Ab Antiquo: Neural Proto-language Reconstruction

Carlo Meloni[*1]   **Shauli Ravfogel**[*2,3]   **Yoav Goldberg**[2,3]

[1]Department of Comparative Language Science, University of Zurich
[2]Computer Science Department, Bar Ilan University
[3]Allen Institute for Artificial Intelligence

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

Abstract

Historical linguists have identified regularities in the process of historic sound change. The comparative method utilizes those regularities to reconstruct proto-words based on observed forms in daughter languages. Can this process be efficiently automated? We address the task of proto-word reconstruction, in which the model is exposed to cognates in contemporary daughter languages, and has to predict the proto word in the ancestor language. We provide a novel dataset for this task, encompassing over 8,000 comparative entries, and show that neural sequence models outperform conventional methods applied to this task so far. Error analysis reveals a variability in the ability of neural model to capture different phonological changes, correlating with the complexity of the changes. Analysis of learned embeddings reveals the models learn phonologically meaningful generalizations, corresponding to well-attested phonological shifts documented by historical linguistics.

1 Introduction

Historical linguists seek to identify and explain the various ways in which languages change through time. Research in historical linguistics has revealed that groups of languages (language families) can often be traced into a common, ancestral language, a "proto-language". Large-scale lexical comparison of words across different languages enables linguists to identify *cognates*: words sharing a common proto-word. Comparing cognates makes it possible to identify rules of phonetic historic change, and by back-tracing those rules one can identify the form of the proto-word, which is often not documented. That methodology is called *the comparative method* ([Anttila](#), [1989](#)), and is the main tool used to reconstruct the lexicon and phonology of extinct languages. Inferring the form of proto-words from existing cognates in daughter languages is possible since historical sound changes within a language family are not random. Rather, the phonological change is characterized by *regularities* that are the result of constraints imposed by the human articulatory and cognitive faculties ([Millar](#), [2013](#)). For example, we can find such regular change—commonly called "systematic correspondence"—by looking at the evolution of the first phoneme of Latin's word for "sky":[1]
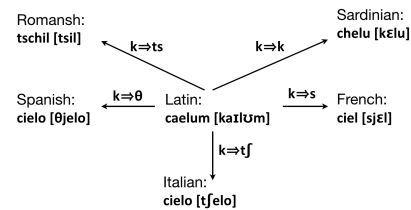

Figure 1: the evolution of Latin word for "sky" is several Romance languages.

The Spanish word's first sound is [θ], while the Italian word begins with [tʃ], the French word with [s], Romansh with [ts] and Sardinian with [k]. This pattern is systematic, and will be found throughout the languages. Working this way, historical linguists reconstruct words in the proto-language from existing cognates in the daughter languages, and determine how words in the proto-language may have sounded.

To what extent can a machine-learning model learn to reconstruct proto-words from examples in this way? And what generalizations of phonetic change will it learn? We focus on the task of proto-word reconstruction: the model is trained on sets of cognates and their known proto-word, and is then tasked with predicting the proto-word for an unseen set of cognates. Our study concentrate on the romance language family[2] and the model is trained to reconstruct the Latin origin. We show

---

[*]Equal contribution

[1]The words as transcribed with International Phonetic Alphabet (IPA) characters.

[2]All the languages that derived from Latin.

that a recurrent neural-network model can learn to perform this task well (outperforming previous attempts).[3]

More interesting than the raw performance numbers are the learned generalizations and error patterns. The Romance languages are widely studied (Ernst (2003); Ledgeway and Maiden (2016); Holtus et al. (1989) among others), and their phonological evolution from Latin is well mapped. The existence of this comprehensive knowledge allows exploring to what extent neural models internalize and capture the documented rules of language change, and where do they deviate from it. We provide an extensive error analysis, relating errors patterns to knowledge in historical linguistics. This is often not possible in common NLP tasks, such as parsing or semantic inference, in which the rules governing linguistic phenomena— or even the suitable framework to describe them— are still in dispute among linguists.

Contributions Inspection of existing datasets of cognates in Romance languages has revealed inherent problems. We thus have collected a new comprehensive dataset for performing the reconstruction task (§4). Besides the dataset, our main contribution is the extensive analysis of what is being captured by the models, both on orthographic and phonetic versions of the dataset (§6). We find that the error patterns are not random, and they correlate with the relative opacity of the historic change. These patterns were divided in different categories, each one motivated by a sound phonological explanation. Moreover, in order to further evaluate the learning of rules of phonetic change, we evaluated models on a synthetic dataset (§6.3), showing that the model is able to correctly capture several phonological change rules. Finally, we analyze the learned inner representations of the model, and show it learns phonologically meaningful properties of phonemes (§6.4) and attributes different importance to different daughter languages (§6.5).

2 Related Work

The related task of cognates detection has been extensively studied. In this task, a set of cognates should be extracted from word lists in different languages. Most effort in Machine learn-

ing approaches to this task has been focused on distance-based methods, which quantify the distance (according to some metric), or the similarity, between a given candidate of cognates. The similarity can be either static (e.g. Levenshtein distance) or learned. Once the metric is established, a classification can be performed either based on hard-decision (words below a certain threshold are considered cognates) or by learning a classifier over the distance measures and other features (Kondrak, 2001; Mann and Yarowsky, 2001; Inkpen et al., 2005; Ciobanu and Dinu, 2014a; List et al., 2016); Mulloni and Pekar (2006) have evaluated an alternative approach, in which explicit rules of transformation are derived based on edit operations. See Rama et al. (2018) for a recent evaluation of the performance of several cognates detection algorithms.

