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

Stephen Bothwell,* **Brian DuSell,†** **David Chiang,*** **Brian Krostenko***

*University of Notre Dame
Notre Dame, Indiana, USA
{sbothwel,dchiang,bkrosten}@nd.edu

†ETH Zürich
Zürich, Switzerland
brian.dusell@inf.ethz.ch

Abstract

Computational historical linguistics seeks to systematically understand processes of sound change, including during periods at which little to no formal recording of language is attested. At the same time, few computational resources exist which deeply explore phonological and morphological connections between proto-languages and their descendants. This is particularly true for the family of Italic languages. To assist historical linguists in the study of Italic sound change, we introduce the Proto-Italic to Latin (PILA) dataset, which consists of roughly 3,000 pairs of forms from Proto-Italic and Latin. We provide a detailed description of how our dataset was created and organized. Then, we exhibit PILA's value in two ways. First, we present baseline results for PILA on a pair of traditional computational historical linguistics tasks. Second, we demonstrate PILA's capability for enhancing other historical-linguistic datasets through a dataset compatibility study.

Keywords: historical linguistics, Latin, Proto-Italic, resource

1. Introduction

All languages change over time, but much work in computational linguistics views language as a static phenomenon. Most natural language processing models (*e.g.*, for named entity recognition or machine translation) are trained on snapshots of languages at a fixed point in time. In contrast, historical linguistics—and, by extension, computational historical linguistics—attempts to track linguistic shifts across various points in time.

The phonological side of historical linguistics examines the relations among *cognate sets*: forms in different languages that are related etymologically. If forms have a common ancestor, or *etymon* (pl. *etyma*), they are known as *reflexes* of that etymon. Historical linguists hypothesize systems of sound change to explain the evolution of language over time. However, as we proceed further into the past, reconstruction becomes more difficult or even speculative due to the dearth of existing evidence.

Consider the examples presented in Table 1 for the Proto-Italic and Latin languages. Each row corresponds to a sound change pattern. *Cluster reduction*, for instance, involves the collapse of a longer consonant cluster into a shorter one. Here, the reconstructed `*takslos` reduces `*ksl` into `l`.[1] How does this collapse occur? Many sequences of rules could be proposed to explain the phenomenon. To reconstruct the true process, historical linguists gather evidence by comparing languages in close temporal and geographic proximity. Yet, to our knowledge, such efforts have not demonstrated the precise sequence of rules that underlies cluster reduction—alongside other sound change patterns.

Pattern	Proto-Italic	Latin	Gloss
Vowel weakening	`*fragelis`	`fragilis`	'brittle'
Cluster reduction	`*takslos`	`taːlus`	'ankle'
Syncope	`*aːzideːjoː`	`aːrdeoː`	'I burn'
Metathesis	`*pauros`	`parvus`	'few'

Table 1: Sampling of sound change patterns present in PILA. Phones affected by the listed pattern are underlined and written in boldface. Here and elsewhere we write Proto-Italic forms in orange (with asterisks), and Latin forms in purple.

To scrutinize systems of sound change and build models to capture them, we need sizable datasets. Toward that end, we introduce PILA, the **P**roto-**I**talic to **La**tin dataset.[2] PILA contributes to the study of historical linguistics in many ways. Namely:

- PILA is the first dataset to contain full (and not partial; see Section 4.2.2) reconstructions of Proto-Italic etyma and their Latin reflexes.
- PILA is one of the largest available datasets of etymon–reflex pairs for a single proto-language. (See Table 3.)
- PILA provides multiple inflections for most lemmata, letting phonological studies consider morphology's influence. (See Section 4.2.5.)
- PILA highlights the presence of non-phonological changes (*e.g.*, analogy) through per-entry annotations. (See Table 5.)

[1] In this paper as well as in our dataset, we use phonetic transcriptions for both languages. We detail our transcription scheme in Section 4.2.3.

[2] Our dataset is available at the following location: https://github.com/Mythologos/PILA.

LREC-COLING 2024, pages 12749–12760

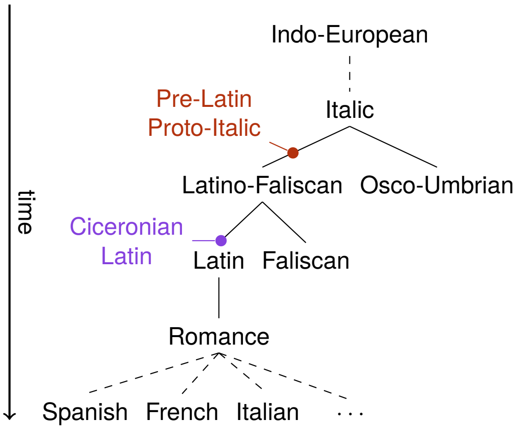

Figure 1: Partial family tree of Italic and Latin. Dashed lines indicate the omission of some intermediate families and branching structures from the tree. The two points represent the time periods covered by our dataset.

After situating our work in Sections 2 and 3, we describe our dataset (Section 4.1) and its development (Section 4.2). Then, in Section 5, we exhibit PILA's applicability through strong baseline results on standard historical linguistics tasks and a careful dataset compatibility study. We further provide a table in our dataset to encourage studies which span multiple historical linguistics datasets. (See Section 5.2 for details.)

2. Background

The Proto-Indo-European (abbreviated PIE) language is a reconstructed, unattested language. Historical linguists theorize that as the language spread, it split up into branches characterized by shared innovations. One set of innovations characterizes a branch conventionally named Proto-Italo-Celtic, which is the parent of later Italic and Celtic languages. One branch of that family, characterized by further innovations (de Vaan, 2008, 8), is dubbed Proto-Italic. That language is the parent of a family of attested languages spoken chiefly in the Italian peninsula, which fall into dialect or family groups, generally along geographical lines.

