# Verba volant, scripta volant? Don't worry! There are computational solutions for protoword reconstruction

**Liviu P. Dinu**[♠,♡] **Ana Sabina Uban**[♠,♡] **Alina-Maria Cristea**[♡]
**Bogdan Iordache**[♠,♡] **Teodor Marchitan**[♠,♡] **Simona Georgescu**[♣,♡] **Laurențiu Zoicaș**[♣,♡]

University of Bucharest, [♠] Faculty of Mathematics and Computer Science,
[♣]Faculty of Foreign Languages and Literatures, [♡]HLT Research Center

{ldinu, auban}@fmi.unibuc.ro, alinaciobanu20@gmail.com,
iordache.bogdan1998@gmail.com, teodormarchitan@gmail.com
{simona.georgescu, laurentiu.zoicas}@lls.unibuc.ro

## Abstract

We introduce a new database of cognate words and etymons for the five main Romance languages (Romanian, Italian, Spanish, Portuguese, French), the most comprehensive one to date with over 19,000 entries. We propose a strong benchmark for the automatic reconstruction of protowords for Romance languages by applying a series of machine learning models and features on these data. The best results reach 90% accuracy in predicting the protoword of a given cognate set, surpassing existing state-of-the-art results for this task and showing that computational methods can be very useful in assisting linguists with protoword reconstruction.

## 1 Introduction and Related Work

Protoword reconstruction, consisting of recreating the words in a proto-language from their descendants in daughter languages, is central to the study of language evolution. As the foundation of historical linguistics (Campbell, 2013; Mallory and Adams, 2006) and the basis for linguistic phylogeny (Atkinson et al., 2005; Alekseyenko et al., 2012; Dunn, 2015; Brown et al., 2008), protoword reconstruction offers important pieces of information concerning the geographical and chronological dimensions of ancient communities (Heggarty, 2015; Mallory and Adams, 2006), at the same time, allowing an insight into the cognitive and cultural world of our ancestors. The traditional process of reconstructing ancient languages consists of the "comparative grammar-reconstruction" method (Chambon, 2007; Buchi and Schweickard, 2014), and the etymological data thus obtained can be used as a source on human prehistory, corroborating the archaeological inventory (Heggarty, 2015), and providing the basis for 'linguistic paleontology' (Epps, 2014). The reconstruction of a word automatically implies a reconstruction of the surrounding realities, both natural and socio-cultural. For example, the presence in different Indo-European languages of obviously related words for 'beech' or 'salmon' allowed the reconstruction of words from Proto-Indo-European and thus information about the elements of nature present in the immediate vicinity of the Indo-Europeans could be extracted. In the absence of any clear documentary or archaeological data, these lexical clues allowed the geographical identification of the Indo-European homeland, also facilitating the chronology of successive waves of separation of Indo-European languages from the common trunk.

In the case of Romance languages, although the mother tongue - Latin - is attested, its presence in written texts is not an exhaustive source for linguistic, social, and historical analysis of the community that spoke it. It is now generally accepted that the spoken language represented a different diastatic, diaphasic, and diamesic variety from written language, used by the few educated people who decided to express themselves in writing (Wright, 2002). The Latin language that we reconstruct from words inherited in Romance languages is thus the only concrete and reliable living variety of the language from which Romance languages originate, whether we call it oral/ vulgar Latin or Proto-Romance. We will opt here for the name "Proto-Romance" when we refer to the language from which the Romance languages originate, as this corresponds to the concept of protolanguage and protoword (Buchi and Schweickard, 2014).

Furthermore, there are still numerous clearly cognate words present in several Romance languages, whose etymon is not attested in Latin (nor in any other language from which it might have been borrowed). For example, in the case of It. *trovare* 'find', Fr. *trouver*, Cat. *trobar*, etymologists have hotly debated over the decades whether one should reconstruct the protoform *tropare or *turbare (Georgescu and Georgescu, 2020). A series of cognates attested in all Romance geograph-



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6314–6326

ical areas, like Rom. *încă* 'moreover', It. *anche*, Old Fr. *anc*, Cat. *anc* etc., has triggered over 15 etymological hypotheses over the last century, still without a generally accepted solution.

Although etymologists' interest in reconstructing the protolanguages has risen over the years, they still encounter numerous gaps when using exclusively the classical, manual methods (Buchi and Schweickard, 2010, 2020). As the task of protoword reconstruction plays an important role in historical linguistics, studies have gone beyond the comparative method in an attempt to automate the process (Atkinson, 2013; Oakes, 2000; Bouchard-Côté et al., 2013; Ciobanu and Dinu, 2019). However, the task has been recognized as difficult and challenging. Computational protoword reconstruction is a fairly new direction of study, and consequently even state of the art approaches have limitations. Complete automation of the reconstruction process is still a desideratum. Oakes (2000) proposed two systems (Jakarta and Prague) that, combined, cover the steps of the comparative method for protolanguage reconstruction, and several other approaches to reconstruct protowords computationally had been attempted previously (Hewson, 1973; Lowe and Mazaudon, 1994; Kondrak, 2002). The work of computational biologists such as Alexandre Bouchard-Côté, Russell Gray, Robert McMahon, and Mark Pagel, and coauthors took the protoword reconstruction one step further by applying methods from computational biology to the problem of the reconstruction of language history, often in collaboration with linguists (Pagel, 1999; Pagel et al., 2013; Bouchard-Côté et al., 2009; Bouchard-Côté et al., 2013). In recent years, researchers have introduced new methods for protoword reconstruction, based on modern computational techniques (for example, CRF, transformers, RNN, deep learning) (Ciobanu and Dinu, 2018; Sims-Williams, 2018; Meloni et al., 2021; Fourrier, 2022; List et al., 2022; He et al., 2023a; Akavarapu and Bhattacharya, 2023; Kim et al., 2023). The computational methods are limited today by 1) the available data (sparse, inconsistent) and 2) by the insufficiency of linguistic knowledge embedded in the systems.