Several studies have gone beyond the stage of cognates extraction, and used resulted list of cognates to reconstruct the lexicon of protolanguages. Most studies in this direction borrowed techniques from computational phylogeny, drawing a parallel between the hypothesized branching of (latent) proto words into their (observed) current forms and the gradual change of genes during evolution. Bouchard-Côté et al. (2007) has applied such a model to the development of the Romance languages, based on a dataset composed of aligned-translations. Bouchard-Côté et al. (2009, 2013) used an extensive dataset of Austronesian languages and their reconstructed proto-languages, and built a parameterized graphical model which models the probability of a phonetic change between a word and its ancestral form; the probability is branch-dependent, allowing for the learning of different trends of change across lineages. While achieving impressive performance, even without necessitating a cognates lists as an input, their model is based on a given phylogeny tree that accurately represents the development of the languages in question.

Wu and Yarowsky (2018) have automatically constructed cognate datasets for several languages, including Romance languages, and used a character-level NMT system to complete missing entries (not necessarily the proto-form). Several works studied the induction of multilingual dictionaries from partial data in related languages. Wu et al. (2020) reconstruct cognates in Austronesian languages (where the proto-language is not attested). Lewis et al. (2020) employ a mixture-

---

[3]We note that the role of the ML model is easier than that of the historical linguist, as it is trained on sets of words that it took the historical linguistics discipline a considerable effort to acquire.

of-experts approach for lexical translation induction, combining neural and probabilistic methods, and Nishimura et al. (2020) translate from a multi-source input that contains partial translations to different languages, concatenated. Finally, Ciobanu and Dinu (2018) have applied a CRF model with alignment to a dataset of Romance cognates, created from automatic alignment of translations (Ciobanu and Dinu, 2014b). The researchers also applied RNNs on the same dataset, but reported negative results.

3 Proto-word Reconstruction

Our proto-word reconstruction is as follows: the training set is composed of pairs (x_i, y_i), where each $x_i = c_i^{\ell_1}, ..., c_i^{\ell_n}$ is a set of cognate words, each tagged with a language ℓ_j, and y_i is the proto-word (Latin word) of that set. We consider an *orthographic task*, where the cognates and proto-words are spelled out as written. As the orthography is often arbitrary and more conservative than spoken language, we consider also a *phonetic task*, in which the cognates and proto-words are represented as their phonetic transcriptions into IPA.

An example of a training instance (x, y) for the orthographic task is:

$x =$lapte[RM], lait[FR], latte[IT], leche[SP], leite[PT]

$y =$lactem

and for the phonetic task is:

$x =$lapte[RM], lɛ[FR], latte[IT], letʃe[SP], lɐjtɨ[PT]

$y =$laktɛm

A cognate in one of the languages may be missing, in which case we represent it by a dash. Here, we are missing the Italian and Romanian cognates:

$x =$–[RM], tʁavaj[FR], –[IT], trabaxo[SP], tɾɐβaʎu[PT]

$y =$trɪpalɛm

At test time, we are given a set of cognates and are asked to predict their proto-word.

4 Comprehensive Romance Dataset

The different experiments described in the paper were performed on a large dataset of our creation, which contained cognates and their proto-words in both orthographic and phonetic (IPA) forms. The dataset's departure point is Ciobanu and Dinu (2014b), which consists of 3,218 complete cognate sets in six different languages: French, Italian, Spanish, Portuguese, Romanian and Latin [4]. We augmented the dataset's items with a

freely available resource, Wiktionary, whose data were manually checked against DIEZ and Donkin (1864) to ensure their etymological relatedness with the Latin source. The entries were transcribed into IPA using the transcription module of the eSpeak library[5], which offers transcriptions for all languages in our dataset, including Latin. The final dataset contains 8,799 cognate sets (not all of them complete), which were randomly splitted into train, evaluation and test sets: 7,038 cognate sets (80%) were used for training, 703 (8%) for evaluation and 1,055 (12%) for testing. Overall, the dataset contains 41,563 distinct words for a total of 83,126 words counting both the orthographic and the phonetic datasets. Vowel lengths were found to be difficult to recover (see Table 1), hence we created the following variations of the dataset: with and without vowel length (for both the orthographic and phonetic datasets), and without a contrast (for the phonetic dataset); see section §6 for further discussion.

A detailed description of the dataset collection process is available at the appendix §A.1. We make our additions to the dataset of Ciobanu and Dinu (2014b) publicly available [6].

5 Experimental Setup

5.1 NMT-based Neural Model

Our proto-word reconstruction setup follows an encoder-decoder with attention architecture, similar to contemporary neural machine translation (NMT) systems (Bahdanau et al., 2015; Cho et al., 2014).

We use a standard character-based encoder-decoder architecture with attention (Bahdanau et al., 2015). Both encoder and decoder are GRU networks with 150 cells. The encoder reads the forms of the words in the daughter languages, and output a contextualized representation of each character. At each decoding step, the decoder attends to the encoder's representations via a dot-product attention. The output of the attention is then fed into a MLP with 200 hidden units, which outputs the next Latin character to generate.

Input representation Each character (a letter in the orthographic case, and a phoneme in the phonetic case) is represented by an embedding vector

---

[4]We thank Ciobanu and Dinu for sharing their data with us.

	Edit Distance						
	0	**≤1**	**≤2**	**≤3**	**≤4**	**Average**	**Avg, norm**
Ortographic	64.1%	84.0%	92.7%	96.8%	98.5%	0.65	0.064
IPA	50.0%	73.9%	85.9%	93.3%	97.0%	1.022	0.100
Ortographic, added vowel lengths	42.5%	72.0%	86.3%	92.5%	96.7%	1.13	0.102
IPA, added vowel lengths	47.9%	62.0%	80.2%	87.1%	93.8%	1.331	0.119
IPA, no contrast	65.1%	80.0%	87.5%	93.6%	96.7%	0.797	0.077

Table 1: Distribution of edit distances between the reconstructed and original Latin form, on the orthographic and transcribed datsaets. Edit distance of 0 corresponds to perfect reconstruction. "Average" refers to average edit distance, and "Avg, norm" to normalized average edit distance.

of size 100. While all Romance languages are orthographically similar, the same letters represent different sounds, and thus convey different kinds of information for the task of Latin reconstruction. A possible approach would encode each language's characters using a unique embedding table. We instead share the character embedding table across all languages (including Latin), but concatenate to each character vector also a language-embedding vector. The final representation of a character c in language ℓ is then $WE[c] + UE[\ell]$ where E is a shared embedding matrix, c is a character id, ℓ is a language id, and W and U are a linear projection layers.

