

# PILA: A Historical-Linguistic Dataset of Proto-Italic and Latin

**Authors: Stephen Bothwell, Brian DuSell,  
David Chiang, and Brian Krostenko**



**ETH** zürich

**LREC-COLING 2024**

# Agenda

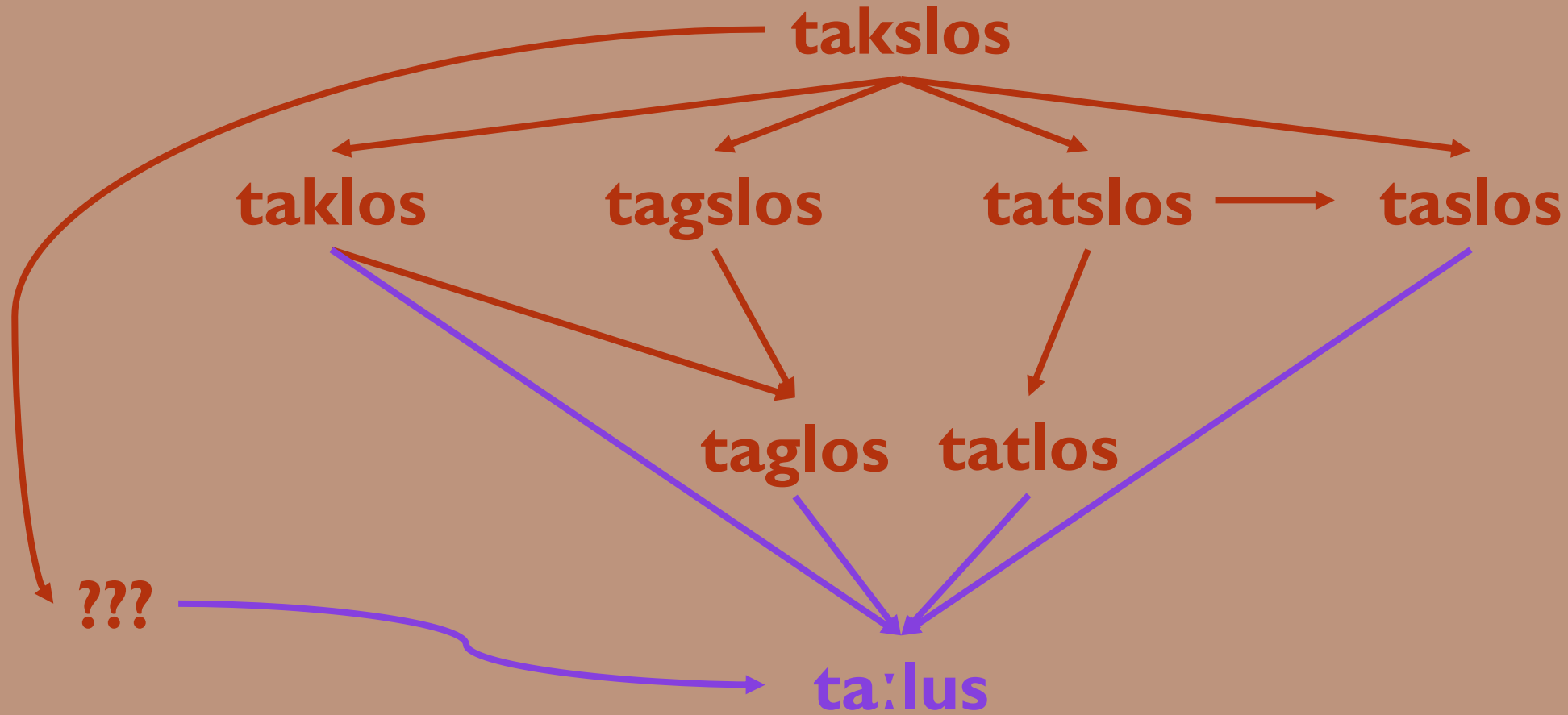
- Introduction: On the Data of Historical Linguistics
- Proto-Italic and Latin: the Context
- Proto-Italic and Latin: the Dataset
- Applications: Phonetic Transduction
- Applications: Dataset Compatibility

# On the Data of Historical Linguistics

# The Nature of Historical Linguistics

- Computational historical linguistics studies language change. In phonetics, one subject of study is *cognate sets*.
  - Cognate sets relate *etyma* (sg. *etymon*) and *reflexes*—ancestor and descendant forms, respectively.
  - Historical linguists determine phonetic relationships and rules by comparing forms related by geographic and temporal circumstances. They even postulate proto-languages on this basis.
- Linguists come to different conclusions about reconstructions, and some patterns postulated by historical linguistics do not have complete explanations.