The language family centered on the Apennine spine is Osco-Umbrian, divided into Oscan in the south and Umbrian in the north. Meanwhile, the language family originally spoken in the lower country along the Tyrrhenian coast between Etruria and the Campania is Latino-Faliscan, divided into Faliscan in the north and Latin in the south. As the city-state of Rome developed into an international power, Latin spread widely, eventually supplanting the other Italic languages. Thus, Latin is undoubtedly the best attested of the older Italic languages, and it becomes the source of the Romance languages in turn. We depict a summarized version of this language family tree in Fig. 1.

Turning to these languages' phonetic inventories, Proto-Italic inherits a fairly small set of phones from PIE (Beekes, 1995, 124). By the time Proto-Italic splits up, laryngeals—a class of sounds with a guttural articulation—disappear. They leave behind traces of various kinds (*e.g.*, long vowels, the creation of `a`). Meanwhile, PIE's voiced aspirates generally become fricatives, and their vocalized nasals become nasal stops (`em`, `en`). Within Latin the fricatives are heavily reordered, sometimes becoming stops; diphthongs have become or are becoming monophthongs; and the labiovelars sometimes have lost labial or velar articulation.

PIE is a highly inflected language; Proto-Italic and Latin follow suit, although these languages gradually simplify PIE's system. Like PIE, Proto-Italic and Latin inflect through patterns of suffixation. PIE has a nominal system with stem classes that fall into three agreement categories, conventionally called genders, and can be suffixed into eight or nine cases. By the time of Latin, only five or six cases (*e.g.*, genitive, accusative) are in full use. As for PIE's verbal system, it employs vowel changes in the stem and suffixation variously to assume different aspects, moods (*e.g.*, indicative, imperative), and persons (1^{st}, 2^{nd}, and 3^{rd}). Latin largely drops the meaning behind vowel changes in verb stems and instead standardizes these changes into four conjugations. Every verb is characterized by number (singular, plural), person, tense (*e.g.*, present, perfect), mood, and voice (*e.g.*, active, passive).

3. Related Work

Recent interest in computational historical linguistics has spurred an uptick in the number of available datasets—including those with proto-languages. Work from Lexibank (List et al., 2022), a large collection of historical linguistics datasets converted into the Cross-Linguistic Data Format (CLDF) (Forkel et al., 2018), has been central to the increase in accessible datasets across language families.

Nevertheless, the number of datasets with proto-languages remains small, and coverage of language families could be greatly improved. Lexibank contains languages from four of Glottolog's 245 language families (Hammarström et al., 2023). With the inclusion of all non-Lexibank datasets that we know of, a total of eight language families have received attention. We list all datasets containing proto-languages with cognate relationships in Table 2. As this table illustrates, our work fills a gap, being the first dataset to provide explicit coverage for Italic. Table 3 further shows that PILA is one of the largest datasets of its kind.[3]

[3]Because it does not differentiate attested Sanskrit

Family	Coverage
Arawakan	Purus† (de Carvalho, 2021)
Austronesian	Austronesian Basic Vocabulary Database† (Greenhill et al., 2008), JAMBU [Munda] (Arora et al., 2023), Micronesian Comparative Dictionary† (Bender et al., 2003a,b), Tukanoan (Chacon, 2014)
Dravidian	JAMBU (Arora et al., 2023)
Indo-European	Indo-European Cognate Relationships (IE-CoR) Database (Heggarty et al., 2023), Germanic (Luo, 2021), Indo-European Lexicon (IELEX) (Linguistics Research Center, 2024), JAMBU (Arora et al., 2023), **PILA** [Italic] (Ours), Slavic (Cathcart and Wandl, 2020)
Mongolic-Khitan	JAMBU [Mongolic] (Arora et al., 2023)
Sino-Tibetan	Bai† (Wang, 2004), Burmish† (Gong and Hill, 2020), Karen† (Luangthongkum, 2019), Lalo† (Yang, 2023), Tujia† (Zhou, 2020)
Uto-Aztecan	Aztecan† (Davletshin, 2012), Corachol and Náhuatl† (Pharao Hansen, 2020)
Turkic	JAMBU (Arora et al., 2023)

Table 2: A depiction of proto-form dataset coverage across language families. If datasets are named, their names are presented followed by the language subfamilies which they concern, if applicable, in square brackets. Otherwise, the name(s) of contained language subfamilies are used to represent the dataset. Items are marked with a † if they are, at the time of writing, made available through Lexibank.

Dataset	Ancestor	Descendant	Pairs
IELEX (LRC, 2024)	Proto-Indo-European (indo1319)	English (stan1293)	14697
Luo (2021)	Proto-Germanic (germ1287)	Old English (olde1238)	4599
JAMBU (Arora et al., 2023)	Proto-Dravidian (drav1251)	Tamil (tami1289)	4276
PILA (Ours)	Proto-Italic (ital1284)	Latin (lati1261)	2916
Cathcart and Wandl (2020)	Proto-Slavic (slav1255)	Russian (russ1263)	1572
MCD (Bender et al., 2003a,b)	Proto-Chuukic (truk1243)	Chuukese (chuu1238)	1460
Yang (2023)	Proto-Lalo	Shuizhuping (west1506)	954
Wang (2004)	Proto-Bai	Enqi (enqi1234)	455
Luangthongkum (2019)	Proto-Karen (kare1337)	Northern Sgaw	366
Zhou (2020)	Proto-Bizic	Cebu (sout2739)	324
Gong and Hill (2020)	Proto-Burmish (burm1266)	Maru (maru1249)	264
de Carvalho (2021)	Proto-Purus (puru1265)	Yine (yine1238)	201
IECoR (Heggarty et al., 2023)	Proto-Indo-European (indo1319)	Old Polish (poli1260)	146
Chacon (2014)	Proto-Tucanoan (tuca1253)	Tukano (tuca1252)	131
Davletshin (2012)	Proto-Aztecan (azte1234)	Classical Nahuatl (clas1250)	84
Pharao Hansen (2020)	Proto-Nahua (azte1234)	Cora (cora1259)	55

Table 3: A comparison of datasets containing proto-languages. Each language pairing presented is the pairing with the maximal number of etymon–reflex pairs for the given dataset. Each language is joined by its provided (or otherwise determinable) Glottocode. Reconstructions are included whether they are *full* or *partial* (as defined in Section 4.2.2), as it is nontrivial to distinguish between these reconstruction types.