5.2 Evaluation Metric

Our main quantitative metric for evaluation is the edit distance between the reconstructed word and the gold Latin word. We use the standard edit distance with equal weight of 1 for deletion, insertion and substitution. We report test set average edit distance and average normalized edit distance (divided by word length), as well as the percentage of instances with less than k edit operations between the reconstruction and the gold, for $k = 0$ to 4.

6 Results and Analysis

Table 1 summarizes our main quantitative results. "Orthographic, added vowel lengths" and "IPA, added vowel lengths" refer to variations of the datasets that include explicit marking of vowel length in Latin words, marked by <ː> after long vowels. The models performance on the orthographic dataset demonstrates a substantial improvement over previously reported results. Our method has achieved average edit distance of 0.65, average normalized edit distance of 0.064, and 64.1% complete reconstruction rate (edit distance of 0). These numbers compare favorably with the edit distance of 1.07, normalized edit distance of 0.13 and 50% complete reconstruction reported

by Ciobanu and Dinu (2018). We note, however, that as our method is different both in the training corpus and in the type of model we employ, it is not clear whether this improvement should be attributed to the quality of the data, to the model, or to both of them.[7]

The performances on the phonetic dataset were lower than those derived from the orthographic one: in the phonetic dataset the average edit distance was of 1.022, and the average normalized edit distance of 0.1, with 50.0% complete reconstruction rate.

This disparity can be explained at least partially by a peculiarity of the phonetic dataset: it implicitly encodes vowel length, which was neutralized in the orthographic dataset. The reason for this difference is that length contrast in Latin co-occurred with quality differences: short vowels tended to be more open than their long counterparts, a contrast also called "tense-lax" (Allen and Allen, 1989). This contrast is not present in Latin orthography, but it appears in its phonetic transcription. This results in a noticeable gap between the results of the orthographic dataset with vowel lengths and without vowel lenghts (0.064 average normalized edit distance vs. 0.119), while the differences between the phonetic IPA dataset with vowel lenghts and without vowels lengths are much smaller. When the contrast "tense-lax" is manually neutralized[8], the performances achieved are similar to the ones on the orthographic dataset (as it is possible to see from the performances on "IPA, no contrast", whose Latin entries do not contain a "tense-lax"

---

[7] When we train a smaller version of our model (75-dimensional GRU) on the original dataset of Ciobanu and Dinu (2014b) we achieve average edit distance of 0.881, average normalized edit distance of 0.103, and complete reconstruction rate of 59.1%. Training a similar model on their dataset after cleaning resulted in average edit distance of 0.612, average normalized edit distance of 0.062 and complete reconstruction rate of 68.8%.

[8] We achieved that by respectively changing the characters <ʊ>, <ɔ>, <ɪ>, <ɛ> to <u>, <o>, <i>, <e> in the Latin words

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

Table 2: Error type distribution based on 650 orthographic and 650 phonological errors.

contrast).

6.1 Error Patterns

The following subsections focus on the model performances on the orthographic and phonetic datasets without explicit vowel length marking. A thorough analysis of both datasets reveals that the model's errors are not arbitrary, but rather tend to correspond to one of a few well-defined linguistic phenomena characterizing the evolution of Latin to its daughter languages. From an analysis of about 1300 errors, equally divided between the orthographic and the phonetic datasets, we find that 80% of the errors of the model on the orthographic dataset, and 75% on the phonetic one can be grouped into one of the following groups: high-mid vowel alternations, segment deletion, segment changes, cluster changes, morphological changes and other vowel changes. Additionally, one error category is unique to the phonetic dataset, tense-lax errors, and one is unique to the orthographic dataset, orthography errors. Table 2 summarizes the results, and Figure 2 visualizes the vowels error patterns on the phonetic dataset.

We briefly discuss each of these groups. [9]

High-mid alternation. The largest number of errors on the orthographic dataset, 18%, can be attributed to confusion between high and mid-high vowels (correspondingly <i>, <u> and <e>, <o>), as shown by the reconstruction <pescarium> instead of the Latin <piscarium> (alternation between <e> and <i>). That error is much rarer in the phonetic dataset, accounting only for 8% of all the errors. The reason of this error can be attributed to the origin of the mid-vowels in the daughter languages: while Latin

long vowels [iː], [eː], [oː] and [uː], always evolved into [i], [e], [o] and [u] in the daughter languages (with minor changes related to syllable structure), Latin short vowels—[ɪ], [ɛ], [ɔ] and [ʊ]—are not deterministically mapped into corresponding vowels in the daughter languages: Latin [ɪ] and [ɛ] both usually became Romance [e] (with alternations related to syllable structure, as diphthongization to [je]), while Latin [ɔ] and [ʊ] have different reflexes in the daughter languages as [u], [o], [ɔ], [ø] or as diphthongs. Because of this complex evolution, which merges different Latin phonemes into the same one in the daughter languages, the model is unable of unequivocally predicting the Latin vowel. Nonetheless, it seems that the tense-lax contrast present in the phonetic dataset eases the task of distinguishing the different phonemes, and enables the network to reconstruct their origin more often.