# The Search Space of Sound Change

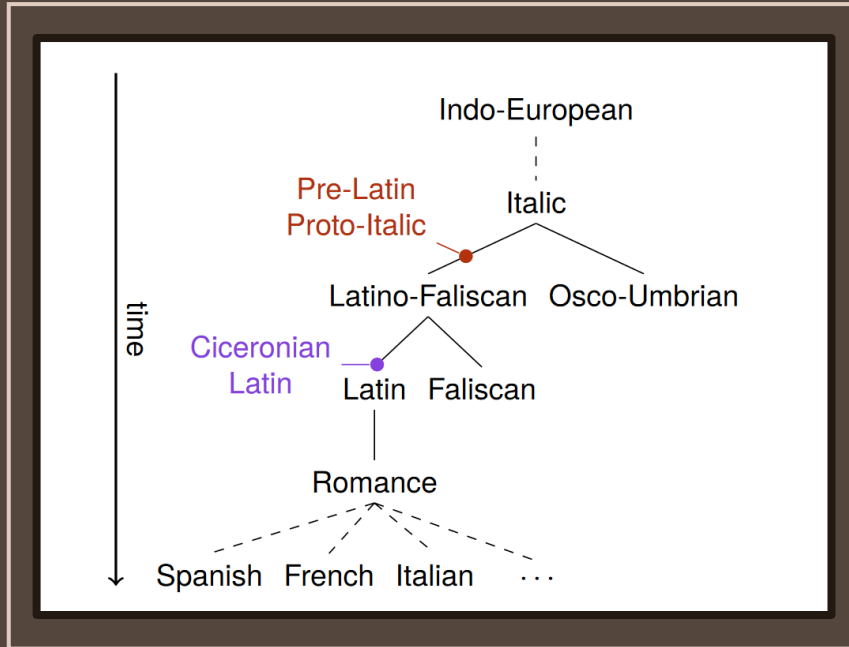


**Figure 1:** Tree of potential sound change paths for *takslos* > *talus*.

# Proto-Italic and Latin: the Context

# The Language Family

- Latin and Proto-Italic are descendants of Proto-Indo-European (PIE)—a reconstructed ancestor for a variety of languages.
- PIE and many of its descendants (*i.e.*, Proto-Italic and Latin) are *highly inflected*: given a stem, they produce a variety of distinct forms fulfilling different grammatical categories.



**Figure 2:** Partial family tree of Italic and Latin. Dashed lines indicate structural omissions. The colored points represent the time periods of PILA’s data points.

# The Availability of Protoform Datasets

- A decent quantity of phonetic historical linguistics datasets have become available—especially through Lexibank (List *et al.*, 2022).
- Of datasets including proto-forms, only eight language families of approximately 245, according to the Glottolog (Hammarström *et al.*, 2023), have any coverage. The Italic subfamily did not previously have coverage.