The latest computational results on Romance protoword reconstruction, in particular, are reported on the database of (Meloni et al., 2021), which contains 8,799 cognates set in Latin, Italian, Spanish, Portuguese, French, and Romanian (not all full cognates set). This is a revision of the dataset of (Dinu and Ciobanu, 2014) (used with very good results in (Ciobanu and Dinu, 2018)) with the addition of cognates scraped from Wiktionary.

Starting with these remarks, our main contributions are:

1. We introduce a comprehensive Romance database for protoword reconstruction by processing RoBoCoP (Dinu et al., 2023), the largest Romance cognate-borrowing database obtained from electronic dictionaries with etymological information of Romanian, Italian, Spanish, Portuguese, and French.

2. We propose a strong benchmark for automatic protoword reconstruction, by applying a set of machine learning models (using various feature sets and architectures) on any cognate set of Romance languages.

The rest of the paper is organized as follows: In Section 2 we present the database that we have created and offer details about the processing steps involved; in Section 3 we introduce our approach for the automatic protoword reconstruction, along with methodological details; the results of our proposed experiments are fleshed out in Section 4; and a comprehensive error analysis is described in Section 5. The last section is dedicated to final remarks.

## 2 Data

A major inconvenience in Historical Linguistics in general, and in computational approaches of protoword reconstruction in particular is the scarcity of available data. Nonetheless, in the last few years, several initiatives have been undertaken in this direction. (Ciobanu and Dinu, 2018) developed a database of Latin protowords, further expanded by (Meloni et al., 2021) with Wiktionary data. Recently, this dataset was extensively used for several studies (Kim et al., 2023; He et al., 2023b; Akavarapu and Bhattacharya, 2023). In 2023, Dinu et al. (2023) published the most comprehensive database of Romance related words, named RoBoCop. It contains cognates and etymons in five Romance languages: Italian, Spanish, Portuguese, Romanian, and French. It has already been used with good results on prominent historical linguistic tasks such as cognate identification (Dinu et al., 2023), cognate-borrowings discrimination (Dinu et al., 2024b), and determining the borrowing di-

| | RO | ES | PT | IT | FR |
|---|---|---|---|---|---|
| **axis** | axă | eje | áxis | asse | ais |
| | axis | axis | áxis | asse | ais |
| | ax | axis | áxis | asse | ais |
| | axis | eje | áxis | asse | ais |
| | axă | axis | áxis | asse | ais |
| | ax | eje | áxis | asse | ais |

Table 1: All cognate tuples present in the ProtoRom dataset for the Latin etymon *axis*.

rection (Dinu et al., 2024a).

### 2.1 The ProtoRom Database

Starting with the RoBoCoP database (Dinu et al., 2023), in order to obtain cognate sets with common etymons in the five Romance languages, we filtered out the words with Latin etymology. We then created maximal tuples of words in the Romance languages with the same etymon $(< w_{L_i} >, e)$, where $L_i$ are all the languages among the five where the etymon $e$ engendered a word, and $w_{L_i}$ are the corresponding words in each of the languages discussed. In cases where multiple words in $L_i$ derive from the same etymon $e$, we created multiple tuples $(< w_{L_i} >, e)$ with all possible combinations of cognate words $< w_{L_i} >$ and the same etymon $e$. For an example of such a case see Table 1.

We curated the obtained data, with the help of linguists. In the process, we discarded sets that contained irrelevant or erroneous information, e.g.: erroneous lexical forms (e.g. Lat. *videre* 'see' - It. *vedere* - Fr. *voir* - Ro. ***videa*** (correct: *vedea*); included a verb form in any mood other than the infinitive (e.g. Lat. *videre* - Sp. ***veas*** (subjunctive) / ***viendo*** (gerundive) / etc.); retained the reflexive form of a verb (e.g. Lat. *ponere* 'put' - It. *porre* - Sp. ***ponerse*** (poner + reflexive pronoun *se*), etc.); or contained words derived on Romance ground (e.g. Lat. *dens* 'tooth' - It. *dente* - Ro. ***dințos*** (= *dinte* + suff.-*os*), etc.).

We were able to apply manual corrections for all these errors for the smaller subset of entries in the database that have a cognate in each of the five languages. For the rest of the full database ProtoRom, we applied a semi-automatic correction by lemmatizing the cognate words, using the default lemmatizers[1] implemented in the *spaCy*[2] library for each of the Romance languages. In all experiments described in the rest of the paper, we use the lemmas of the cognates instead of the original forms found in the dictionary.

In addition to the correct series thus retained, we integrated the database created by Reinheimer-Rîpeanu (2001), a high quality collection of cognate series manually selected from the etymological dictionaries of each Romance language, some of which still not digitized (which probably explains why certain cognate sets from this collection were not among ones in the RoBoCoP database). We thus obtained a new database of cognate sets.

The proposed database contains a total 39,973 full or partial cognate sets along with their etymons. For the experiments in this paper, we focus on the 19,222 entries with at least 2 cognates. We choose this subset in order to ensure the robustness of our experiments, focusing on Latin etymons that engendered at least two cognates in two different languages, and we ignore the entries with only one cognate for a given etymon. Going further, this restricted dataset will be referred to as ProtoRom[3]. A cognate set is composed of a tuple of words in different languages with a common etymon, where the tuple can be either a full set of 5 cognates or a partial set of 2 to 4 cognates, where the cognate in one or more of the languages is missing (the Latin etymon did not produce an attested word in these languages according to our sources).