Segment deletion. examples of these errors are the reconstruction of <aspargum> instead of Latin <asparagum>, and the reconstruction of [abɪlɪtatɛm] instead of Latin [habɪlɪtatɛm]. During the evolution from Latin to Romance languages, unstressed syllables tended to be dropped. This phenomenon was not systematic, and occurred in different ways among and within the languages. Such process could affect either whole syllables (consonant + vowel) or only the vowel, creating new consonant clusters. Because of the erratic nature of this process, it seems that the network struggles with the exact reconstruction of segments eliminated in the daughter languages. A special kind of deletion is that of the consonant [h]. This consonant did not survive in any Romance languages (although it may be represented orthographically), and hence many times the network does not reconstruct it.

Segment changes. this category encompasses errors in the reconstruction of consonants— such as voicing changes (reconstructing <faculdadem> vs. Latin <facultatem>), assimilation ([wessarɛ] vs. [weksarɛ]) and gemination ([agrɛgatɪonɛm] vs. [aggrɛgatɪonɛm]). All these errors reflect processes that took place in all of the daughter languages, that obscures the original form of the proto-word.

Cluster changes. These are changes that occur with two contiguous consonants. Consider, for example, the reconstruction of [rɛatɪonɛm] instead of Latin [rɛaktɪonɛm], and of <sennorem> instead of Latin <seniorem>. The former is an

instance of cluster simplification, while the latter is an instance of cluster palatalization. In many of the daughter languages clusters of two different sounds underwent simplification, either by the dropping of one of the sound or the assimilation of one of them. Palatalization is the process by which certain sounds tend to be pronounced more closely to the palate, usually because of an adjacent front vowel. This change occurred in all Romance languages, even though its orthographic representation may vary among them.

Morphological changes. Latin had a very developed morphology, with several classes of special conjugations and irregular forms. The network struggles to reconstruct correctly irregular forms, as these forms were mostly lost in the daughter languages. An instance of such irregular verbs is <praeferre>, reconstructed as <praeferire> by the network. Moreover, other special morphological classes, such as Latin neuters, tend to be reconstructed as more usual forms. Another interesting class of errors is change of morphological category: some nouns have suffixes reminiscent of those of verbs, and hence are wrongly reconstructed as such. A separated case is that of Greek words: Latin contained several Greek loanwords that conserved their original morphology, different from the Latin one. Since these peculiarities were, for most part, not retained in the daughter languages, the network reconstructs them with normal Latin suffixes. For example, the greek [syntaksɪs] was reconstructed as [syntaksɛm], with the normal Latin suffix.

Other vowel changes. Latin contained several diphthongs, among them the diphthongs [aɪ] and [ɔɪ]. These sounds did not survive in any of the daughter languages (although in some rare cases they may be represented in the orthography), and both changed into [e] in the different Romance languages. This lef to reconstruction errors such as reconstructing <egrum> instead of Latin <aegrum>. Some changes also occurred with the vowel [a], which was reconstructed as a different vowel.

Greek orthography. some Latin words from Greek origin retained some orthographic conventions alien to Latin, such as the use of <y>, <ph>, <th>, <rh> etc. These conventions were only partially retained in the daughter languages, which creates some inconsistencies in their reconstruction by the network.
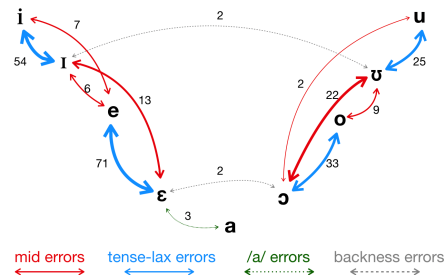
Tense-Lax alternation. this is the largest cate-

Figure 2: Phonological mistakes resulting from alternations between *vowels*, on the phonetic dataset. The numbers signify the number of errors, excluding singleton errors.

gory found in the networks errors on the phonetic dataset – up to 26% of all errors. As said previously, the tense-lax contrast reflects vowel length in Latin, which is not entirely predictable based on the daughter languages. The network tends to confuse between the lax and the tense vowels.

Figure 2 shows clearly that the network's errors are internally consistent and not random: all the vowel errors fit neatly in one of the aforementioned categories, while other possible errors do not occur.

Orthographic vs. Phonetic Importantly, the phonetic and orthographic tasks differ in their error distributions: while the performance of the network on the orthographic task displays many syllable changes – changes that alter the structure of the syllable (mostly changes in consonant clusters and deletion of segments) – on the phonetic tasks the model tends to retain syllable structure, but perform more segment-related errors (i.e., changing a specific vowel or consonant for another one). The IPA performance contained more idiosyncratic errors that could not be categorized in one of the main categories. Such errors tended to occur when the network had only one or two cognates from the daughter languages. Even though the orthographic performance also exhibited poorer reconstructions in these cases, it seems that the IPA performance was even more affected by the singular words, leading to more erratic reconstructions.

6.2 Learnt generalizations

This section will focus on the phonetic dataset. A closer inspection of the errors made by the model, and of those that do not occur in the data, can shed light on the processes of phonological change learnt by the model. We will first focus on the vowels. The Latin vowel [a] is quite resilient to

changes, and most of the daughter languages retain it without change (only in French and Romanian some phonological changes occur, in certain phonological environments). Indeed, the network has almost no mistakes in recovering it, apart from some isolated cases that derives from insufficient cognates in the daughter languages. The network also makes virtually no errors regarding the reconstruction of vowel backness – here also the only few cases are caused by the paucity of cognates and by assimilation processes in the daughter languages that make the Latin source opaque (metaphony processes). All in all, the network learns correctly the phonological changes that occurred in Latin vowels, and the main errors are a result of changes that cannot be fully reverted from the daughter languages.