Outside of Lexibank, there are other datasets that document etymon–reflex relations between attested languages. For example, the WikiHan dataset compares Middle Chinese to eight Chinese subgroups (Chang et al., 2022). Another group of related datasets compare Latin with five descendant Romance languages (Ciobanu and Dinu, 2014; Meloni et al., 2021). One of two Coglust (Cognate Clustering) datasets similarly deals with Latin and Romance languages, adding Catalan to their set; the other contains Turkic and six of its descendants (Wu and Yarowsky, 2018).

Yet other datasets have less of a focus on modeling the relationship between specific ancestor and descendant languages and instead identify cognate sets over a wide range of languages. Namely, the Indo-European Lexicon (IELEX) (Linguistics Research Center, 2024), the Indo-European Cognate Relationships (IE-CoR) database (Heggarty et al., 2023), CogNet (Batsuren et al., 2019, 2022), the JAMBU database for South Asian languages (Arora et al., 2023), and a pair of cognate datasets descending from Proto-Germanic and Latin from Luo all contain over 100 languages.

and unattested proto-forms, an Old Indo-Aryan dataset (Cathcart and Rama, 2020) is left out from Tables 2 and 3.

	Latin	Proto-Italic	All
# Forms	2860	2916	5776
# Phones	15974	18779	34753
# Phone Types	33	41	48
Avg. Length	5.6 ± 1.4	6.4 ± 1.8	6.0 ± 1.7

Table 4: Collection of PILA dataset statistics relative to its component languages. The first section of the table consists of frequencies, whereas the second contains averages with standard deviations.

4. PILA: Proto-Italic to Latin Dataset

4.1. Overview

PILA is intended to document etymon–reflex relationships between Proto-Italic and Latin. However, languages change over time even as they bear the same name. Therefore, to assure consistency in etymon–reflex relationships, PILA specifically captures Proto-Italic and Latin at stages of development we call "Pre-Latin Proto-Italic" and "Ciceronian Latin" (see Fig. 1).

We selected the former to focus PILA on phonetic changes, as this stage of Proto-Italic has already undergone many non-phonological changes which would obscure sound change laws. As for Ciceronian Latin, we chose it not only because it serves as a standard version of Latin (promoted by the famed orator, Cicero) but also because its distance from Pre-Latin Proto-Italic allows for many meaningful phonetic changes to have occurred.

To store our dataset, we use the Cross-Linguistic Data Format (CLDF) (Forkel et al., 2018). This format consists of CSV, JSON, and BibTEX files that conform to a specification built on the CSV on the Web (CSVW) standard. By providing explicit and uniform structural requirements, CLDF enforces a principled style of data table design and allows adherent datasets to readily work with libraries supporting the data format (*e.g.*, the historical linguistics library LingPy (List and Forkel, 2021)).

Our dataset builds on CLDF's `Wordlist` module. Disregarding the dataset's JSON table metadata file and its BibTEX file to store cited sources, seven tables (CSV files) constitute the dataset:

1. `languages.csv`: a collection of languages contained in the dataset and their attributes.

2. `forms.csv`: a collection of phonetic sequences and tokenized phones. It contains identifiers to link forms with all other tables in PILA. Statistics for these forms are collected in Table 4.

3. `cognates.csv`: a collection of numeric identifiers which link each form to its cognate set.

4. `lemmata.csv`: a collection of lemmata to which our forms are morphologically related. Currently, only Latin forms have lemmata.

5. `glosses.csv`: a collection of notes and tabulated irregular phenomena for our forms. For our irregularity categories, see Table 5.

6. `tags.csv`: a collection of groups of morphological tags which correspond to Latin forms in our dataset. See Section 4.2.5 for details.

7. `overlaps.csv`: a collection of identifiers linking forms in PILA and in others' datasets together. See Section 5.2 for more details.

4.2. Development

In this section, we describe our dataset development procedure. We worked closely with an expert in historical linguistics for Proto-Italic and Latin (the last author) and a graduate student with a decade of Latin experience (the first author) to scrape, trim, normalize, augment, and annotate the data. We describe each step below. Although this procedure is framed in the context of Latin and Proto-Italic, many of its steps (*e.g.*, our scraping procedure) could be performed with minor contextual changes for any language pair.

4.2.1. Scraping

We initially extracted data from Wiktionary. Because of public availability and decent etymological curation, Wiktionary was deemed to be a suitable starting point for PILA. We scraped all pages tagged as "Latin terms derived from Proto-Italic" for their Latin headwords and Proto-Italic forms (tagged with the `lang` attribute "itc-pro") (Wiktionary contributors, 2017).[4] For each "etymology" heading in the Latin section of a headword's page, we extracted an etymon–reflex pair. (Multiple etymologies are possible if multiple Latin words happen to have the same spelling.) We automatically cleaned the natural-language formatting of etymologies. We filtered out affix headwords (as they do not occur in natural speech), mislabeled headwords, and reconstructed Latin headwords from our data.

4.2.2. Trimming

We manually classified etymon-reflex pairs as *partial* or *full* reconstructions. Partially-reconstructed pairs have an etymon either with mismatching parts of speech or with mismatching or missing morphemes. For example, Wiktionary reported (at the time of data collection) an etymology for the Latin verb `audeo:` 'I dare' as the Proto-Italic adjective

[4] The scraping was done on February 15th, 2022.

