

# Large-Scale Computerized Forward Reconstruction Yields New Perspectives in French Diachronic Phonology

Clayton Marr  
The Ohio State University

David Mortensen  
Carnegie Mellon University

## Abstract

Traditionally, historical phonologists have relied on tedious manual derivations to sequence the sound changes that have shaped the phonological evolution of languages. However, humans are prone to errors, and cannot track thousands of parallel derivations in any efficient manner. We demonstrate *computerized forward reconstruction* (CFR), deriving each etymon in parallel, as a task with metrics to optimize, and as a tool which drastically facilitates inquiry. To this end we present DiaSim, an application which simulates “cascades” of diachronic developments over a language’s lexicon and provides various diagnostics for “debugging” those cascades. We test our method on a Latin-to-French reflex prediction task, using a newly compiled, publicly available dataset *FLLex* consisting of 1368 paired Latin and modern French forms. We also introduce a second dataset, *FLLAPS*, which maps 310 reflexes from Latin through five attested intermediate stages up to Modern French, derived from Pope (1934)’s periodic development tables. We present publicly available rule cascades: the baseline *BaseCLEF* and *BaseCLEF\** cascades, based on Pope (1934)’s widely-cited view of French development, and *DiaCLEF*, made from incremental corrections to *BaseCLEF* aided by DiaSim’s diagnostics. DiaCLEF outperforms the baselines by large margins, improving raw accuracy on FLLex from 3.2% to 84.9% of etyma, with similarly large improvements for each of FLLAPS’ periods. Changes were made to build DiaCLEF considering only the baseline and DiaSim’s diagnostics, but they often independently reproduced past work in French diachronic phonology, corroborating both our procedure and past endeavors; we discuss the implications of some of our findings in detail.

**Keywords:** phonology, computational historical linguistics, French, sound change, forward reconstruction

## 1 Introduction

When reconstructing the phonological history of a language, linguists most often operate under the Neogrammarian assumption that sound change operates on an input defined by its phonetic characteristics, can be conditioned based on its phonetic context, and results in a predictable output,

with no exceptions.<sup>1</sup> This paradigm operationalizes sound change as a classical function: input information maps to a unique output. In aggregate, the sequence of the operations of these sound change functions naturally form an algorithm. Furthermore, diachronic sound change is inherently ordered, as different changes happened at different times. As such, they may create or perturb the necessary environments for later changes, ultimately causing a cascading effect that makes the ordering of sound changes critical for the outcome.

One could imagine that, upon learning this, a non-linguist might guess that most current work in historical phonology uses computerized simulations for these diachronic algorithms, to test whether their understanding of a language’s history actually produces the correct outcomes. However, while the theoretical underpinnings for such a method, termed *computerized forward reconstruction* (Sims-Williams 2018) (henceforth *CFR*), do exist, for reasons discussed in more depth in §2.2, it has failed thus far to achieve widespread usage. Instead, much of the research in diachronic phonology has tended to analyse at high resolution the specifics of certain types of sound changes cross-linguistically, and rarely explicitly and holistically tackles how they fit together in the whole of any one language’s phonological history. To verify our understanding of that latter “bigger picture”, the diachronic phonologist would have to either write down or memorize the effects of hundreds or thousands of rules operating over millennia, mapping onto native vocabularies of at least thousands of etyma – a task no human could possibly do efficiently. It is unsurprising, then, that most current work tends to “zoom in” on one phenomenon, and that if the “larger picture” is tackled, it is with regard to the generalizing typology for how sound changes may occur across languages, rather than how they fit together to form the history of one language.

These typological discussions are useful and informative, but they must remain grounded in our knowledge of the larger histories of the languages in question or we may end up relying on false premises due to incorrect understandings of the state of a language at a given time. Furthermore, the phonological histories of the majority of the world’s languages, which likely will not survive the next century, remain mysterious. While valuable work continues to be done in reconstructing them, this task would certainly be much more efficient if aided by computers. While humans without prior knowledge can take a matter of months to correctly execute the mapping of thousands of etyma across millennia, a computer can perform the task in a matter of seconds. Transparent computerized forward reconstruction furthermore greatly facilitates thoroughness in checking both the claimed results and coverage of the native lexicon. Building on the example of earlier now abandoned projects discussed in §2.2, we present *DiaSim*<sup>2</sup>, a transparent forward reconstruction application which offers generalizability as well as various diagnostic capabilities, in hopes that it could help ameliorate the current situation.