The model learnt well the mapping of consonants between Latin and its daughter languages, and vowel reconstruction errors are considerably more prevalent. Focusing on one type of errors, palatalization, shows that the network failed to reconstruct the original consonant in opaque contexts, that is, when phonological cues crucial for the right reconstruction were lacking. Specifically, the network confused between the consonants [t] and [k] in the Latin reconstruction, since they palatalize to the same segments in Spanish and French. Without the other daughter languages, it is impossible to reconstruct correctly the original sound in Latin.

Finally, the network correctly generalized the occurrence of nasals in Latin clusters. Latin nasal tended to assimilate to the place of articulation of the adjacent consonant, deriving clusters such [ŋk], [mp] and [nt]. When the network deleted a consonant in a cluster containing a nasal, or changed a consonant adjacent to a nasal, the nasal consonant always changed to match the place of articulation of the following consonant. Hence, by deleting [k] in the cluster [ŋkt], the network reconstructs [nt]. Similarly, by changing [p] to [t] in the cluster [mp], the nasal consonant accordingly: [nt].

6.3 Evaluating Rules of Phonetic Change

To what extent did the model learn known rules of phonetic change?

The evolution of the Romance languages is well studied and linguists documented the set of phonological transformations that underwent between Latin and its daughter languages. We collected 33 of these phonological change rules, and used them to create a "synthetic" test set, containing syllable examples each focusing on a different phonological change. An example of a row in this dataset, corresponding to the rule of change of Latin [j] at word initial, is:

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

$$y = \text{ja}$$

Since the model was trained on complete words, isolated syllables tended to be unnatural for the network, and the output often contained additional consonants (usually morphological endings). When evaluating the model output we focus on the specific phonemes involved in the phonological change, and we ignore additional phonological material.

Results The complete list of synthetic examples and predictions is available at Table 3. The network correctly predicted 22 out of the 33 phonological rules (66.67% of the changes). The results are compatible with the results of the main reconstruction experiment: In both experiments, the network correctly reconstructed phonemes retained with little or no changes in all languages (e.g. [a] in different -phonological environments). Another class of phonemes correctly reconstructed in both cases are those which changed in a predictable way in each one of the daughter languages. Thus, [w] was correctly reconstructed since it predictably changed to [v] in all the daughter languages (apart from Spanish, which merged it with [b]). Phonemes that tended to change differently, but consistently, were also faithfully recovered: even though Latin [k] tended to change differently depending on the daughter language ([s] in French and Portuguese, [θ] in Spanish and [tʃ] in Italian and Romanian), it was reconstructed correctly because of the consistence of the change in each daughter language. The phonemes wrongly reconstructed tended to be those whose phonological change was "opaque". The "opaqueness" of their change can be ascribed to the fact that they were neutralized in the daughter languages, making it impossible to recover them without additional information. Relevant to this case are mostly vowels and diphthongs, as Latin [e] and [ɪ], which both became [e] in all the different daughter languages (with variants influenced by the phonological environment).

4466

Latin phoneme	Romanian	French	Italian	Spanish	Portuguese	Latin	Latin - reconstruction	Correct
/e/ blocked syllable	pep	pep	pep	pep	pep	pep	pɪp	no
/o/ blocked syllable	pop	pup	pop	pop	pop	pop	pʊp	no
/ɛ/ blocked syllable	pjep	pɛp	pɛp	pjep	pɛp	pɛp	pep	no
/kt/ medially, before nasals	-	anta	anta	anta	anta	an*k*ta	antam	no
/aɪ/	pe	pe	pe	pe	pe	p**aɪ**	pɛm	no
/ɔɪ/	pe	pe	pe	pe	pe	p**ɔɪ**	pɛm	no
/b/ intervocalic	aa	ava	ava	aβa	ava	a**b**a	awam	no
/e/ free syllable	pe	pwa	pe	pe	pe	pe	pɛm	no
/o/ free syllable	po	pø	po	po	po	po	pʊm	no
/ɪ/ free syllable	pe	pwa	pe	pe	pe	pɪ	pɛm	no
/n/ before front vowels	ɲi	ɲi	ɲi	ɲi	ɲi	ni	ɲidɛm	no
/a/ before nasal	pɨn	pan	pan	pan	pan	pan	pan	yes
/a/ blocked syllable	pap	pap	pap	pap	pap	pap	pap	yes
/i/	pi	pi	pi	pi	pi	pi	pi	yes
/u/	pu	py	pu	pu	pu	pu	pu	yes
/ɪ/ blocked syllable	pep	pep	pep	pep	pep	pɪp	pɪp	yes
/ʊ/ blocked syllable	pup	pup	pop	pop	pop	pʊp	pʊp	yes
/ɔ/ blocked syllable	pop	pɔp	pɔp	pwep	pɔp	pɔp	pɔp	yes
/k/ before front vowels	tʃi	si	tʃi	θi	si	ki	ki	yes
/sk/ before front vowels	ʃti	si	ʃi	θi	ʃi	ski	ski	yes
/kt/ medially, elsewhere	a**pt**a	ata	atta	atʃa	ata	a**kt**a	aktam	yes
/aʊ/	pau	pɔ	pɔ	po	po	pau	paʊm	yes
/pl/ word initial	**pl**a	**pl**a	**pj**a	ʎa	ʃa	**pl**a	**pl**am	yes
/a/ free syllable	pa	pa	pa	pa	pa	pa	pam	yes
/ɛ/ free syllable	pje	pje	pje	pje	pɛ	pɛ	pɛm	yes
/w/	**v**a	**v**a	**v**a	ba	**v**a	**w**a	**w**am	yes
/b/ word initial	**b**a	**b**a	**b**a	**b**a	**b**a	**b**a	**b**am	yes
/j/ word initial	**ʒ**a	**ʒ**a	**dʒ**a	xa	**ʒ**a	**j**a	**j**am	yes
/f/ word initial	**f**a	**f**a	**f**a	a	**f**a	**f**a	**f**am	yes
/f/ elsewhere	a**f**a	a**f**a	a**f**a	a**f**a	a**f**a	a**f**a	a**ff**am	yes
/ʊ/ free syllable	pu	pø	po	po	po	pʊ	pʊpʊm	yes
/ɔ/ free syllable	po	pø	pwɔ	pwe	pɔ	pɔ	pɔdɛm	yes
/l/ before front vowels	ji	ji	ʎi	xi	ʎi	li	ɡɪlʊm	yes