There are 1,245 full cognate sets in the database, the rest being partial cognate sets. To facilitate distinguish- ing between the two settings, we name the first one ProtoRom-all5, and the second one ProtoRom. When we leave out one of the languages, we can obtain more full sets of 4-tuples (sets with at least 4 cognates) as follows: 1,480 if we leave out Italian, 2,493 if we leave out French, 1,489 when we leave out Portuguese, 1,504 when we leave out Spanish, and 1,946 by leaving out Romanian. The statistics detailing the number of partial cognate sets in all combinations are shown in Table 2.

ProtoRom is the largest database of cognate sets for Romance languages so far, significantly exceeding the widely used database for this task (Meloni et al., 2021), containing 8,799 cognate sets of Romanian, French, Italian, Spanish, Portuguese words and the corresponding Latin form (which, in turn, is an extension of Ciobanu and Dinu (2018)'s orig-

[1]using the models, for each language L

[2]https://spacy.io/usage/models

[3]The dataset is available for research purposes upon request at: https://nlp.unibuc.ro/resources.html#protorom

inal dataset of 3,218 cognate sets, by adding data from Wiktionary).

## 3 Methodology and Experiments

### 3.1 Experimental Setting

For our experimental trials, we consider two settings: In the first one, we limit our dataset to only the full cognate sets (i.e. 5-tuples of cognates from each of the five languages, that originate from the same Latin etymon), while in the second one we consider all cognate sets (with at least two cognates from different languages, per etymon, as previously mentioned). The second setting uses the full breadth of our proposed dataset (ProtoRom-all5), whereas the first one is a strict subset (ProtoRom).

**Data splitting.** In order to train and validate our models, we split our datasets into 80% : 10% : 10% train-dev-test subsets. Because of the nature of the cognate sets, generating a language-level stratified split is a non-trivial task. Since a Latin etymon can produce more than one reflex in a given language, we end up with $\prod_i \max(1, n_{L_i})$ cognate sets for a given etymon, where $n_{L_i}$ is the number of reflexes generated by that etymon in language $L_i$.

We propose a random split methodology that achieves the following properties: A Latin etymon and all of its cognate sets are not allowed to be part of more than one split; the raw number of cognate sets (i.e. entries in the dataset) follows the 80 : 10 : 10 distribution; the distribution of unique Latin etymons is also 80 : 10 : 10; for each of the five languages; and computing the distribution of unique reflexes in that language yields the same ratio across the splits. In other words, if we only keep the Latin etymons and their reflexes in only one language, we obtain a monolingual task with the same 80 : 10 : 10 split.

In order to perform these splits, we construct for each Latin etymon a 5-dimensional vector $(n_{L_i})_i$, using the previous definition of $n_{L_i}$. In order to obtain a split of ratio $0 < p < 1$, we want to select such vectors that, when summed together, equal $p \cdot (N_{L_i})_i$, where $N_{L_i}$ is the total number of unique reflexes from language $L_i$. In other words, we face a task equivalent to a five-dimensional knapsack problem, which is not feasible given the large total capacities. Considering that these vectors contain particularly small values, and are somewhat uniformly distributed, plus the large capacities that we have to fill, we are able to randomly select etymons and their associated cognate sets and add them to any of the three splits, as long as they fit. This approach yields the original split distribution with some small deviations ($< 1\%$).

Also note that after splitting the ProtoRom-all5 dataset, containing only the full cognate sets, we can use it as a starting point for splitting the rest of the ProtoRom dataset, thus ensuring that no training examples from one setting leaks into the validation of the other one.

**Features.** The proposed approaches can be split into two main categories: models for reconstructing the orthographical representation of the protowords using the orthographical form of modern cognates, and models that reconstruct the phonemic representation from phonetic transcriptions of modern cognates. Our extracted dataset essentially provides the necessary examples for the former, while for the latter we employ the *eSpeak*[4] library to automatically generate the phonemic representations.

### 3.2 Models

We use a variety of machine learning models, including classical, neural, and transformer-based (pretrained and trained from scratch for the task). We include methods used in previous papers on the topic and evaluate them on our larger dataset in order to provide a benchmark for the task of protoword reconstruction for Romance languages.

We experiment with a variety of models, including pre-trained large language models (LLMs) and current state-of-the-art models for protoword reconstruction with various architectures (probabilistic RNN, character-level transformer) adapted to our new database, as well as original solutions. In this way, we aim to provide a benchmark for the task of protoword reconstruction.

**CRF + reranking** We used an approach that relies on conditional random fields (CRFs), based on the method proposed by Ciobanu and Dinu (2018). Firstly, we applied a sequence labeling method that produces the form of the Latin ancestors, for each modern language. The modern words are the sequences, and their characters are the tokens. We used character n-grams from the input words as features. We employed pairwise sequence alignment (Needleman and Wunsch, 1970) between modern words and protowords to obtain

[4]https://github.com/espeak-ng/espeak-ng

| | | | | | |
|---|---|---|---|---|---|
| It: 5,197 | It-Fr: 2,807 | It-Ro: 3,439 | It-Es: 6,820 | It-Pt: 4,605 | -It: 1,480 |
| It-Fr-Ro: 1,842 | Fr: 4,992 | Fr-Ro: 3,898 | Fr-Es: 4,413 | Fr-Pt: 2,797 | -Fr: 2,493 |
| It-Fr-Es: 2,270 | It-Ro-Pt: 2,926 | Ro: 5,685 | Ro-Es: 5,117 | Ro-Pt: 3,394 | -Ro: 1,946 |
| It-Fr-Pt: 2,390 | It-Es-Pt: 3,988 | Fr-Ro-Pt: 1,782 | Es: 6,820 | Es-Pt: 4,543 | -Es: 1,504 |
| It-Ro-Es: 2,913 | Fr-Ro-Es: 3,503 | Fr-Es-Pt: 2,311 | Ro-Es-Pt: 2,919 | Pt: 5,202 | -Pt: 1,489 |

Table 2: Number of cognate sets that are descendants from the same Latin word, for each language combination. *x-y* means the number of cognate sets for languages *x* and *y*; *x-y-z* means the number of cognate sets for languages *x*,*y*, and *z*; *x* means how many descendants are from Latin for language *x*; *-x* means the number of cognate sets for all languages except *x*.

the labels for each token. Secondly, we defined several ensemble methods to take advantage of the information provided by all languages, in order to improve performance. We employed fusion methods based on the ranks in the n-best lists and the probability estimates provided by the individual classifiers for each possible production, in order to combine the outputs of the classifiers (n-best list of possible protowords) and to leverage information from all modern languages. For each word in the productions list, we multiply the rank of it with the confidence score given by the CRF model for each language; we sum up the multiplication scores for each word in the list and then rerank the productions based on these results.

