CS 221 Project Proposal:

Toward a new frontier in historical comparative linguistic reconstruction

Andrew Yang ycm, Ken Hong kenhong, Nash Luxsuwong nashlux

Background

Comparative linguistic reconstruction is a process by which qualities of an ancestor language is reconstructed from the qualities of existing languages. *Phonological* reconstruction is linguistic reconstruction that focuses on the *phonology*, or sound system, of a language. A major part of *comparative*, *phonological reconstruction*, then, is using data from existing languages to recreate the sounds that would have been a part of an ancestral language.

Comparative phonological reconstruction is usually done by comparing sound correspondences between related words—*cognates*—in related languages, and proposing *phonological rules* to predict how the ancestral language would have sounded.

Proto-Indo-European, for example, is the hypothetical ancestral language to many languages in Europe and South Asia; languages such as English, French, and Hindi are all thought to have descended from Proto-Indo-European. In English, French, and Hindi, the word for *two* starts with a dental or alveolar consonant; via comparative reconstruction, we would hypothesize that the Proto-Indo-European word for *two* also begins with a dental or alveolar consonant. Indeed, Sihler 1995 reconstructs Proto-Indo-European *two* as /d(u)wóh/.

Scope of project and motivation

We will create a machine learning model that reconstructs syllables in Middle Chinese from the phonology of existing Sinitic and Sino-Xenic languages such as Mandarin, Hakka, and Korean. Whereas the traditional approach involves long hours of studying pronunciation patterns in descendant languages, we hope that a machine learning approach will be able to accurately identify sound correspondences much more quickly. We focus on Middle Chinese (in particular, the Karlgren-Li reconstruction) because current reconstructions from established linguists do not vary greatly, whereas there tends to be large disagreements in reconstructions of Old Chinese. Thus, we will be able to use widely accepted Middle Chinese reconstructions as our "ground truth." If successful, our results could have profound implications for the field of historical linguistics, and our methodology can be applied to previously unconstructed languages.

Related work

Work in this area is relatively sparse; two selected works that are tangentially related are "Automated reconstruction of ancient languages using probabilistic models of sound change" (Bouchard-Côté et al., 2013) which attempts to reconstruct Austronesian languages using a probabilistic model of sound change and "Reconstructing language ancestry by performing word prediction with neural networks" (Dekker, 2018) which attempts to retrieve regularities in sound change patterns with neural networks. To our knowledge, these related works differ in their approaches and do not use Sinitic and Sino-Xenic languages as a part of their metrics.

Data

The pronunciations for words in various Sinitic and Sino-Xenic languages will be crawled from Wiktionary. We expect to crawl pronunciations for at least 5,000 characters that appear throughout history (i.e. characters that existed when Middle Chinese was spoken). We will preprocess the pronunciations so that they may be separated into their constituent sounds.

Inputs and outputs

The raw data will be processed and standardized into a format that is compatible with the model that we decide to use. An input would consist of several cognates of the target reconstruction from the related languages, from which we would expect an output similar to an actual reconstructed pronunciation by a historical linguist. For each pronunciation, we separate it into beginning consonant, vowel, and ending consonant, which we will refer to as *onset, nucleus,* and *coda* respectively.

An example is provided below: (Data obtained from https://en.wiktionary.org/wiki/讓)

讓 "yield"	Language	Pronunciation	Onset	Nucleus	Coda
	Mandarin	/rang/	/r/	/a/	/ng/
	Cantonese	/joeng/	/j/	/oe/	/ng/
	Hakka	/ngiong/	/ng/	/io/	/ng/
	Japanese	/jou/	/j/	/ou/	/∅/
	Korean	/yang/	/y/	/a/	/ng/

The format for a sparse input vector for 讓 might look like:

```
Φ("讓")  =  {
                "mandarin_onset_r":   1,
                "mandarin_nucleus_a": 1,
                ...
                "korean_coda_ng":     1
            }
```

From this data, we would try to predict the Middle Chinese reconstruction of the character pronunciation, which should yield the output: /ŋʱɐŋ/. (The symbols in /ŋʱɐŋ/ are sounds in the International Phonetic Alphabet.)

Baseline

First, we will experiment with multiple logistic regression; given a list of features, we would like to have a model that returns the most likely Middle Chinese onset, nucleus, and coda. We expect this baseline to achieve average performance, where basic syllables like 漢 /han/ that do not vary much cross-linguistically are predicted correctly, while characters like 日 (/ri/, /jit/, /jat/, /il/ etc.) that vary wildly will be classified incorrectly.

Oracle

In historical linguistics, reconstructions are created by formulating a set of phonological rules. For instance, a phonological rule might be initial /h/ deriving /k/ from Middle Chinese to Japanese. Real reconstructions are based from long chains or phonological rules (look up *feeding order* and *bleeding order*). Thus, our oracle would be a rule-based paradigm that is able to produce exactly the reconstructions made by a historical linguist. We don't know for certain what Middle Chinese sounded like, so we will treat reputable reconstructions (such as the Karlgren-Li reconstruction) as the "ground truth."

Ideas for models:

We think that a neural network will do better in this task, because we believe there are complex vowel-consonant interactions that a logistic regression model may not be able to exploit. We will also test with other models like random forests.

Evaluation metrics:

We are currently considering various different evaluation metrics. The first is raw accuracy (# correctly reconstructed / total), where an example counts as correct if the entire reconstruction matches the "ground truth" (the known Karlgren-Li reconstruction, for instance). Another metric is using partial correctness divided by number of examples, where partial correctness depends on whether any of the onset, nucleus, and coda were reconstructed correctly.