Table 3: the set of test phonemes used to evaluate the model's generalizations. Each row represents a distinct rule of phonetic change, which focuses on a single phoneme. The phoneme in question is bolded, and other consonants / vowels are added to simulate the phonological environment of the rule. The added consonants / vowels were chosen because they did not affect the evolution of the examined phonemes from Latin to the Romance languages. "correct" signifies whether the network's prediction were correct.
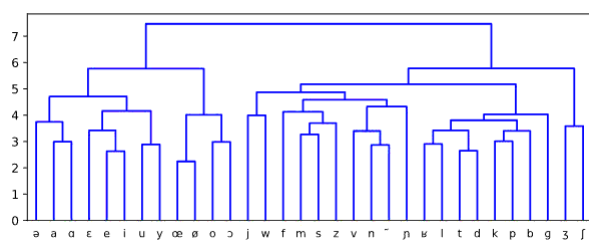

Figure 3: Hierarchical clustering of French phoneme embeddings

6.4 Learnt phoneme representations

Does training on proto-word reconstructions implicitly encourage the model to acquire phonologically-meaningful representations? We visualize the representation learned by network on the phonetic task by performing hierarchical clustering on the characters embedding vectors using the sklearn (Pedregosa et al., 2011) implementation of Ward variance minimization algorithm (Ward Jr, 1963).

Here we will briefly discuss the learned French phoneme representations (Figure 3). For all other languages, see appendix §5. As can be seen, the primary division that the network performs is between vowels and consonants, displayed on two different branches of the tree. On a lower level other phonologically motivated groupings are found: the network tends to place under the same node pairs of voiced and unvoiced consonants (as [ʃ] and [ʒ], [d] and [t]), allophones ([œ] and [ø]) or phonemes of the same category (as the glides [j] and [w]). To conclude, the results demonstrate the learning of a phonologically meaningful taxonomy of phonemes, without explicit supervision.
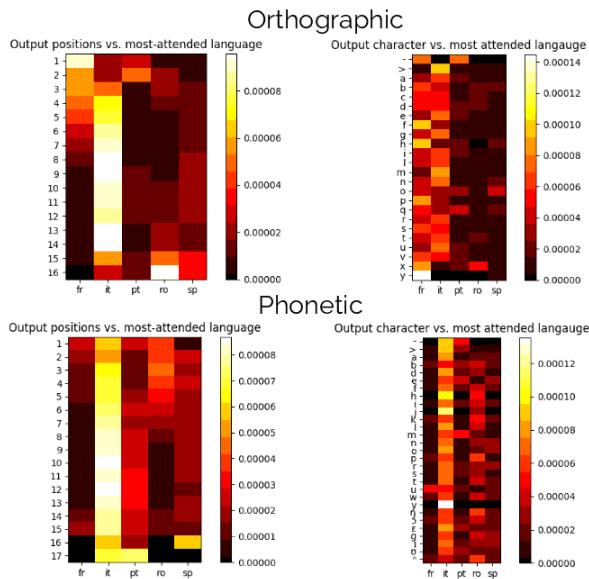
Figure 4: position in output vs. most attended language (left) and output letter vs. most attended language (right) for the orthographic (upper) and phonetic (lower) tasks.

6.5 Attention analysis

Since different languages can diverge to a varying extent from their proto-language, we hypothesize that the 5 daughter languages we use in this work would be of different importance for the model. To test this hypothesis, we inspect the learned attention weights. We focus on the most attended input character at each time step (the character having the largest attention weight) and count the number of times each of the 5 input languages is the most attended language, as a function of the location in the output and of the identity of the Latin character produced in that time step. We normalize the count with respect to time step, letter frequency and language frequency in the corpus.

Results The results for the phonetic and orthographic tasks are presented in Figure 4. In both cases, Italian is the most attended language. There are some differences between the settings, however. For the orthographic task, the network focuses noticeably more on French than in the phonetic task. This tendency can be attributed to the very conservative orthography of French, that masks the phonological innovations that occurred in the language. Indeed, the network focuses exclusively on French for the reconstruction of the characters <h> and <y>, which are consistently represented only in French orthography, disappearing from the written form of the other Romance languages. The comparison to the atten-

tion of the phonetic dataset shows that the network tends to actually ignore French, favoring other sources instead. Similarly, in the orthographic dataset, French is favored in the initial positions, a tendency that disappears in the phonetic dataset. Finally, an interesting trend in the phonetic dataset is a tendency to attend to Romanian at the initial positions and to Portuguese at later ones.

7 Conclusions

In this work, we introduce a new dataset for the task of proto-word reconstruction in the Romance language family, and used it to evaluate the ability of neural networks to capture the regularities of historic language change. We have shown that neural methods outperform previously suggested models for this task. Analysis of the linguistic generalizations the model acquires during training demonstrated that the mistakes are related to the complexity of the phonetic change. A controlled experiment on a set of rules for phonetic alternations between Latin and its daughter languages demonstrated the model internalizes some of the systematic processes that Latin had undergone during the evolution of the Romance languages. Visualizing the learned phoneme-embedding vectors has revealed a hierarchical division of phonemes that reflects phonological realities, and inspection of attention patterns demonstrated the model attributes different importance to different languages, in a position-dependent manner.