**Probabilistic LSTM** We conducted experiments using a combination of recurrent neural networks with different dynamic programs and expectation-maximization techniques, as described in He et al. 2023b. The overall system can be split in two stages: a) a modelling stage, where we model the evolution of words by making small character-level edits to the ancestral form; for each language in the study, the distribution over newly created words is computed; b) an expectation-maximization stage, where the ancestral form is inferred; using words sampled from the posterior distribution, the expected edit count is computed and further used by the character-level recurrent neural network in order to optimize the next round of samples; the final reconstruction is the maximum likelihood word forms. This model requires a full tuple of cognates to be passed as input, so we only compute results for experiments on the ProtoRom-all5 set. Like the original authors, we only apply this model on the phonemic forms of words, since the probability distributions of edit operations used in the algorithm rely on a set of manually set features for each phoneme that are not similarly available for orthographical characters.

**Character-level transformer** The next experiments conducted in this research are based on the transformer model, proposed by Kim et al. 2023. Some critical changes in the architecture were made in order to be able to accept our samples format: multiple modern word sequences (one for each language) correspond to a single protoform sequence. A positional encoding is applied to each individual modern word sequence before concatenation. An additive language embedding is applied to the token embeddings alongside the positional encoding in order to make a difference between input tokens of different languages.

**Pre-trained LLM (Flan-T5)** We finally evaluate the capabilities of pretrained Large Language Models (LLMs) to solve our task. While LLMs are currently obtaining state-of-the-art performance across NLP tasks, our specific goal is unlike usual tasks included in benchmarks or in training data for LLMs, and it is strongly multilingual (including one dead language), so we suspect it might be a difficult task for an LLM. We choose to use a pretrained model and fine-tune it on our own training data in order to increase its chances to perform well. We use a "base" variant of the Flan-T5 model (Chung et al., 2024), and fine-tune the model using instructions including the prompt: "What is the etymon given the following cognates:", followed by a list of cognate and language pairs formatted as "$< L_i >: < w_i >$" and separated by new lines, where the list of cognate words $w_i$ is their respective languages $L_i$ can be arbitrarily long (from 2 to 5 cognates, in the case of our experiments). For evaluation, we attempt to generate multiple output sequences, which are used as a ranking for the etymon prediction.

One limitation of pretrained LLMs that we cannot overcome through fine-tuning is its alphabet, which contains mostly characters in the Latin graphical alphabet, which means that we can only

use this model with othographical features. Using phonemic features would require retraining the model from scratch and we would lose the benefit of pertaining which is usually the strong point of LLMs.

## 4 Results

The previously described methods have been applied on both ProtoRom and ProtoRom-all5 datasets, using the orthographical form of the cognates and Latin etymon, or alternatively the autogenerated phonemic representations (where the models were able to accommodate them).

We also provide a comprehensive human evaluation of the results. Linguists from our team manually analyzed the entire list of results, and we present the most significant observations regarding the models' successes and failures. The linguists did not correct the protoforms proposed by the models, but only evaluated and commented on them in relation to current knowledge in the field of historical linguistics.

The metrics used include accuracy, (normalized) edit distance, and $Cov_i$, with $i \in \{1, 5, 10\}$, which stands for an extended version of the accuracy metric, where a correct prediction is one where the model found the correct etymon within the first $i$ etymons predicted by our method (this metric is computed for models that are able to output a ranked list of predictions - Flan-T5 and CRF-based models).

### 4.1 ProtoRom-all5 Results

Results obtained on the ProtoRom-all5 set are shown in Table 3. In terms of accuracy (or $Cov_1$), the best results are obtained using the orthographical forms, with the CRF-rerank model, reaching 60.4%. From the perspective of the $Cov_i$ metrics, it is remarkable that the CRF-rerank model obtains a $Cov_{10}$ score above 82%.

The experiments using the phonemic forms produce weaker results, with the best accuracy reaching 55.8% in the top 1 predictions scenario. Nevertheless, the CRF approach is able to achieve an accurracy close to 80% when we consider the top 10 best ranked predictions.

The probabilistic RNN models achieve very poor performances, reaching a mean edit distance of 3.11 when trained on the phonemic representations.

### 4.2 ProtoRom Results

The best accuracy when training the orthographical models is achieved in this scenario by the Transformer model, closely surpassing 73% (Table 4). As for the $Cov_i$ metrics, the Flan model remarkably obtains a $Cov_{10}$ accuracy score of 85.4%, and an edit distance of 0.23.

Similarly to the previous scenario, the experiments using the phonemic forms produce weaker results, with the best accuracy reaching 66.8% via the Transformer model. These results represent a collection of baselines for protoword reconstruction using our proposed dataset configurations.