*awidos. Meanwhile, a fully-reconstructed pair has matching parts of speech and morphemes; in PILA, we have such pairs in *awide:o: and audeo: and also in *awiðos and avidus 'desirous'. Having fully-reconstructed forms is desirable because these forms fully represent phonological developments with respect to both stems and affixes, whereas partially-reconstructed forms do not. Therefore, we dropped the partially-reconstructed forms, producing a set of 1,205 fully-reconstructed pairs.

Once these pairs were gathered, our experts examined them for quality. All Proto-Italic forms were checked against standard etymological dictionaries and grammars (Sihler, 1995; Meiser, 2010; Leumann, 1977; Walde and Hofmann, 1938; Ernout and Meillet, 2001), such as the Etymological Dictionary of Latin (henceforth EDL) (de Vaan, 2008). This resulted in the deletion of various forms. Such forms largely fell into one of two categories.

First, a form could have either fallen out of use before our Ciceronian Latin period or could have become attested only after that period. Examples of these include aevita:s 'age, lifetime', which was superseded by aeta:s, for early attestation and genimen 'progeny' for late attestation. Second, a form could have been a proper name. Proper names have a tendency to contain non-Latin or non-Indo-European phonetic sequences, diverging from PILA's focus on Italic phonetics. An example is the Latin ti:bur, the name of a town.

We also made minor alterations to some forms to reflect known Latin developments. These tend to have case-by-case explanations, which we store among our notes in the glosses.csv table.

4.2.3. Normalization

We normalized the Latin and Proto-Italic data to consistent phonetic representations. This step was necessary because Wiktionary contributors followed different standards in different entries.

Proto-Italic forms were normalized to EDL's conventions, as this work is a current exemplar for Latin historical linguistics (de Vaan, 2008). However, a few alterations were made to improve the phonetic representations. First, because voicing assimilation appears to be a consistent feature of Pre-Latin Italic, *s before a voiced consonant or between vowels was written as *z. For example, the etymon for the adjective numerus 'number' is *nomezos instead of *nomesos. Second, fricatives that resulted from the loss of PIE's voiced aspirates were reconstructed as voiceless in initial position and as voiced between vowels or before a voiced consonant. The word faber 'craftsman' exhibits this well: its etymon, *faβer, has an initial voiceless fricative (*f) and a medial voiced fricative (*β).

Meanwhile, for Latin, we largely preserved its orthography but again made alterations when the orthography hid the actual pronunciation. We changed n or g to ŋ where they were actually pronounced as ŋ. For instance, magnus 'great' was written as maŋnus and frango: 'I shatter' as fraŋgo:. Similarly, we duplicate i when the orthography conceals that it is used both as a vowel and as a consonant, as in maiior 'greater' and cuiius 'whose'.

4.2.4. Augmentation: Novel Forms

We added a variety of forms to the cleaned, fully-reconstructed set. In consultation with EDL (de Vaan, 2008), we completed some of the partial reconstructions and appended additional forms. Some appended forms were chosen by our experts to provide representation for missing morphological features. Specifically, the scraped pairs did not include any perfect-tense forms. Thus, we added some in both the active and passive voices.

The perfect active "third principal part" for standard Latin verbs largely falls into one of four main categories: reduplicating perfects (*e.g.*, fefelli: 'I deceived' for fallo: 'I deceive'); root aorists (*e.g.*, ru:pi: 'I broke' for rumpo: 'I break'); s aorists (*e.g.*, scri:psi: 'I wrote' for scri:bo: 'I write'); and vi:/ui: perfects (*e.g.*, ama:vi: 'I loved' for amo: 'I love' or monui: 'I warned' for moneo: 'I warn'). Some of these forms pose problems.

Regarding reduplicating perfects, the inherited reduplicating vowel is e (*e.g.*, memordi: 'I bit' for mordeo: 'I bite'). Apparently, when a verbal root itself contained an *e, the generative principle for the reduplicated syllable was reinterpreted to incorporate the vowel of the root instead of *e with the initial consonant. That became the standard pattern in reduplicated perfects (*e.g.*, cucurri: 'I ran'—and not cecurri:—for curro: 'I run'), and we reconstructed reduplicating perfects accordingly.

Meanwhile, concerning vi:/ui: perfects, their origins are not clear. Latin is the only Italic language that uses that suffix for the perfect. A form like *monawai producing monui: may therefore not be a truly Proto-Italic feature but a late, "Pre-Latin" one. However, given the importance of these perfects in the Latin system and the Pre-Latin character of our Proto-Italic, we decided to include them.

Turning to perfect passive forms, we added the Latin paradigm-defining *to participle. In some cases this participle's form is due to paradigm leveling (see Table 5) or other types of analogical change (*e.g.*, perhaps pulsus 'having been struck' after the perfect active pulsi: 'I struck'). Such forms were excluded, and older forms, where known or securely reconstructable, were used (*e.g.*, pultus, which must have existed, as it is the base of the frequentative pulta:re 'to knock').

Through this process, we created a set of 1,515 Latin forms and 1,548 Proto-Italic forms.

Category	Definition	Example		
		Etymon	Expected	Actual
Association	A form with a change attributable to its association with another word or semantic set.	*kleiments	cli:me:ns	cle:me:ns
Borrowing	A form with a sound change different from standard Latin and/or from another Italic dialect.	$*g^{w}$o:s	vo:s	bo:s
Morphology	A form where morphosemantics undid, blocked, or otherwise interfered with expected sound changes.	*mensen	me:nsen	me:nsis
Paradigm Leveling	A form which exhibits changes taken from another part of its own paradigm or lexical family.	*weznos	ve:nus	ve:ris
Phonology	A form with an uncommon or unexpected phonological phenomenon.	*dikitos	dicitus	digitus

Table 5: PILA sound change irregularity categories, definitions, and examples.