While the task examined in this paper is commonly called "proto-word reconstruction", in practice the task the model faces is considerably less challenging than the work of historical linguists, as the model is trained in a supervised setting. A future line of work we suggest is applying neural models for the end task of proto word reconstruction, without relying on cognates lists, in a way that would more naturally model the historical linguistic methodology.

Acknowledgements

References

Ti Alkire and Carol Rosen. 2010. *Romance languages: A historical introduction*. Cambridge University Press.

W Sidney Allen and William Sidney Allen. 1989. *Vox latina*. Cambridge University Press.

Raimo Anttila. 1989. *Historical and comparative linguistics*, volume 6. John Benjamins Publishing.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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*, pages 65–73. Association for Computational Linguistics.

Alexandre Bouchard-Cote, David Hall, Thomas L Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Alexandre Bouchard-Côté, Percy Liang, Thomas L. Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 887–896.

Peter Boyd-Bowman. 1980. *From Latin to Romance in sound charts*. Georgetown University Press.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111.

Alina Maria Ciobanu and Liviu P Dinu. 2014a. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 99–105.

Alina Maria Ciobanu and Liviu P Dinu. 2014b. Building a dataset of multilingual cognates for the romanian lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC*, pages 1038–1043.

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*, pages 1604–1614.

J Halvor Clegg and Willis C Fails. 2017. *Manual de fonética y fonología españolas*. Routledge.

Josette Rey Debove and Alain Rey. 2000. *Le Nouveau Petit Robert: Dictionnaire alphabetique et analogique de la langue francaise*. Dictionnaires Le Robert.

Friedrich Christian DIEZ and TC Donkin. 1864. *An Etymological Dictionary of the Romance Languages; chiefly from the German of F. Diez. By TC Donkin*. Williams and Norgate.

Gerhard Ernst. 2003. *Romanische Sprachgeschichte: Ein Internationales Handbuch Zur Geschichte Der Romanischen Sprachen Und Ihrer Erforschung/Manuel International Sur L'Histoire Et L'Etude Linguistique DES Langues Romanes*. Mouton de Gruyter.

Félix Gaffiot and Pierre Flobert. 1934. *Dictionnaire latin-français*. Hachette Paris.

Robert A Hall. 1944. Italian phonemes and orthography. *Italica*, 21(2):72–82.

Martin Harris and Nigel Vincent. 2003. *The romance languages*. Routledge.

Günter Holtus, Michael Metzeltin, and Christian Schmitt. 1989. *LRL*, volume 3. M. Niemeyer.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Adam Ledgeway and Martin Maiden. 2016. *The Oxford guide to the Romance languages*, volume 1. Oxford University Press.

Charlton T. Lewis and Charles Short. 1879. A latin dictionary. *Perseus Digital Library*.

Dylan Lewis, Winston Wu, Arya D. McCarthy, and David Yarowsky. 2020. Neural transduction for multilingual lexical translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4373–4384. International Committee on Computational Linguistics.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 599–605.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL 2001, Pittsburgh, PA, USA, June 2-7, 2001.*

Maria Helena Mateus and Ernesto d'Andrade. 2000. *The phonology of Portuguese*. OUP Oxford.

Robert McColl Millar. 2013. *Trask's historical linguistics*. Routledge.

Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographics cues for cognate recognition. In *LREC*, pages 2387–2390.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2020. Multi-source neural machine translation with missing data. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:569–580.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 393–400.

Mika Sarlin. 2014. *Romanian grammar*. BoD-Books on Demand.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Winston Wu, Garrett Nicolai, and David Yarowsky. 2020. Multilingual dictionary based construction of core vocabulary. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4211–4217. European Language Resources Association.

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 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

A Appendix

A.1 Dataset Creation

In order to perform the reconstruction task, we required a large dataset of cognates and their proto-words, in both orthographic and phonetic (IPA) forms.

Despite growing interest in recent years, high-quality digital resources for the tasks of proto-word reconstruction and cognates detection are scarce. Our departure point is the dataset provided by Ciobanu and Dinu (2014b), which, to the best of our knowledge, is the most extensive dataset for proto-word reconstruction of a well-attested proto-language. The dataset contains 3,218 complete cognate sets in five Romance languages (Spanish, Italian, Portuguese, French, Romanian) together with their Latin etymological ancestor. Although being a valuable resource, this dataset was constructed via automatic method of cognate extraction, and a comparison with references on the development of Romance languages (Boyd-Bowman, 1980; Alkire and Rosen, 2010) reveals some problems, such as false cognates, truncated forms, non-existent words and mismatch between the part of speech of the cognates and the ancestor. Another salient problem of the dataset regarded the grammatical case of Latin nouns: Romance languages derived their words from the accusative Latin case (Harris and Vincent, 2003), while in the dataset Latin words were displayed in the nominative case, an inconsistency making the reconstruction inherently more challenging.

Lastly, as neural models often requires large amounts of training data, we aimed to expand the dataset. We thus created a cleaned and extended dataset by Wiktionary scrapping, followed by manual validation and cleansing.

Wiktionary scraping We augment the existing dataset with a freely available resource: Wiktionary. Wiktionary entries for Latin words usually contain inflection tables, and often list the descendants in Romance languages; these descendents are, by definition, cognates. We scraped all Latin entries from Wiktionary, and extracted the forms of the daughter languages (available in the "Descendants" section). This resulted in 5,598 additional comparative entries, for a total of 22,361 new individual words. Contrary to the previous dataset, the Wiktionary-derived cognates are not based on automatic alignment between translations, but rather on direct human annotation. On the other hand, the Wiktionary-based entries are often incomplete, and include cognates in only a subset of the daughter languages.