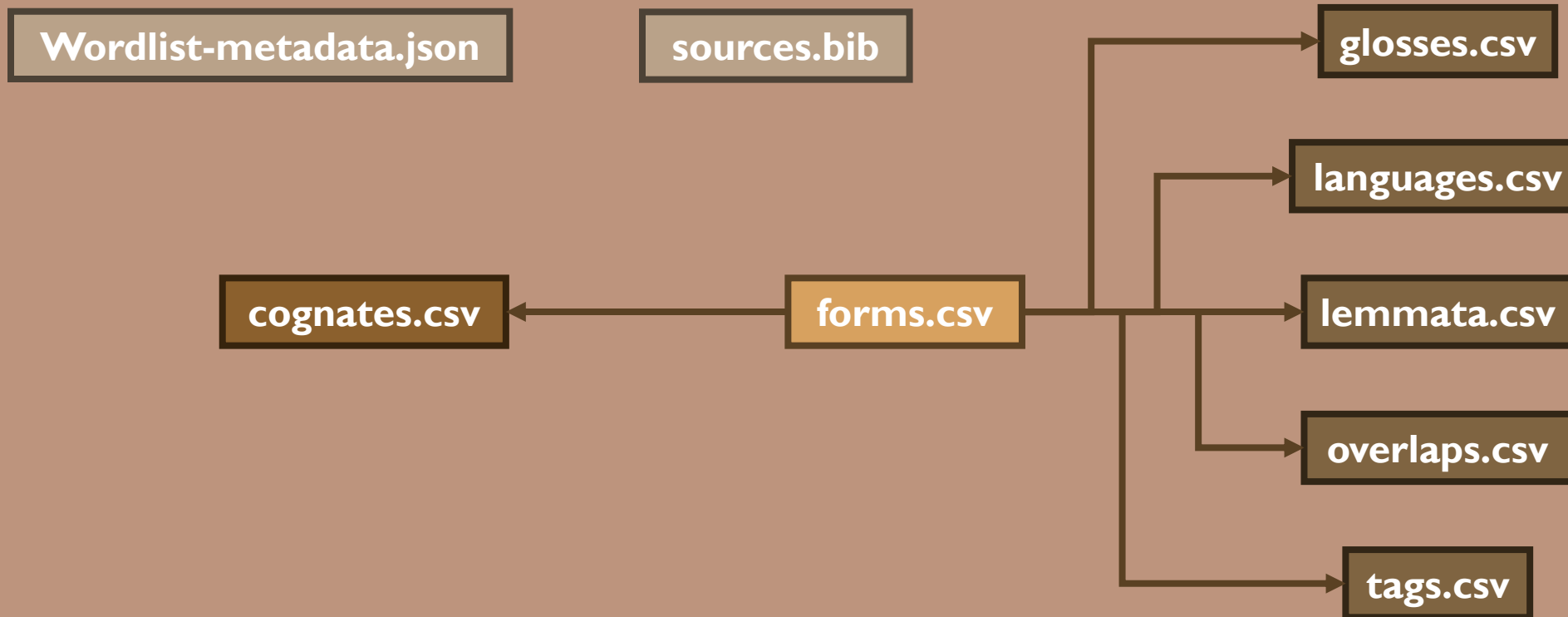
| Dataset                            | Ancestor            | Descendant  | Pairs  |
|------------------------------------|---------------------|-------------|--------|
| IELEX (LRC, 2024)                  | Proto-Indo-European | English     | 14,697 |
| Luo (2021)                         | Proto-Germanic      | Old English | 4,599  |
| JAMBU (Arora <i>et al.</i> , 2023) | Proto-Dravidian     | Tamil       | 4,276  |
| PILA (Ours)                        | Proto-Italic        | Latin       | 2,916  |
| Cathcart and Wandl (2020)          | Proto-Slavic        | Russian     | 1,572  |
| [11 datasets omitted]              |                     |             |        |

**Table 1:** Display of maximal pair counts of proto-languages and their descendants in historical linguistic datasets.

# Proto-Italic and Latin: the Dataset

# Dataset Structure

- **Template:** Cross-Linguistic Data Format [CLDF] (Forkel *et al.*, 2018)



**Figure 3:** Schema for the PILA dataset.

# Dataset Development: Scraping

- To begin developing our dataset, we scraped data from Wiktionary.



The image shows a screenshot of a Wiktionary page for the Latin word "vēlum". The page is structured with sections: "Latin", "Etymology", "Pronunciation", and "Noun".

- Latin** [edit]
- Etymology** [edit]

From Proto-Italic **\*wekslom** note the Latin term's diminutive form *vēxillum* (as in *pālus* > *pāxillus*), which lends credence to this reconstruction), with two competing theories:

  - From Proto-Indo-European *\*wegslom*, from *\*weg-* ("to weave, bind"). Cognate with English *wick*, Welsh *gweu* ("to weave").<sup>[1]</sup>
  - Others refer it to *\*weǵh-* ("to ride"), thus "that which propels"; in this case, cognate with Proto-Slavic *\*veslo* ("oar"). This is semantically less attractive than the above theory.
- Pronunciation** [edit]
  - (Classical) IPA<sup>(key)</sup>: /ˈʋeː.lum/, [ˈʋeːt̪õ]
  - (modern Italianate Ecclesiastical) IPA<sup>(key)</sup>: /ˈve.lum/, [ˈvɛːlum]
- Noun** [edit]