4.2.5. Augmentation: Inflections

We elected to further augment this data with inflected forms. This seemed crucial for our target languages, as they are rife with inflections. Moreover, sound changes may occur differently in the citation form and inflected forms of a word. For instance, the Latin cru:s 'leg' has a genitive form cru:ris 'of the leg'. The final s in cru:s and the medial r in cru:ris arise from the same original *s; however, because the genitive suffix, is (Proto-Italic *es), begins with a vowel, the stem's *s found itself between two vowels, causing it to rhotacize. Many other Latin forms display similar behaviors (*e.g.*, flo:s 'flower' and flo:ris 'of the flower').

Because nouns, adjectives, participles, and verbs are the primary subjects of inflection in Latin, we added inflections for words in those categories.

Nominal Forms To most nouns, adjectives, and participles, we added genitive singular forms. In all declensions, these endings pose certain difficulties.

In the first or a: declension, Latin inherited the PIE ending *a:s, preserved in archaic phrases and older texts. But the analogical ending *a:i:, with *i: imported from the second declension, became the standard ending. We used that ending, as it is the result of non-phonological changes that lead into Pre-Latin Proto-Italic. Meanwhile, in the second or o declension, Latin preserves the inherited ending *osyo quite rarely. In attested Latin, the ending is i:; thus, we used that ending in Proto-Italic as well.

Lastly, the Latin third declension represents a fusion of PIE i-stems and consonant stems. These i-stems affixed consonant stem endings to an ablauting medial *ej. The earlier Latin i-stem genitive ending was i:s from *ejes. But that affix was replaced by the is of the consonant stems. That affix, in turn, represents the generalization of the outcome of inherited *es over the outcome of inherited *os (attested, very rarely, as us). All Proto-Italic i-stems and consonant stems were thus reconstructed with genitive *es.

Verbal Forms The 3rd-person singular present of all indicative verbs was added. It is, on some accounts, the least marked member of a paradigm. A problem is posed by the so-called *hic et nunc* ("here and now") marker *i, which tagged present tense forms in PIE but disappeared in Proto-Italic. The contrast in Old Latin third person endings *ti and/or *t producing Latin t and/or d suggests that the origin of t is (partly) phonological. Therefore, verbs were reconstructed with that ending.

Morphological Tags To provide a way to analyze PILA with respect to its inflections, we incorporate a set of morphological tags into our dataset through the `tags.csv` file. This file lists common morphological properties for all our Latin forms. We adapted the tagset used by the Perseus Project's morphological analyzer, Morpheus (Crane, 1991), for our purposes. Namely, we considered the part-of-speech, person, number, tense, mood, voice, gender, case, and degree for each form. Furthermore, we added a tag for each form's *inflection class*—that is, the paradigm of conjugation or declension to which a word belongs (if any).

Although we largely adhered to Morpheus' tagset, we made a few adjustments to suit the ambiguity that comes with examining forms devoid of sentential (and, thus, syntactic) context. For instance, although many Latin adjectives change their inflection class to conform to the gender of the noun that they modify, some—such as the word fe:li:x 'fortunate'—do not change at all. As a result, we include all possible combinations of genders as potential tags. We detail other aspects of our tagging scheme in our dataset repository.

4.2.6. Annotation

To account for the influence of irregular (*i.e.*, rare or non-phonological) changes in phonological studies, we annotated PILA's etymon–reflex pairs with tags to flag the presence of such changes. We present the categories for these tags in Table 5. In our dataset, a gloss accompanies each tag to justify its attachment to a given form. One application of these tags could be to filter PILA to a set of forms which have undergone a specific kind of change to study the effects of that change.

5. Applications

So far, we have detailed the PILA dataset. In this section, we exhibit PILA's applicability to standard computational historical linguistics studies (Section 5.1) and PILA's capacity to enhance existing datasets through overlapping forms (Section 5.2).[5]

5.1. Sample Tasks

In this section, we perform two traditional computational historical linguistics tasks. Suppose that we have a pair of languages, an ancestor E and a descendant R, both considered as sets of forms. Then we can define a set of etymon–reflex pairs $C \subseteq E \times R$. We note that both $e \in E$ and $r \in R$ could be used in more than one pair.

For this set C, we can define two tasks. First, we define *reflex prediction* as the task of predicting r, given e. Second, we define *etymon reconstruction* as the task of predicting e, given r. We perform both of these tasks with PILA below.

5.1.1. Procedure

To perform the reflex prediction and etymon reconstruction tasks, we grouped our data by lemma (citation form) to ensure that inflected forms of the same lemma stay together, and we randomly split the lemmata into training (80%), validation (10%), and test (10%) splits. This resulted in sets of 2,331, 298, and 287 etymon–reflex pairs, respectively.

We train a Transformer encoder–decoder model (Vaswani et al., 2017), and we match most of the hyperparameter settings used in that work. We use PyTorch's Transformer layer implementation (Paszke et al., 2019). Unlike Vaswani et al. (2017), we use pre-norm instead of post-norm (Wang et al., 2019; Nguyen and Salazar, 2019) and apply layer normalization to the last layer's output. We use 6 layers in both the encoder and decoder and 8 attention heads per layer. We initialize the output layer with Xavier uniform initialization (Glorot and Bengio, 2010). For layer norm, we initialize all weights to 1 and all biases to 0. We initialize all other parameters by sampling uniformly from [–0.01, 0.01].

