Verified Computational Rule-based Historical Phonology in Standard ML and Isabelle/HOL

This paper introduces an implementation of rule-based phonology in Standard ML and a formal definition and verification of the core components of such phonology in Isabelle/HOL, an interactive theorem prover. This phonology is used to automatically derive modern reflexes of Spanish, Portuguese, Chinese and Sino-Korean from their ancestral etyma using one underlying model. The architecture of this program is as such: We first implement a featureful segmental inventory, which means that each segment is not merely a character literal but data types with feature information. Then we define the constituents of syllables and syllables themselves by gluing the segments into nested lists. Once syllables are defined, phonological words come naturally, as they can be implemented as lists of syllables. Finishing the definition of phonological words means that we can represent all of the etyma and reflexes in the languages that we are investigating in this project. The rest of the program deals with the definition of the operations on those data, namely sound changes, and convenient utilities to help us define all of the sound changes happened in the history of these four languages.

A sound change in our system is represented as function mapping a phonological word to another, which matches the intuition of a working linguist. One may stop here and start implementing all of the sound changes as recursive ML functions on lists, as the phonological words are so represented in our system, but this approach is tedious, certainly may result in a lot of boilerplate, for that many sound changes in world languages only differ in some details and are structurally similar; it is also error-prone, as the reduplication of similar routines in the code base usually is. Thus instead of writing those sound changes entirely by hand, most of the sound changes in our system are created through schema that decouples the problem into manageable modular pieces. Here is the concrete explanation: just like we have roughly three tiers in the representation of a phonological word: the segments (which themselves are products of features), syllables (that have constituents like onset, nucleus, and coda), and phonological words. Our strategy is to define utilities that would rewrite one tier at a time and compose them into an actual sound change.

These two components, the component that represents data and the component that rewrites data, constitutes the trusted kernel of the program; this kernel is what we are going to verify in Isabelle/HOL. Isabelle is a member of HOL family of theorem provers. It is based on Higher-Order Logic, which a battle-tested logical system that is more than enough to verify our system to secure the desired behaviors. This verified kernel is shared among all the languages whose history we implemented; the only two language dependent parts of our system are the etyma that are to be rewritten and the respective sound changes in those languages. It should be noted that although there are four languages in our project, we only need two sets of etyma: Latin for Spanish and Portuguese, Chinese for both Chinese and Sino-Korean. Even more, Spanish and Portuguese share all the sound changes in our system until their split in Medieval times. Spanish and Portuguese are chosen precisely because of their similarities, so that we can demonstrate another feature of our system: our program is able to output and parse reflexes coupled with their history represented a list of sound changes, which enabled us to: define sound changes that are shared among Spanish and Portuguese, apply them to the etyma, and store the results so that the now separate Spanish and Portuguese modules can deal with the Romance etyma respectively.

The entirety of this program is written in Standard ML '97 and is able to be compiled both by the SML/NJ and MLton compilers. The proof scripts are written in Isabelle/HOL 2022. The source code of this project will be released under the BSD-3-Clause license.

[Wil38] Edwin B. Williams. *From Latin to Portuguese: Historical Phonology and Morphology of the Portuguese Language*. University of Pennsylvania Press, 1938.

[Llo87] Paul M. Lloyd. *From Latin to Spanish*. American Philosophical Society, 1987.

[Pau90] Lawrence C. Paulson. "Isabelle: The Next 700 Theorem Provers". In: *Logic and Computer Science* (1990), pp. 361–386.

[Pau96] Lawrence C. Paulson. *ML for the Working Programmer*. Cambridge University Press, 1996.

[Pen02] Ralph Penny. *A History of the Spanish Language*. Cambridge University Press, 2002.

[Har03] Steven Lee Hartman. *Phono (Version 4.0): Software for Modeling Regular Historical Sound Change*. 2003. URL: https://langnhist.weebly.com/files/theme/ver40.pdf (visited on 10/18/2022).

[Hua05] José Ignacio Hualde. *The Sounds of Spanish*. Cambridge University Press, 2005.

[Sch09] Axel Schuessler. *Minimal Old Chinese and Later Han Chinese: A Companion to Grammata Serica Recensa*. University of Hawai'i Press, 2009.

[BS14] William H. Baxter and Laurent Sagart. *Old Chinese: A New Reconstruction*. Oxford University Press, 2014.