**vēlum** ? (genitive **vēlī**); *second declension*

  1. a cloth, covering, curtain, veil, awning [quotations ▼]
  2. (usually in the plural) the sail of a ship [quotations ▼]
  3. (anatomy) the soft palate

**Figure 4:** Sample Wiktionary page with Proto-Italic etymon and Latin reflex.

# Dataset Development: Trimming & Normalization

- We cut some pairs extracted from Wiktionary. Examples of trimmed items include:

□ Partially-Reconstructed Forms:

|                   | Etymon         | Reflex            |
|-------------------|----------------|-------------------|
| PoS Mismatch      | <b>wezor</b>   | <b>ve:rna:lis</b> |
| Morpheme Mismatch | <b>mene:o:</b> | <b>e:mineo:</b>   |

Table 2a: Collection of trimmed partial reconstructions.

- We normalized our phonetic representations to the notation used in the current exemplar for Latin historical linguistics: the *Etymological Dictionary of Latin and Other Italic Languages* (de Vaan, 2008).

# Dataset Development: Augmentation

- The data collected from Wiktionary filled only so many categories of forms in Latin and Proto-Italic. Because of this, we augmented our dataset in two ways:

## □ Novel Forms:

| Etymon         | Reflex        |
|----------------|---------------|
| <b>mowawai</b> | <b>mo:vi:</b> |
| <b>piktos</b>  | <b>pictus</b> |

Table 3a: Examples of added novel forms.

## □ Inflected Forms:

| Etymon             | Reflex             |
|--------------------|--------------------|
| <b>keiwita:tes</b> | <b>ci:vi:tatis</b> |
| <b>wide:eti</b>    | <b>videt</b>       |

Table 3b: Examples of added inflected forms.

# Dataset Development: Annotation

- To account for the role of non-phonetic changes in our dataset, we tagged entries with up to five different categories of irregularity. Glosses were provided to indicate the reason for such tags.

| Category          | Example                 |                |                |
|-------------------|-------------------------|----------------|----------------|
|                   | Etymon                  | Expected       | Actual         |
| Borrowing         | <b>g<sup>w</sup>o:s</b> | <b>vo:s</b>    | <b>bo:s</b>    |
| Paradigm Leveling | <b>weznos</b>           | <b>ve:nus</b>  | <b>ve:ris</b>  |
| Phonology         | <b>dikitos</b>          | <b>dicitus</b> | <b>digitus</b> |

**Table 4:** Sampling of PILA's sound change irregularity categories with examples. Definitions deferred to paper and documentation.

# PILA: Dataset Summary

- We initially began with a dataset of 1,205 pairs after scraping and trimming etymological data from Wiktionary.
- The subsequent steps produced a dataset with 2,916 pairs.

|                  | Latin         | Proto-Italic  | All           |
|------------------|---------------|---------------|---------------|
| Forms            | 2860          | 2916          | 5776          |
| Phones           | 15974         | 18779         | 34753         |
| Phone Types      | 33            | 41            | 48            |
| Avg. Seq. Length | $5.6 \pm 1.4$ | $6.4 \pm 1.8$ | $6.0 \pm 1.7$ |

**Table 5:** Collection of statistics regarding PILA. The first three statistics are counts, whereas the last statistic is an average with standard deviation.

# Applications: Phonetic Transduction

# Phonetic Sequence Transduction

- We define two (traditional) computational historical linguistics tasks:

□ **Reflex Prediction:** Given an etymon, predict a reflex.

*s e k<sup>w</sup> o n t i n o s*  *s e c u n d u s*

□ **Etymon Reconstruction:** Given a reflex, predict an etymon.

*s e k<sup>w</sup> o n t i n o s*  *s e c u n d u s*

# Modeling and Results

- We apply two models to transduce sequences:
  - **Copying**: The input sequence is copied to the output sequence.
  - **Transformer**: A standard Transformer encoder-decoder model is trained to produce the output sequences from the input sequences (Vaswani *et al.*, 2017).