Form normalization Using the Wiktionary-provided inflection tables, we decline the Latin nouns to the accusative case, and conjugate verbs to the infinitive form. We do this both to the Wiktionary-based entries and to the ones in the original dataset. We selected a sample of around 100 Latin words to check the accuracy of the automatic conjugation, against Gaffiot and Flobert (1934), finding them all correct. Finally, Latin words in the Ciobanu and Dinu (2014b) dataset for which we did not find a Wiktionary entry were conjugated "manually" by consulting (Lewis and Short, 1879; Gaffiot and Flobert, 1934).

Manual verification and cleaning After the collection of the Wikitionary dataset, we went manually through all the Latin words contained in Ciobanu and Dinu (2014b), checking them against Lewis and Short (1879); Gaffiot and Flobert (1934). Additionally, we went over the some suspicious-looking words from the daughter language and verified them against DIEZ and Donkin (1864) to ensure their etymological relatedness with the Latin source, fixing if necessary.This sort of fix was not performed systematically, but we did fix or remove around 170 words.

Finally, we sample 300 entries from the original (Ciobanu and Dinu, 2014b) dataset prior to cleaning and 300 words from our cleaned and unified version of the dataset, and manually verified them. We find 43 mistakes in the original dataset and only 4 in our version, indicating that, while still not perfect, it is of substantial higher quality.

IPA transcription To obtain the phonetic transcriptions into IPA, we utilized the transcription module of the eSpeak library, which offers transcriptions for all languages in our dataset, including Latin. While a human transcription would be preferable, a manual evaluation of 200 of the resulting transcriptions by comparing them against several sources (Allen and Allen, 1989; Hall, 1944; Debove and Rey, 2000; Clegg and Fails, 2017; Mateus and d'Andrade, 2000; Sarlin, 2014) show high accuracy: all the 200 words were correct, except for minor systematic changes which we fixed globally to better suit the transcription to phonological conventions. Specifically, we deleted the vowel symbols <ʊ> and <ɪ> in Italian and Romanian, which resulted to be alien to

those languages, changed the sequence <rr> to <r> in Spanish, and regularized the Portuguese transcriptions, which showed some phonological traits of Brazilian Portuguese.

Final dataset The resulting dataset, used for all experiments in this work, contains 8,799 entries. The dataset was randomly splitted into train, evaluation and test sets, with 7,038 examples (80%) used for training, 703 (8%) for evaluation and 1,055 (12%) for testing.

Overall, the dataset contains 41,563 distinct words across the different languages (for a total of 83,126 words counting both the orthographic and the phonetic datasets), with 7,384 Italian words, 7,183 Spanish words, 6,806 Portuguese words, 6,505 French words and 4,886 Romanian words. As vowel lengths were found to be difficult to recover, we created the following variations of the dataset: with and without vowel length (for both the orthographic and phonetic datasets), and without a contrast (for the phonetic dataset).

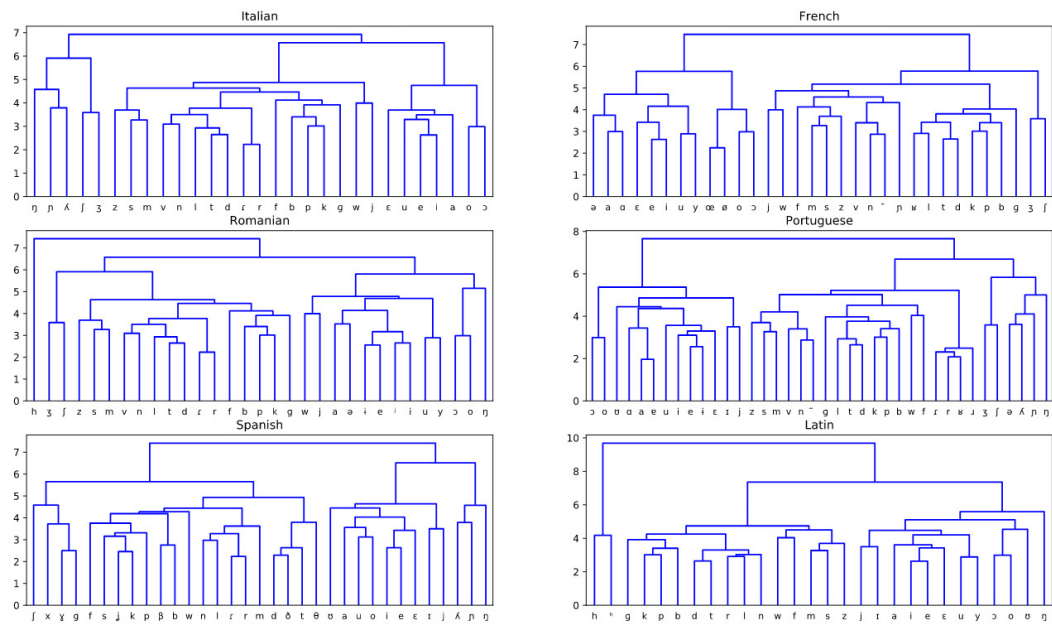## A.2 Phoneme representations

Figure 5

In this appendix, we show the hierarchical clustering created for all the languages in our dataset. As it can be noted from figure 5, the results for the different languages exhibit representations similar to those found in the French clustering: the primary division in each language is between vowels and consonants. In Portuguese, Latin, Spanish and Romanian some consonants are grouped together with vowels. These consonants are restricted to nasals, liquids or glides. The inclusion of these consonants can be explained by the peculiarity of their nature: all of them have a special phonological status, displaying similarities in their behavior to vowels. In all languages phonologically related phonemes tend to be group under the same nodes. Among the others, glides are either found together with each other (as in French, Italian and Romanian) or with their vocalic counterparts (Latin, Spanish and Portuguese), consonants differentiated only in voicing are usually paired ([ʃ] and [ʒ]), front and back vowels forms clusters and allophones usually shares the same node (Italian, French, Romanian, Spanish, Portuguese).