was constructed to serve as a baseline for analysis. This model returns the most frequent class every time.

LOGISTIC REGRESSION

- For each syllabic unit within each category, we trained a logistic regression classifier and weighed it against every other classifier.
- Performed significantly better than baseline model.

DEEP NEURAL NETWORK

- For each category, we trained a neural net with two 32-neuron hidden layers.
- Outperformed logistic regression by ~5% on average.

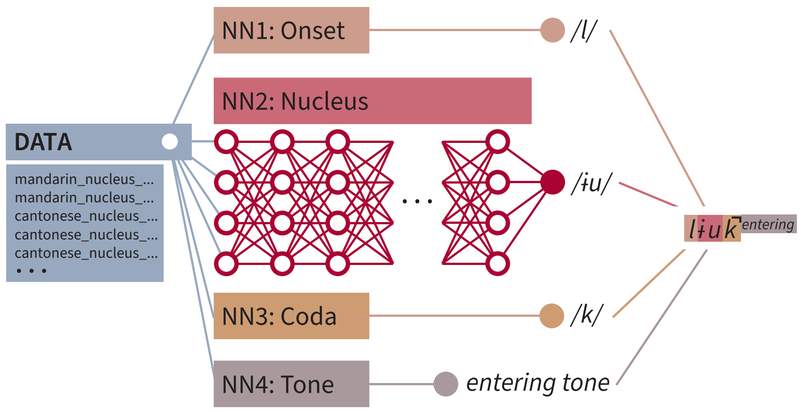
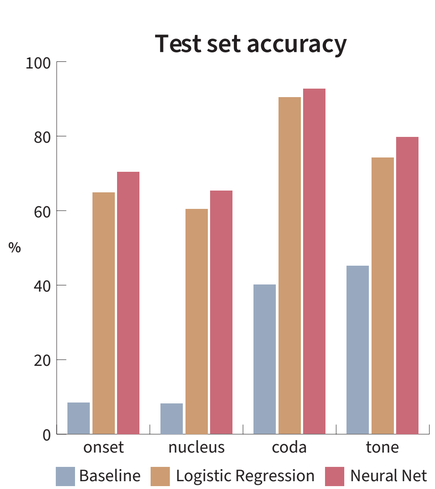
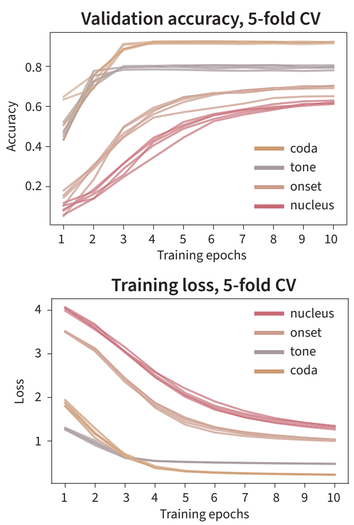


Figure 3. Each neural network is responsible for predicting one part of the final syllable.

Preliminary Results





Left: Figure 4. Test set accuracy across our three models.
Right: Figure 5. Validation accuracies and training losses with 5-fold CV for the Neural Network.

Analysis and Discussion

- Results for *tone* and *coda* are significantly better than *onset* and *nucleus* across the board. This is expected, as there are fewer classes in these categories.
- The top performer was the neural network, achieving the best results across all four categories.
- NN attained 65% accuracy for *nucleus* — a significant result considering there are 62 possible predictions.
- Results limited because most characters lacking pronunciation data for one or more languages.
- We hope to add data from more related languages and to refine and retrain our models after adding phonetic series data for each Chinese character.
- Overall, machine learning seems to be promising for further applications in phonological reconstruction.