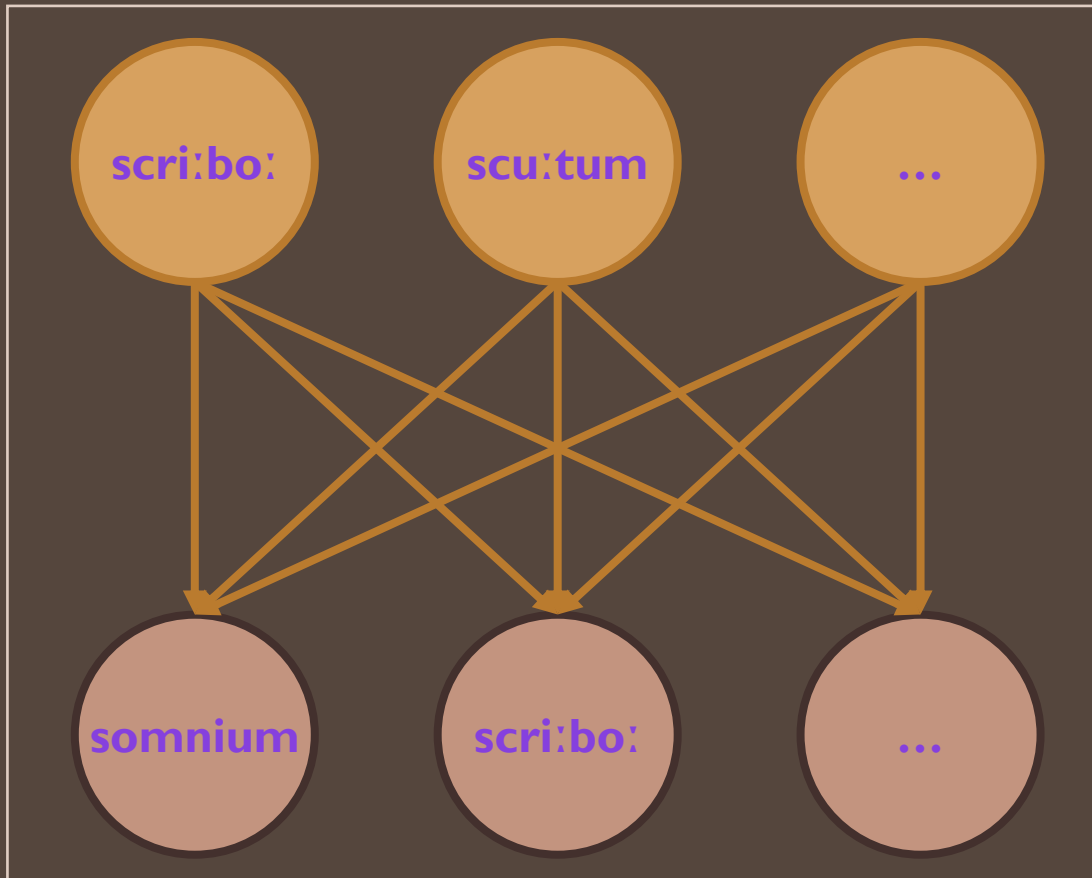
| Model       | Proto-Italic → Latin |         | Latin → Proto-Italic |         |
|-------------|----------------------|---------|----------------------|---------|
|             | PER (↓)              | WER (↓) | PER (↓)              | WER (↓) |
| Copying     | 0.53                 | 0.98    | 0.46                 | 0.97    |
| Transformer | 0.18                 | 0.52    | 0.24                 | 0.73    |

**Table 6:** Results for phone prediction tasks on PILA’s test set.

Applications:  
Dataset  
Compatibility

# The Algorithm

- Using the Hungarian algorithm (Kuhn, 1955; Munkres, 1957), we construct and compute a one-to-one matching between forms of a given language in two datasets.
- We compute the *overlap* by tallying valid matches. We apply this procedure twice:
  - ❑ **Direct Overlap:** only the normalized headword forms of each word will be matched.
  - ❑ **Indirect Overlap:** we remove vowel lengths from PILA's headwords and derive inflections for them.