We believe the higher accuracy observed on the full dataset is simply due to the larger amount of available data. While ProtoRom-all5 is a subset that contains only complete cognate sets from each of the five studied languages (totaling 1,245 sets) the ProtoRom dataset includes sets of two, three, or four cognates, resulting in significantly more sets (19,222). This larger dataset allowed the models to learn more phonetic correspondences, thereby improving the reconstruction process. Even though they are not full sets of five cognates, the additional cognate sets in the full database seem to help the models learn more about their protowords. This learning process is closely similar to the human method of learning: with more examples, linguists can be more certain of particular correspondences or phonetic changes and can apply them in the reconstruction with much greater confidence.

## 5 Error analysis

This section is dedicated to a deeper dive into qualitatively quantifying the errors produced by the previously proposed models. Our objective is separating purely wrong predictions from "near misses", which may still provide value for linguists for the reasons discussed below.

The error analysis was manually conducted by the linguists from our team, who specialize in Romance languages. They did not modify the protoforms provided by the models in any way. Their only intervention was to distinguish forms that were genuinely erroneous from those whose differences from the dictionary form were either insignificant or represented a correct adjustment to the reality of Latin pronunciation. In the final quantitative analysis, forms in this category were therefore included in the list of correct predictions without any changes to their structure.

| | | Accuracy | | | Edit/NEdit | | |
|---|---|---|---|---|---|---|---|
| | | $Cov_1$ | $Cov_5$ | $Cov_{10}$ | $Cov_1$ | $Cov_5$ | $Cov_{10}$ |
| Gr | Flan | 55.0 | 70.5 | 75.9 | 1.03/0.15 | 0.55/0.08 | 0.43/0.06 |
| | CRF | **60.4** | **78.2** | **82.1** | 0.80/0.12 | **0.38/0.05** | **0.31/0.04** |
| | Transformer | 59.92 | — | — | **0.72/0.11** | — | — |
| Ph | CRF | 55.8 | **75.9** | **79.8** | 0.86/0.13 | 0.4/0.06 | 0.33/0.05 |
| | Transformer | **47** | — | — | 0.98/0.16 | — | — |

Table 3: Reported results for protoword reconstruction on the ProtoRom-all5 dataset via orthographical representations (Gr) and via phonemic representations (Ph), respectively. We report the reconstruction accuracy along with the mean edit distance (Edit) and mean normalized edit distance (NEdit). The $Cov_i$ values for the edit distances are computed by selecting the minimum distance between the true etymon and the top $i$ predictions, then averaging over these minima for all of the test examples. For the Flan and CRF models, we look at the top 1, 5, and 10 predictions when computing these metrics.

| | | Accuracy | | | Edit/NEdit | | |
|---|---|---|---|---|---|---|---|
| | | $Cov_1$ | $Cov_5$ | $Cov_{10}$ | $Cov_1$ | $Cov_5$ | $Cov_{10}$ |
| Gr | Flan | 65.5 | **81.7** | **85.4** | 0.73/0.09 | **0.30/0.04** | **0.23/0.03** |
| | CRF | 55.0 | 71.3 | 79.1 | 1.06/0.16 | 0.55/0.08 | 0.42/0.06 |
| | Transformer | **73.1** | — | — | **0.51/0.08** | — | — |
| Ph | Transformer | **66.8** | — | — | **0.67/0.10** | — | — |

Table 4: Similar to Table 3 we report the same evaluations when using the complete ProtoRom dataset.

Through analyzing the errors, we have identified some patterns that typically reflect either an insufficient number of examples to support a particular phonetic change or the irregularity of the change itself. For example, the short tonic $/u/$ develops into Spanish $/o/$ in half of the cases, while it remains $/u/$ in the other half. In such scenarios, the model may not know which phonetic treatment the cognates underwent and might choose the wrong variant. Similarly, in cases of phonetic accidents, which are by nature irregular and unpredictable, the model cannot reconstruct the pre-accident form. Instead, it reconstructs the intermediate form between the classical word and its Romance descendants. Identifying and systematizing these errors can help improve future results by broadening the input with information related to sound changes.

Before analysing the errors, a few preliminary points should be made. Romance lexicography as a whole is graphocentric - it considers the written, classical Latin (CL) lexical variants as the basis for the Romance vocabulary, even though it goes without saying that vernacular languages, oral par excellence, developed from an oral language, in our case Proto-Romance (PR) (Chambon, 2007). In the latest methodology used in Romance etymology, developed within the DÉRom project (Buchi and Schweickard, 2014), the etymological identification is based strictly on the comparative grammar - reconstruction method, starting from the lexical forms that were used uninterruptedly in Romance languages. The lexemes attested in Classical Latin are only a written correlate, possibly further evidence of the existence of the form obtained by the methods of comparative historical linguistics.

In the light of these considerations, we find that some of the reconstructed variants classified as errors should actually be considered as positive results and evidence that the machine could work at the same level as a linguist applying traditional methods. By positive results instead of errors we mean cases - not a few - where the machine reconstructed exactly the phonetic form valid for oral Latin, at the expenses of the standard orthographical form as it is lemmatized in classical Latin dictionaries.

Cases where the word obtained and the one given by the dictionary did not completely match were automatically considered as errors, although sometimes it was not a mistake as such. Therefore, there are a number of protoforms which, although they appear in the list as inadvertences, are variants that should be taken into account with full attention by linguists. Some are no more wrong than the form in the dictionary, some are closer to the actual oral form than those provided by lexicographers, while some are exactly the form that historical linguists would have reconstructed using traditional meth-

ods based on the sound laws of each language (we discuss each case below). Therefore, protoforms obtained by the automatic methods proposed here are sometimes preferable to the lemmatized ones, and this is the most important thing we can expect from the machine.

Below, we provide a list of situations categorized as errors, but where the the automatic protoword reconstruction is either comparable or better than the version proposed by the dictionary, as it represents exactly the linguistic variant we should consider as intermediate between classical Latin and Romance languages.