To demonstrate what computerized forward reconstruction (CFR) can contribute to the field, we present the fruits of our work in using *DiaSim* to “debug” the early 20th century received un-

---

<sup>1</sup>That is, as long as the sound change in question is purely phonetically motivated, all other things being equal. Many changes in sound which are not proper “sound changes” do not need to be “regular”, if they are not phonologically motivated; examples include analogy, folk etymology, homophony or taboo avoidance, hypercorrection, contamination, and other sociolinguistic factors. These same factors can also end up systematically perturbing the former regularity and phonetic conditioning of sound changes (Janda & Joseph 2003; Mazzola 2013), or *rules*. This model also requires one to specify the order in which rules apply, and whether they operate forwards or backwards over a word.

<sup>2</sup>First presented in Marr & Mortensen (2020).

derstanding of French phonological history, as represented by Pope (1934). French has a large historical corpus and extensive prior research, so one might be confused by the selection of a language for which a diachronic simulation is less necessary; however, in this way, French offers us the best way to show that such a system can lead us in the same direction of improved understanding that later scholarship building on Pope (1934) did, as well as generate novel insights. The corrections we generated with DiaSim’s help often agree with the later literature on French phonological history; in cases where they don’t, we present arguments why these corrections represent new hypotheses that should be given consideration, as DiaSim may have revealed what was previously a blind spot in the field even in such a well-studied language such as French – such as a new regular explanation for the voicing of Gallo-Roman initial /k/, a phenomenon traditionally thought to be idiosyncratic or sporadic.

We additionally present the newly compiled datasets *FLLex* and *FLLAPS* (described in §6), with which we demonstrate the use of DiaSim to debug not just with respect to a final outcome, but also to match any information from any intermediate forms that may be available. We present results on the measured performance of baseline rule cascades BaseCLEF and BaseCLEF\*, and the improved version of DiaCLEF, which essentially represents our understanding of the entire phonological history of French, built on Pope (1934) using DiaSim.

Overall, while the baseline model was accurate for 3.2% of etyma, the corrected cascade achieved accuracy on 84.9% of etyma, with similar improvements across each stage featured in FLLAPS. Results are discussed at further length in §8. All of these resources are available online.<sup>3</sup>

## 2 Background

### 2.1 French phonological history

Romance philology was formalized in the works of Diefenbach (1831) and Diez (1836). In the late 19th century, a second generation of Romance linguists informed by Neogrammarianism worked on foundational work for French (Bourciez 1889; Meyer-Lübke 1908; Nyrop 1899). Early studies focusing on specific periods included Thurot (1881) for the early modern period, Suchier (1893) for Old French, and Marchot (1901) for Gallo-Roman. Later major works include Brunot and Charlier (1927), Richter (1934) and of course Pope (1934). Significant work was done later (Fouché 1952/1966; Martinet 1970; Straka 1970), and to an extent continues to be revised (Buckley 2003; Machonis 1990; Morin 2009; Posner 1994; Zink 1986). Attention is increasingly paid to French historical sociolinguistics (Lodge 2004, 2013; Lusignan 1986), French orthographical history, and “protofrançais” (Banniard 2001; Noske 2011). The magnitude of phonological deviation from Latin (Grimes & Agard 1959; Pei 1949) within Romance is greatest in French (Posner 1996:199), giving French the greatest wealth of sound change to study and model.

Traditional methodology combined Neogrammarianism with the principle that “the history of a language should be related as closely as possible to the study of texts” (Pope 1934). This often involved tracing orthographic changes especially as relevant to phonemes and morphemes (“flexion”), and taking the remarks of then-contemporary writers and grammarians as objective evidence.

---

<sup>3</sup><https://github.com/clmarr/DiaSim>

In the case of French, writing on how the language is or (more often) *should* be pronounced is quite abundant from the late thirteenth century onward, and Pope is often reliant on the words of esteemed Renaissance writers such as Palsgrave, Meigret, Bèze, and Estienne. We, like others (Fouché 1952/1966; Posner 2011), see these texts as *prescribing* how French *should* be pronounced, rather than *describing* contemporary realities.<sup>4</sup> Rather than rely on these sources, we privilege foremost the sequence of rules that leads to the best resulting accuracy, and only after determining that sequence do we consider its implications with respect to prior knowledge, including remarks by such historical writers.