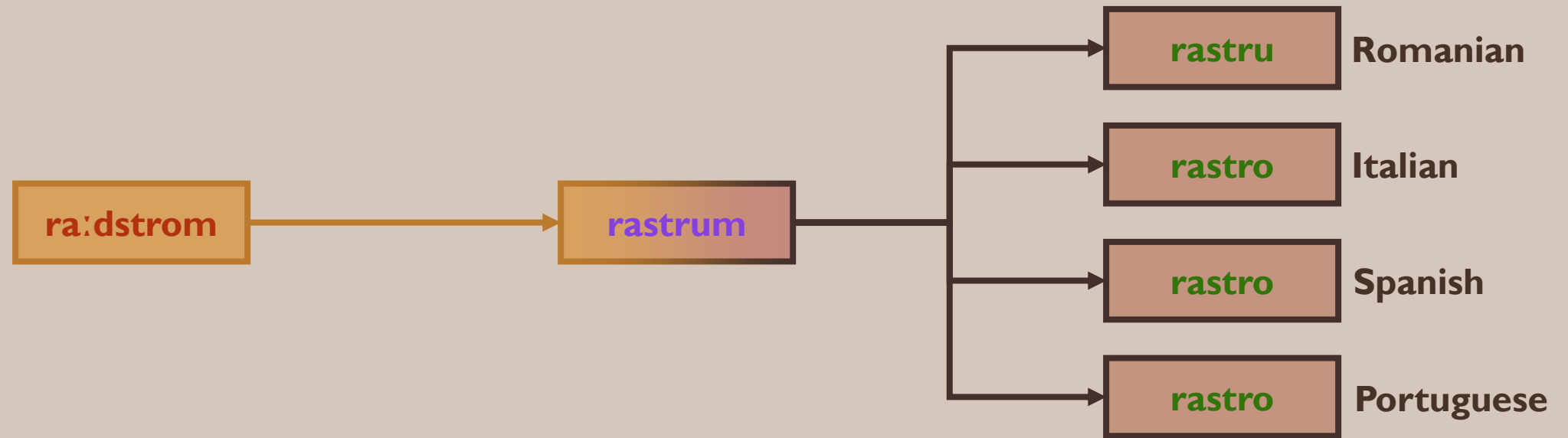
**Figure 5:** Visualization of matching procedure in dataset compatibility algorithm.

# Dataset Compatibility Results (I)

| Dataset                                | Latin Forms | Overlaps |          |       |
|--|-------------|----------|----------|-------|
|  |             | Direct   | Indirect | Total |
| Ciobanu and Dinu (2014)                | 3218        | 147      | 31       | 178   |
| Meloni <i>et al.</i> (2021): Additions | 5419        | 68       | 580      | 648   |
| Meloni <i>et al.</i> (2021): Full      | 8799        | 135      | 847      | 982   |

**Table 7:** Dataset compatibility study results for three related datasets.  
Values are all counts.

# Dataset Compatibility Results (II)



**Figure 6:** Visualization of sample data points connected through the compatibility study. The Latin form is present in both datasets. Meanwhile, the Proto-Italic form is in PILA and the Romance language forms are in Meloni *et al.*'s additions.

# Conclusion and End Matter

# Conclusion

- **In this work, we:**
  - ❑ ... established a sizable phonetic dataset of etymon-reflex pairs between Proto-Italic and Latin—a previously unaddressed language pair.
  - ❑ ... detailed our development process so as to lay groundwork for the creation of future datasets.
  - ❑ ... provided baseline results on traditional historical-linguistic transduction tasks.
  - ❑ ... linked our dataset with other datasets to promote studies of systemic sound change processes.

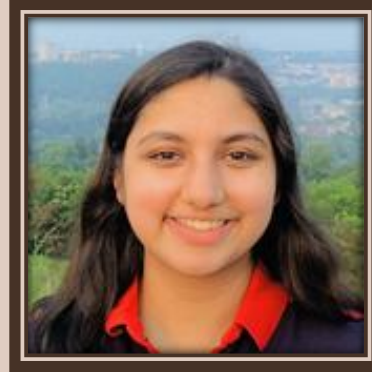
# Acknowledgements



**Darcey  
Riley**



**Ken  
Sible**



**Aarohi  
Srivastava**



**Chihiro  
Taguchi**



**Andy  
Yang**



**Arora  
Aryaman**



**Alina Maria  
Cristea**



**Liviu P.  
Dinu**



**Todd  
Krause**



**Shauli  
Ravfogel**

# Thanks for Watching!



Link: <https://bit.ly/pila-dataset>

Dataset: <https://github.com/Mythologos/PILA>

Code: <https://github.com/Mythologos/PILA-Code>