We optimize parameters with Adam (Kingma and Ba, 2015). We clip gradients with a threshold of 5 using L^2 norm rescaling. We train the model by minimizing the decoder's cross-entropy (summed over all timesteps) on the training set. We use label smoothing with a weight of 0.1. We take a checkpoint every 2,000 examples to evaluate the decoder's per-token cross-entropy on the validation set. After two checkpoints with no improvement, we multiply the learning rate by 0.5; after two more such checkpoints, we stop training early. We train for up to 100 epochs. We use the checkpoint with the best validation cross-entropy.

For each epoch, we randomly shuffle examples and group examples of similar lengths into the same minibatch. We limit the number of tokens in the source or target side of a batch to B tokens, including padding, BOS, and EOS symbols. We use beam search decoding with a beam size of 4. We apply length normalization to hypothesis probabilities before selecting the top k hypotheses for the next beam. We do this by dividing the log-probability of each hypothesis by the number of tokens generated by that hypothesis so far (including EOS).

We perform a random hyperparameter search (Bergstra and Bengio, 2012) over 10 runs. For each run, we randomly sample four hyperparameters: the initial learning rate, the batch size B, the model size s, and the dropout rate. See Table 6 for our chosen distributions. With the model size, we set d_{model} to $8 \cdot s$ and the size of the feedforward hidden layers to $4 \cdot d_{model}$. We apply dropout as PyTorch does, which follows Vaswani et al. (2017) and also applies it to feedforward sublayers' hidden units and attention probabilities.

Hyperparameter	Distribution	Range
Batch Size	Uniform	$[32, 256]$
Dropout Rate	Uniform	$[0, 0.2]$
Learning Rate	Log-Uniform	$[0.0001, 0.01]$
Model Size	Uniform	$[4, 64]$

Table 6: Collection of hyperparameters and search spaces used in our sample tasks.

To evaluate our models, we use multi-reference word error rate (WER) and phoneme error rate (PER). "Multi-reference" means that when there are multiple references (multiple etyma for one reflex), we take the minimum error across all references. Note that for WER, the number of errors is either 0 or 1, so WER is one minus the exact match rate. For PER, we use micro-averaging: we sum the total number of edits and total number of reference symbols over the whole test set, reporting their ratio.

5.1.2. Results

In Table 7, we report results on the test set from the hyperparameter search's best-performing model on the validation set. Corresponding hyperparameters for the best models are shown in Table 8.

[5]The code for these applications, as well as for various dataset creation and analysis utilities, is available at this location: `https://github.com/Mythologos/PILA-Code`.

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

Table 7: Results for phone prediction tasks on PILA's test set. The "Copying" baseline copies the input to the output. All "Transformer" results are derived from the best model from our hyperparameter search.

Task	d_{model}	Dropout Rate	Learning Rate	Batch Size B
Proto-Italic → Latin	112	0.1665	0.00021969	138
Proto-Italic ← Latin	496	0.0875	0.00010027	81

Table 8: Randomly-searched hyperparameters of the best Transformer models from Table 7.

We compare the Transformer model discussed in the previous section to a "Copying" baseline. In this baseline, the input is simply copied to the output. This baseline is somewhat reasonable because, although Proto-Italic phones undergo many sound changes in becoming Latin, some remain recognizably unchanged. Thus, by comparing a model to the "Copying" baseline, we examine whether it learns any nontrivial sound change rules. As Table 7 shows, the Transformer baseline vastly outperforms the "Copying" baseline in both directions, indicating that this is the case—and that, in fact, PILA provides a learnable signal.

5.2. Dataset Compatibility

In this section, we show to what extent PILA's entries can be linked to those of other datasets, allowing for models to be built with longer chains of sound changes and additional linguistic metadata.

To measure PILA's capacity to link to other datasets, we tally the overlap between PILA's and other datasets' Latin forms. Because datasets organize data differently, it is nontrivial to extract overlap counts. Moreover, because datasets vary in their attention to phonetic features such as vowel length, the legitimacy of matches can be murky (*e.g.*, without vowel length, PILA's adjectives `leːvis` 'smooth' and `levis` 'light' would be indistinguishable).

To account for this, we define two categories of overlap. We say that an overlap is *direct* if phonological or morphological information is not lost in performing the match. Conversely, an overlap is *indirect* if some such information is lost. For instance, a form may need to be inflected differently, resulting only in a partial compatibility (or perhaps a false positive match due to homography) between the information stored in each dataset.

5.2.1. Procedure

Algorithm 1 sketches our overlap computation procedure. This procedure centers around the Hungarian or Kuhn–Munkres algorithm (Kuhn, 1955; Munkres, 1957), which solves the *linear sum assignment* problem. In our pseudocode, this algorithm is named LSA, and in practice we use SciPy's implementation (Virtanen et al., 2020). Given a weighted bipartite graph, the Hungarian algorithm seeks the one-to-one matching that maximizes the sum of those edges' weights.

Algorithm 1 Dataset Compatibility Algorithm

```
procedure SCORECOMP(first, second)
    M, N ← |first|, |second|
    maxSize ← max(M, N)
    scores ← zeros(maxSize, maxSize)
    for i ∈ 1 . . . M do
        for j ∈ 1 . . . N do
            if first[i] ⋂ second[j] ≠ ∅ then
                scores[i][j] ← 1
            else
                scores[i][j] ← 0
    maxEntries ← LSA(scores)
    score ← 0
    for (i, j) ∈ maxEntries do
        score ← score + scores[i][j]
    return score
```

The function SCORECOMP takes as arguments two lists of nodes (one for each dataset), and each node contains one or more forms. We create a bipartite graph whose nodes are the aforementioned nodes, and if node u and node v have some form in common, there is an edge between u and v with weight 1. Applying a linear-sum assignment algorithm results in a list of edges. The number of edges is called the *overlap*.