- Protowords ending in *-um* instead of standard *-us* (*lupum* instead of *lupus*). The difference between the endings *-us* / *-um* did not properly exist in Proto-Romance, as the final consonant *-s*/*-m* was no longer pronounced. Thus, if the etymological dictionaries provide the classical nominative form *lupus* as an etymon for Ro. *lup*, It. *lupo*, Fr. *loup* etc., but the computer reconstructs *lupum* – this latter variant is more correct from a grammatical point of view, since in general nouns are inherited from the accusative form (in our case ending in *-um*) and not from the nominative (ending in *-us*). Moreover, if it reconstructs *lupu*, this form is even more correct, being the real one, that reflects the pronunciation in the spoken language.
- The automatically reconstructed protoforms reflect phonetic features specific to Proto-Romance: monophthongation (au > o, e.g. CL *auca* vs PR *oca*; œ > e, e.g. *pœna* vs *pena*; æ > e, e.g. *hæsitare* vs *esitare*); reduction of geminate consonants (*addictus* vs *adictum*); loss of the initial or intervocalic /h/ (*hæsitare* vs *esitare*; *cohærente* vs *coerente*); phonetic adaptation of Greek loanwords to the Latin pronunciation (y > i, e.g. CL *byzantinus* vs PR *bizantinus*, the aspirate consonants become oclusive, th > t (CL *citharoedu* vs PR *citaredu*), ph > f (CL *phalange* vs PR *falange*); assimilations (CL *admonere* vs PR *ammonire*); simplification of consonant clusters (CL *sculptore* vs PR *scultore*, *temptare* vs *tentare*, *unctura* vs *untura*); changes in the pronunciation of vowels (CL *guttu* vs PR *gotu*, *misculare* vs *mescolare*, *siccare* vs *sec(c)are*, *occidere* vs *ucidere*, *calcea* vs *calcia*).
- Certain reconstructed etyma retain accidental phonetic changes that must be presupposed for a particular geolinguistic area (Sp. *queso*, Pt. *queixo* imply the metathesis PR *caesu* instead of CL *caseu*, Ro *plop*, It. *pioppo*, Sp. *chopo* lead to the protoform with metathesis *plopu*, correctly identified by the machine, instead of CL *populus*), or for the global PR variety (Ro. *doamnă*, It. *donna*, Sp. *doña*, lead to the syncopated protoform *domna*, reconstructed by the machine, instead of CL *domina*, registered in lexicography).
- The automatically reconstructed protoforms may mirror morphologic changes that underlie the subsequent Romance developments: nouns of the 5th declension undergo a shift to the 1st declension (CL *canities* vs PR *canitia*, *species* vs *specia*); verbs shifting from middle-passive to the active voice (CL *renasci* vs PR *renascere*).
- The computer has reconstructed the oblique case forms representing the basis from which the Romance nouns were inherited (nominative *flos* vs oblique case *flore-* > Ro. *floare*, It. *fiore*, Fr. *fleur*, etc.; *civitas* vs *civitate* > Ro. *cetate*, Sp. *ciudad*, etc.), or the plural instead of the singular form, when the Romance lexemes descend from the former (sg. *capitium* vs pl. *capitia* > Sp. *cabeza*, Pt. *cabeça*).

The real errors in the experiments we developed stem primarily from lexicographic omissions or mistakes, as well as in the imprecise methodology employed by the Ibero-Romance dictionaries consulted, namely the lack of any distinction between inherited and borrowed Latin words (Buchi and Dworkin, 2019). This latter inaccuracy leads to a misinterpretation of the phonetic correspondences by the computer, given that only the inherited words, not the borrowed ones, underwent regular sound change. Therefore, if we put together Ro. *roată*, Sp. *rueda*, Pt. *roda*, with Ro. *rotație*, Sp. *rotacion*, Pt. *rotação*, the computer will not be able to correctly infer the correspondence $t/d/d$ and will confuse it with $t/t/t$, also assuming the series $d/d/d$. Therefore, some reconstructions, especially in the case of words circumscribed only to Ibero-Romance languages, could not take this sound law into account (e.g., on the basis of Sp. *miedo*, Pt. *medo*, the computer could not reconstruct *metus*, but proposed *medus*, which is wrong). This kind of shortcomings will be easily overcome in the

future, firstly by clearly establishing, in the ProtoRom database, the inheritance-borrowing distinction, and secondly by extending the input provided to the computer with a number of basic phonetic laws.

**Revised performance scores.** Looking at the best reported predictions, we can apply the linguistic observations stated in the previous section and count which wrong predictions can be actually considered acceptable errors. Thus using these recovered predictions, the best models' scores would change as follows:

- the orthographical Transformer accuracy for the ProtoRom dataset increases from 73.1% to 82.7% (135 out of the 575 original errors were recovered).
- the Flan model's $Cov_{10}$ accuracy on ProtoRom increases from 85.4% to 89.6% (90 out of the 311 original errors were recovered).
- the $Cov_{10}$ accuracy for the orthographical CRF model trained on ProtoRom-all5 increases from 82.1% to 90.7% (11 out of the 23 original errors were recovered).

## 6 Conclusion

In this paper, we built a new dataset for automatic protoword reconstruction, consisting of 19,222 cognate sets from five Romance languages (Romanian, Italian, Spanish, Portuguese, French). This is to date the largest database of its kind, surpassing its predecessor which totals 8,799 cognate sets.

We also proposed a series of comprehensive benchmarks ranging from deep-learning approaches, using LLMs and Transformer-based architectures, to more classical algorithms such as CRFs, some of which achieved performances of more than 85% accuracy when allowing multiple generated reconstructions.

An in-depth linguistic analysis of the erroneous reconstructions was also performed using the predictions of the best performing models. This attempt shed some light on the various categories of mistakes, out of which several could be considered acceptable. When ignoring the aforementioned acceptable errors, we were able to surpass 90% accuracies. We consider this an important distinction, since in our view similar tools should aim at assisting linguists in their scientific endeavours. Raw metrics are useful to compare computational methods, but, in order to assess their usability, a more qualitative inspection of the results should be performed. We hope through our research to incentivize further analysis.

