Ab Antiquo: Proto-word Reconstruction with RNNs

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg

carlomeloni@mail.tau.ac.il shauli.ravfogel@gmail.com yoav.goldberg@gmail.com

1. Motivation & Task

Can neural sequence models learn the regularities that govern historic sound change in human languages?

x =lapte$^{\text{RM}}$, lait$^{\text{FR}}$, latte$^{\text{IT}}$, leche$^{\text{SP}}$, leite$^{\text{PT}}$

↓ Reconstruct (orthographic)

y =lactem

x =lapte$^{\text{RM}}$, lɛ$^{\text{FR}}$, latte$^{\text{IT}}$, letʃe$^{\text{SP}}$, lɐjti$^{\text{PT}}$

↓ Reconstruct (phonetic)

y =laktɛm

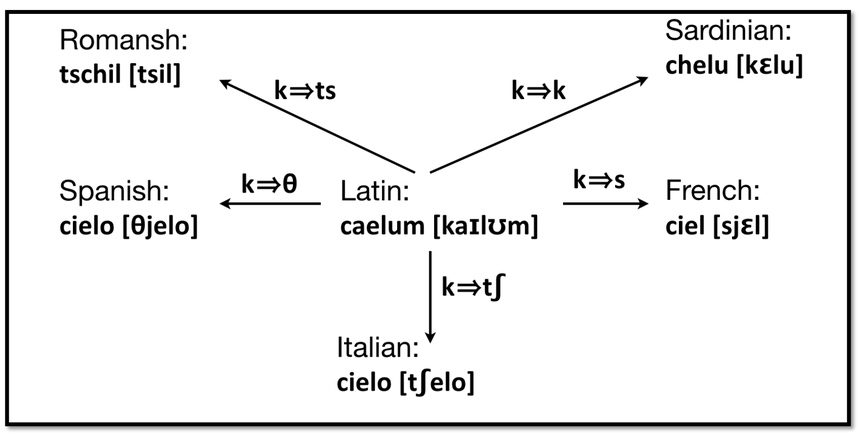
- Previous works: word reconstruction on different languages, using probabilistic graphical models.
- We train RNNs on phonetic and orthographic reconstruction in Romance languages.

2. Contributions

- A novel dataset: over 8,000 human-annotated entries in 6 Romance languages, derived from Wiktionary.
- Extensive error analysis links the opacity of the historic change and the performance of the model.
- A synthetic evaluation set is used to assess the learnability of documented rules of sound change.
- Analysis of learned representation reveals the learning of phonologically meaningful representations without direct supervision.

3. Background: Historical Linguistics

- Historical linguists identify and explain historic linguistic change.
- A family of languages can often be traced into a common, ancestral language – a proto-language.
- Languages in the same family show regularities of phonetic change:



- By back-tracing those rules one can reconstruct proto-words

4. Model and Experimental Setup

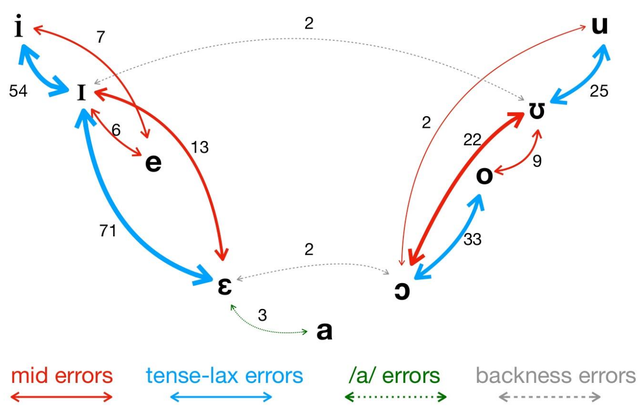
- Standard encoder-decoder architecture: character-level LSTM+attention
- Phoneme and language embeddings enable transfer across languages.
- Evaluation metric: edit distance

4. Main Results & Analysis

- Average edit distance: 0.65 on the orthographic task VS 1.022 on the phonetic task. The phonetic task is significantly harder.
- Several recurrent error types were detected, showing the errors are related to the opacity of the phonological change:

Error type	Orthographic	Phonetic
High-mid	18%	8%
Deletion	14%	6%
Consonant	13%	15%
Cluster	12%	3%
Morphology	11%	10%
Vowel	7%	8%
Length	—	26%
Orthography	5%	—
Other	20%	24%

- The analysis of vowel's errors demonstrates that they are grounded in substantial phonological factors, such as tense-lax distinction:



5. Evaluating rules learning

- To what extent does the model internalize rules of phonetic change?
- A synthetic rules-evaluation dataset was manually constructed, containing 33 instances, each expressing a specific rule of sound change as documented by linguists:

Rule: change of Latin [j] at word initial

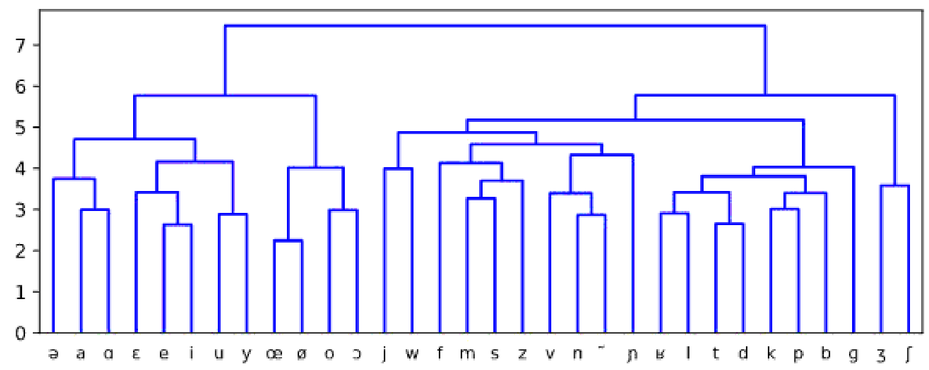
x =ʒa$^{\text{RM}}$, ʒa$^{\text{FR}}$, dʒa$^{\text{IT}}$, xa$^{\text{SP}}$, ʒa$^{\text{PT}}$

y =ja

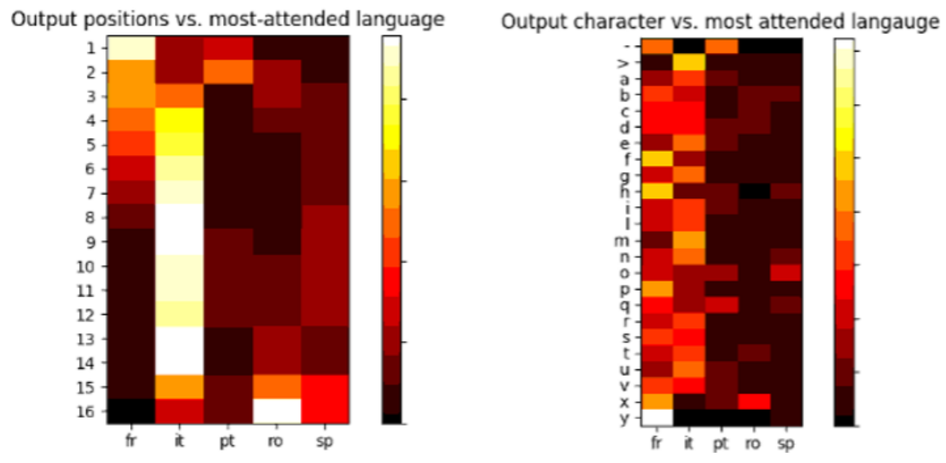
- We find that 66% of the rules were correctly identified by the model.
- Rules learnability is influenced by deterministic mapping between Latin and its daughter languages, as expressed by the rule.

6. Learned representations & Attention

- Hierarchical clustering of phoneme representations demonstrates implicit learning of phonologically meaningful hierarchy:



- This probably reflects the fact that different classes of phonemes undergo different sound change processes.
- Attention analysis: we inspect the most attended language for each position and output character:



- The model almost entirely focuses on French and Italian