We apply this algorithm twice. First, we take both normalized datasets, place each form in its own node, and apply Algorithm 1 to obtain the *direct overlap*. Then, we remove all entries from both datasets that participated in the first matching. We take each dataset's remaining nodes, remove long marks (ː) from each form, and add that form back

Dataset	Latin Forms	Direct	Indirect	Total
Ciobanu and Dinu (2014)	3218	147 (4.6%)	31 (0.1%)	178 (5.5%)
Meloni et al. (2021) – Additions	5419	68 (1.3%)	580 (10.7%)	648 (12.0%)
Meloni et al. (2021) – Full	8799	135 (1.5%)	847 (9.6%)	982 (11.2%)
Coglust (Wu and Yarowsky, 2018)	27645	760 (2.8%)	521 (1.9%)	1281 (4.6%)
CogNet (Batsuren et al., 2019, 2022)	6960	354 (5.1%)	458 (6.6%)	812 (11.7%)
IE-CoR (Heggarty et al., 2022)	266	97 (36.5%)	44 (16.5%)	141 (53.0%)
IELEX (Linguistics Research Center, 2024)	10110	558 (5.5%)	621 (6.1%)	1179 (11.7%)
Jambu (Arora et al., 2023)	4	2 (50.0%)	0 (0.0%)	2 (50.0%)
Luo (2021) – Romance	10866	489 (4.5%)	676 (6.2%)	1165 (10.7%)

Table 9: Dataset compatibility study results. Columns measure degrees of overlap relative to PILA. Integers are counts; percentages are relative to the number of data points in that row's dataset. The top three datasets are from related works and are comprised of similar data.

to each node. For PILA's nodes exclusively, we use Collatinus (Ouvrard and Verkerk, 2014; Verkerk et al., 2020) as integrated into the Classical Language Toolkit (Johnson et al., 2021), to generate and include other inflections of each node's form.[6] We apply Algorithm 1 again to get the *indirect overlap*. Finally, we add the direct and indirect overlap to obtain the total overlap.

We examine nine other datasets which contain Latin forms. We also searched for fully-reconstructed Proto-Italic forms, but we found no dataset with any such forms.

5.2.2. Results

We present the results of our study for direct and indirect overlap in Table 9. In general, we find that PILA can match over 100 forms among all datasets that we examine (except for Jambu, as it contains less than 100 Latin forms), indicating a nontrivial overlap between PILA and other datasets.

While direct matches generally account for most overlapping forms, indirect matches are prominent for Meloni et al.'s datasets, for CogNet (Batsuren et al., 2019, 2022), for IELEX (LRC, 2024), and for Luo's Romance cognate dataset . For Meloni et al.'s datasets, this is because they use the accusative singular inflection for nouns, adjectives, and participles and the present infinitive for verbs as their headwords, as these inflections more strongly link the Latin forms to their Romance language counterparts. Meanwhile, for the other datasets, many matches stem from either the removal of long marks, as their use is inconsistent, or the addition of morphologically-related forms.

In light of this study's results, we created an additional `overlaps.csv` table for PILA that facilitates sharing data between datasets. This table relates the indices of matched forms and designates the type of matching attained between them.

[6] We also correct an error where Collatinus only generates inflected endings without any stem.

6. Conclusion

This paper introduced PILA, a historical-phonological dataset of etymon–reflex pairs in Proto-Italic and Latin. It described PILA's development process and organization. It provided baseline results for PILA on two historical linguistics tasks and showed PILA's capacity to enhance other datasets in a compatibility study.

Future work could expand the scope of PILA to encourage deeper historical linguistics studies. For instance, it could broaden its coverage of languages in the Italic region. Although they have scant extant data, languages like Umbrian (Dehouck, 2022) as well as Cisalpine Celtic, Faliscan, Oscan, and Venetic (Murano et al., 2023) have received some attention in computational literature.

In another direction, PILA could incorporate sets of phonological rules to support the task of automatic sound law induction (ASLI) (Luo, 2021; Chang et al., 2023). However, many issues arise when considering how to select and store such rules. For example, what formalism should be used to organize sound change rules? Although historical linguists have a standard notation for sound laws, certain features and complex conditions do not have agreed-upon, computationally-friendly notation (Luo, 2021). An examination of prior ASLI work and the sound law databases UNIDIA (Hamed and Flavier, 2009) and PBase (Mielke, 2008) could serve as a starting point for this direction.

7. Acknowledgements

Regarding the datasets used in this work, we would like to thank Todd Krause for providing us with a version of the IELEX dataset (Linguistics Research Center, 2024). We would also like to thank Alina Maria Ciobanu and Liviu P. Dinu for allowing us to use their dataset (Ciobanu and Dinu, 2014), as well as Shauli Ravfogel for providing us with their revisions to it (Meloni et al., 2021).

For their helpful comments and discussions concerning this work, we would also like to thank Darcey Riley, Ken Sible, Aarohi Srivastava, Chihiro Taguchi, and Andy Yang.

8. Bibliographical References

Robert S. P. Beekes. 1995. *Comparative Indo-European Linguistics: An Introduction*. John Benjamins Publishing Company, Amsterdam; Philadelphia, PA.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Gregory Crane. 1991. Generating and parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the Other Italic Languages*. Number 7 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden; Boston.

Alfred Ernout and Alfred Meillet. 2001. *Dictionnaire étymologique de la langue latine*. Klinksieck, Paris.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):180205.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Manu Leumann. 1977. *Lateinsche Laut- Und Formenlehre*. C.H. Beck, Munich.

Johann-Mattis List and Robert Forkel. 2021. LingPy. A Python library for historical linguistics. Max Planck Institute for Evolutionary Anthropology.

Gerhard Meiser. 2010. *Historische Laut- Und Formenlehre Der Lateinischen Sprache*. Wissenschaftliche Buchgesellschaft, Darmstadt.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Francesca Murano, Valeria Quochi, Angelo Mario Del Grosso, Luca Rigobianco, and Mariarosaria Zinzi. 2023. Describing inscriptions of ancient Italy. The ItAnt project and its information encoding process. *Journal on Computing and Cultural Heritage*, 16(3).