As for future work, we are looking into an additional refinement of the current cognate sets, but also extending the database with more examples, including properly validated monolingual Latin reflexes that were excluded from our experiments for robustness sake. We also intend to expand past the proposed benchmarks with more novel approaches, relying on both the proposed dataset and the additional contents of its parent database, RoBoCoP.

## Limitations

One limitation of the current work stems from the automatic generation of the phonetic representations via a third-party library (eSpeak). Although this approach was employed successfully in previous studies, the quality of the generated phonemes has a higher variance when comparing high-resourced languages to lower-resourced ones (such as Romanian, or even Latin).

Also, in this study we used the generated phonetic forms without any extra preprocessing steps, in order to have a representation of the pronunciation that is as accurate as possible. Removing phonetic markers (such as stress markers) from these representations may turn the generation task into a somewhat easier one, since currently the phonetic models are tasked with predicting the stressed sounds too.

In terms of resources, existing LLMs are mostly targeting orthographical texts, making any reasonable attempt at generating phonetic ones very difficult.

## Ethics Statement

There are no ethical issues that could result from the publication of our work. Our experiments comply with all license agreements of the data sources used. We make the contents of our package available for research purposes upon request.

## Acknowledgements

This work was supported by a mobility project of the Romanian Ministery of Research, Innovation and Digitization, CNCS - UEFISCDI, project number PN-IV-P2-2.2-MC-2024-0461, within PNCDI IV.

We want to thank the reviewers for their useful suggestions and Diana Grigore, Cosmin Petrescu, Ioana Pintilie for their help in developing the algorithms.

## References

V. S. D. S. Mahesh Akavarapu and Arnab Bhattacharya. 2023. Cognate transformer for automated phonological reconstruction and cognate reflex prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6852–6862. Association for Computational Linguistics.

Alexander V. Alekseyenko, Quentin D. Atkinson, Remco Bouckaert, Alexei J. Drummond, Michael Dunn, Russell D. Gray, Simon J. Greenhill, Philippe Lemey, and Marc A. Suchard. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337:957–960.

Quentin Atkinson, Geoff Nicholls, David Welch, and Russell Gray. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103(2):193–219.

Quentin D Atkinson. 2013. The descent of words. *Proceedings of the National Academy of Sciences*, 110(11):4159–4160.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2009)*, pages 65–73, Boulder, Colorado. Association for Computational Linguistics.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc. Natl. Acad. Sci. USA*, 110(11):4224–4229.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308.

Éva Buchi and Steven N Dworkin. 2019. Etymology in romance. *Oxford Research Encyclopedia of Linguistics*.

Éva Buchi and Wolfgang Schweickard. 2010. À la recherche du protoroman: objectifs et méthodes du futur dictionnaire étymologique roman (dérom). In *XXVe CILPR Congrès International de Linguistique et de Philologie Romanes: Innsbruck, 3–8 septembre 2007*, pages 6–61. de Gruyter.

Éva Buchi and Wolfgang Schweickard. 2014. *Dictionnaire étymologique roman (DÉRom): génèse, méthodes et résultats*, volume 381. Walter de Gruyter GmbH & Co KG.

Éva Buchi and Wolfgang Schweickard. 2020. *Dictionnaire étymologique roman (DÉRom) 3: entre idioroman et protoroman*, volume 443. Walter de Gruyter GmbH & Co KG.

Lyle Campbell. 2013. *Historical Linguistics*. Edinburgh University Press.

Jean-Pierre Chambon. 2007. Remarques sur la grammaire comparée-reconstruction en linguistique romane (situation, perspectives). *Mémoires de la Société de linguistique de Paris*, 15:57–72.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alina Maria Ciobanu and Liviu P. Dinu. 2018. Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2019. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Liviu P. Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1038–1043. European Language Resources Association (ELRA).

Liviu P. Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive Romance borrowing cognate package and benchmark for multilingual cognate identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7610–7629. Association for Computational Linguistics.

Liviu P. Dinu, Ana Sabina Uban, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2024a. It takes two to borrow: a donor and a recipient. who's who? In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6023–6035. Association for Computational Linguistics.

Liviu P. Dinu, Ana Sabina Uban, Ioan-Bogdan Iordache, Alina Maria Cristea, Simona Georgescu, and Laurentiu Zoicas. 2024b. Pater incertus? there is a solution: Automatic discrimination between cognates and borrowings for Romance languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12657–12667, Torino, Italia. ELRA and ICCL.

Michael Dunn. 2015. Language phylogenies. *The Routledge handbook of historical linguistics*, pages 190–211.

Patience Epps. 2014. Historical linguistics and sociocultural reconstruction. In *The Routledge Handbook of Historical Linguistics*, pages 579–597. London: Routledge.

Clémentine Fourrier. 2022. *Neural Approaches to Historical Word Reconstruction. (Approches Neuronales pour la Reconstruction de Mots Historiques)*. Ph.D. thesis, PSL University, France.

Simona Georgescu and Theodor Georgescu. 2020. Fr.«trouver», occ.«trobar» etc.: un dossier étymologique ouvert à nouveau. *Revue de linguistique romane*, 84(1):83–98.

Andre He, Nicholas Tomlin, and Dan Klein. 2023a. Neural unsupervised reconstruction of protolanguage word forms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1636–1649. Association for Computational Linguistics.

Andre He, Nicholas Tomlin, and Dan Klein. 2023b. Neural unsupervised reconstruction of protolanguage word forms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1636–1649, Toronto, Canada. Association for Computational Linguistics.

Paul Heggarty. 2015. Prehistory through language and archaeology. In *The Routledge Handbook of Historical Linguistics*, pages 598–626. Routledge.

John Hewson. 1973. Reconstructing prehistoric languages on the computer: The triumph of the electronic neogrammarian. In *COLING 1973 Volume 1: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed protoform reconstruction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–38, Toronto, Canada. Association for Computational Linguistics.

Grzegorz Kondrak. 2002. Algorithms for language reconstruction, phd thesis, university of toronto.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022. A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In *3rd International Workshop on Computational Approaches to Historical Language Change 2022*, pages 89–96. Association for Computational Linguistics (ACL).

John B Lowe and Martine Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics*, 20(3):381–417.

James P Mallory and Douglas Q Adams. 2006. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. Oxford University Press on Demand.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4460–4473. Association for Computational Linguistics.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Michael P Oakes. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–243.

Mark Pagel. 1999. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Mark Pagel, Quentin D Atkinson, Andreea S. Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.

Sanda Reinheimer-Rîpeanu. 2001. Lingvistica romanică. lexic, morfologie, fonetică. *Bucureşti: All.*

Patrick Sims-Williams. 2018. Mechanising historical phonology. *Transactions of the Philological Society*, 116(3):555–573.

Roger Wright. 2002. *A Sociophilological Study of Late Latin*. Brepols Publishers.

# A Appendix

## A.1 Hyperparameters and infrastructure

### A.1.1 Conditional Random Fields

The implementation of the CRF models follows the description provided by Ciobanu and Dinu 2018.

The CRF algorithm relies on the Mallet library implementation, version 2.0.8[5].

The only training hyperparameters that were tuned are:

- the window size $w \in \{1, 2, 3, 4, 5\}$
- the number of CRF training iterations $i \in \{25, 50, 100\}$

For the orthographical and phonetic training scenarios, the hyperparameters were selected by training on the ProtoRom-all5 training split, and evaluating on the dev one. Because of the long training time on the complete ProtoRom dataset, we ended up reusing the same hyperparameters found during the previous step.

The selected hyperparameters are as follows:

- for the orthographical CRF:
  - Spanish: $w = 1, i = 100$
  - French: $w = 4, i = 100$
  - Italian: $w = 2, i = 100$
  - Portuguese: $w = 1, i = 100$
  - Romanian: $w = 1, i = 100$
- for phonetic CRF:
  - Spanish: $w = 1, i = 100$
  - French: $w = 4, i = 100$
  - Italian: $w = 3, i = 100$
  - Portuguese: $w = 1, i = 100$
  - Romanian: $w = 4, i = 100$

The training was performed on a Ryzen 5 3600X 4GHz CPU, parallelized on 8 threads, the total training time being:

- orthographical CRF for ProtoRom-all5 (including grid search): 15 hours
- phonetic CRF for ProtoRom-all5 (including grid search): 22 hours
- orthographical CRF for ProtoRom: 102 hours

#### A.1.2 Probabilistic LSTM

We conducted experiments with the same architecture used in He et al. 2023b (*GitHub repository*[6]) and following hyperparameters:

- lstm input size: 64
- lstm hidden size: 64
- context window: 10
- number of epochs: 30

For the training we used the following configuration:

- number of rounds: 8
- learning rate: 0.01
- optimizer: Adam
- weight decay: 0.01

All the training was done on an Apple M2 Pro chip and the total training time was 2 hours.

#### A.1.3 Transformer model

The architecture we used in our experiments is the same as Kim et al. 2023 (*GitHub repository*[7]). The hyperparameters used for both orthographical and phonetic experiments are as follows:

- embedding size: 128
- number of encoder layers: 3
- number of decoder layers: 3
- number of attention heads: 8
- feed forward layer size: 128
- dropout: 0.202

Training hyperparameters used for both orthographical and phonetic experiments:

- number of epochs: 200
- batch size: 1
- learning rate: 0.00013
- loss: cross entropy loss
- optimizer: Adam
- scheduler: polynomial decay scheduler with warmup
- warmup epochs: 50
- weight decay: 0

In terms of trainable parameters:

[5] https://mimno.github.io/Mallet/

[6] https://github.com/AndreHe02/historical_release/tree/master

[7] https://github.com/cmu-llab/acl-2023/tree/main

- orthographical experiments: $\approx 817,869$ parameters
- phonetic experiments: $\approx 854,877$ parameters

The training was done using an RTX 2080 Ti GPU. Training time:

- ProtoRom-all5 dataset: 2.5 hours
- ProtoRom dataset: 5 days

#### A.1.4 Flan-T5

Flan-T5 was trained using early stopping based on the $Cov_1$ metric on the validation set.

The configuration used and optimal hyperparameters are as follows:

- batch_size: 50
- epochs: 300,
- learning_rate: 1e-4,
- patience: 3,
- max_seq_len: 64,
- weight_decay: 1e-5,
- warmup_steps: 500,
- lr_scheduler_type: polynomial,
- num_return_sequences: 10,
- num_beams: 10,
- classifier_dropout: 0.0,
- d_ff: 2048,
- d_kv: 64,
- d_model: 768,
- decoder_start_token_id: 0,
- dense_act_fn: gelu_new,
- dropout_rate: 0.1,
- eos_token_id: 1,
- feed_forward_proj: gated-gelu,
- initializer_factor: 1.0,
- is_encoder_decoder: true,
- is_gated_act: true,
- layer_norm_epsilon: $1e-06$,
- max_length: 64,
- model_type: t5,
- n_positions: 512,
- num_beams: 10,
- num_decoder_layers: 12,
- num_heads: 12,
- num_layers: 12,
- num_return_sequences: 10,
- output_past: true,
- pad_token_id: 0,
- relative_attention_max_distance: 128,
- relative_attention_num_buckets: 32,