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*.

Yves Ouvrard and Philippe Verkerk. 2014. Collatinus, un outil polymorphe pour l'étude du Latin. *Archivum Latinitatis Medii Aevi*, 72(1):305–311.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035.

Andrew Sihler. 1995. *New Comparative Grammar of Greek and Latin*. Oxford University Press, New York; Oxford.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*.

Philippe Verkerk, Yves Ouvrard, Margherita Fantoli, and Dominique Longrée. 2020. L.A.S.L.A. and Collatinus: A convergence in lexica. *Studi e saggi linguistici*, 58(1):95–120.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Alois Walde and J.B. Hofmann. 1938. *Lateinisches Etymologisches Wörterbuch*. Carl Winter, Heidelberg.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep Transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822.

9. Language Resource References

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2023. JAMBU: A historical linguistic database for South Asian languages. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 68–77.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2022. A large and evolving cognate database. *Language Resources and Evaluation*, 56(1):165–189.

Byron W. Bender, Ward H. Goodenough, Frederick H. Jackson, Jeffrey C. Marck, Kenneth L. Rehg, Ho-min Sohn, Stephen Trussel, and Judith W. Wang. 2003a. Proto-Micronesian reconstructions–1. *Oceanic Linguistics*, 42(1):1–110.

Byron W. Bender, Ward H. Goodenough, Frederick H. Jackson, Jeffrey C. Marck, Kenneth L. Rehg, Ho-min Sohn, Stephen Trussel, and Judith W. Wang. 2003b. Proto-Micronesian reconstructions–2. *Oceanic Linguistics*, 42(2):271–358.

Chundra Cathcart and Taraka Rama. 2020. Disentangling dialects: A neural approach to Indo-Aryan historical phonology and subgrouping. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)*, volume 24, pages 620–630.

Chundra Cathcart and Florian Wandl. 2020. In search of isoglosses: Continuous and discrete language embeddings in Slavic historical phonology. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, volume 17, pages 233–244.

Thiago Chacon. 2014. A revised proposal of Proto-Tukanoan consonants and Tukanoan family classification. *International Journal of American Linguistics*, 80(3):275–322.

Kalvin Chang, Chenxuan Cui, Youngmin Kim, and David R. Mortensen. 2022. WikiHan: A new comparative dataset for Chinese languages. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3563–3569.

Kalvin Chang, Nathaniel Robinson, Anna Cai, Ting Chen, Annie Zhang, and David Mortensen. 2023. Automating sound change prediction for phylogenetic inference: A Tukanoan case study. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 129–142.

Alina Maria Ciobanu and Liviu Dinu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 1038–1043.

Albert Davletshin. 2012. Proto-Uto-Aztecans on their way to the Proto-Aztecan homeland: Linguistic evidence. *Journal of Language Relationship*, 8(1):75–92.

Fernando O. de Carvalho. 2021. A comparative reconstruction of Proto-Purus (Arawakan) segmental phonology. *International Journal of American Linguistics*, 87(1):49–108.

Mathieu Dehouck. 2022. The IKUVINA treebank. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 38–42.

Xun Gong and Nathan Hill. 2020. Materials for an etymological dictionary of Burmish. Zenodo.

Simon J. Greenhill, Robert Blust, and Russell D. Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics Online*, 4:271–283.

Mahé Ben Hamed and Sébastien Flavier. 2009. A database for deriving diachronic universals: UNIDIA. In Monique Dufresne, Fernande Dupuis, and Etleva Vocaj, editors, *Historical Linguistics 2007: Selected Papers from the 18th International Conference on Historical Linguistics, Montreal, 6–11 August 2007*, Current Issues in Linguistic Theory, pages 259–268. John Benjamins Publishing Company, Montreal, Quebec, Canada.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. glottolog/glottolog: Glottolog database 4.8.

Paul Heggarty, Cormac Anderson, and Matthew Scarborough. 2022. Indo-European Cognate Relationships database project (IE-CoR).

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroński, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Kühnert, and Russell D. Gray. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656):eabg0818.

Linguistics Research Center. 2024. Indo-European Lexicon: PIE etyma and IE reflexes.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Theraphan Luangthongkum. 2019. A view on Proto-Karen phonology and lexicon. *Journal of the Southeast Asian Linguistics Society*, 12(1):i–lii.

Jiaming Luo. 2021. *Automatic Methods for Sound Change Discovery*. Ph.D. thesis, Massachusetts Institute of Technology.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473.

Jeff Mielke. 2008. *The Emergence of Distinctive Features.* Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford.

Magnus Pharao Hansen. 2020. ¿Familia o vecinos? Investigando la relación entre el Proto-Náhuatl y el Proto-Corachol. In Rosa H. Yañez Rosales, editor, *Lenguas Yutoaztecas: Historia, estructuras y contacto lingüístico*, pages 75–108. Universidad de Guadalajara, Mexico.

Feng Wang. 2004. *Language Contact and Language Comparison: The Case of Bai*. Ph.D. thesis, City University of Hong Kong, Hong Kong.

Wiktionary contributors. 2017. Category:Latin terms derived from Proto-Italic.

Winston Wu and David Yarowsky. 2018. Creating large-scale multilingual cognate tables. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Cathryn Yang. 2023. *Lalo Regional Varieties: Phylogeny, Dialectometry, and Sociolinguistics*. Ph.D. thesis, La Trobe University, Bundoora, Victoria, Australia.

Yulou Zhou. 2020. *Proto-Bizic. A Study of Tujia Historical Phonology.* Bachelor's Thesis, Stanford University, Stanford, CA.