Research on the diachronic phonology of French has functioned more or less to fit proposed sequences of sound changes (*diachronic cascades*) to the observed development of French. Two hundred years ago, at a time when the concept of Vulgar Latin was novel and that French came from Latin at all was still disputed (Blom 2009; Posner 1996), our knowledge of this diachronic cascade remained crude. In the early 20th century, the major aspects of French diachrony were clear, but many details remained fuzzy. Pope’s 1934 opus is recognized as the “invaluable” (Posner 1997:3) baseline against which new theories in French are presented as improving upon (Short 2013). In this, it sits among a set of other pertinent analogous works covering French diachrony, of which any would make a fine baseline (Bourciez & Bourciez 1967; de la Chaussée 1974; Fouché 1952/1966; Rheimfelder 1975; Richter 1934; Straka 1970). Perhaps the critically useful aspect of works like Pope (1934) is a holistic account of sound change “from Classical Latin to Modern French”, paired with extensive philological research for all well attested periods. Our aim in this paper is twofold. Alongside the goal of demonstrating the power of CFR, we also aim to, like Pope, provide a holistic account of the inherited French diachronic cascade. Ultimately, our vision is that a publicly available cascade for every language of interest may be improved upon whenever a correction becomes accepted in the field. With this in mind, we publicly release DiaCLEF, our resulting French cascade, to be critiqued as necessary by peers.<sup>5</sup>

## 2.2 Computerized forward reconstruction (CFR)

Soon after the mid-20th century emergence of computational historical linguistics with the works of scholars like Swadesh and Gleason (Dunn 2015; Gleason 1959; Swadesh 1952), the first computerized forward reconstructions (coarsely) derived Russian from Proto-Indo-European (Smith 1969) and Old French from Latin (Burton-Hunter 1976). Others looked at Medieval Ibero-Romance from Latin (Eastlack 1977), Latin from Proto-Indo-European (Maniet 1985), Old Church Slavonic from PIE (Borin 1988), Bantu (Hombert et al. 1991), and Polish from Proto-Slavic (Kondrak 2002). These systems were not intended to be generalizable, lacked sufficiently expressive rule formalisms,

---

<sup>4</sup>In fact, many grammarians were concerned with ensuring that French would be and remain suitable for metrical poetry. Many erroneously believed that French descended from Gaulish or Germanic rather than neo-Latin (Blom 2009:46–47) and thus aimed to force it into line with the “civilizational wellspring” of Latin or Greek. One notorious case is Henri Estienne, who explicitly endeavored in his *Traicté de la conformité du langue françois avec le grec* to make French adhere to Ancient Greek metric structure. These writers were also typically conservative voices that advocated for spelling pronunciations. Often, more can be gleaned from what pronunciations these authors condemn than the ones they support, but even the former category present a biased sample inevitably drawn disproportionately from speech forms that upper class urban older males would likely encounter.

<sup>5</sup><https://github.com/clmarr/DiaSim>

and operated on orthography rather than underlying phones (Piwowarczyk 2016), having “no notion of phonology” (Kondrak 2002).

*Phono*, a phonologically-motivated and phoneme-mediated CFR system, appeared in 1996 and was applied to Spanish and Shawnee (Hartman 2003; Muzaffar 1997), but no further work was published using it (Sims-Williams 2018). Despite an “explosion” in computational modeling in other diachronic fields (Dunn 2015) alongside rapid improvements in modern computing, CFR fell out of fashion by the 1980s and 1990s (Lowe & Mazaudon 1994), and most of the old derivation systems are incompatible with modern computers (Kondrak 2002).<sup>6</sup> Reasons for this decline are varied, including the mid-20th century popularity of non-Neogrammarian approaches, and a view that CFR systems were not “serious” research tools (Sims-Williams 2018).

### 3 Contributions

This paper aims to show that a sufficiently generalizable forward reconstruction system can be used for serious work in diachronic phonology. It is recognized (Sims-Williams 2018) that human memory cannot compete with the speed, accuracy, and thoroughness with which such systems can detect omissions from and exceptions to any given ordered rule sequence (henceforth, *rule cascade*). We acknowledge that computerized forward reconstruction has been done before but did not achieve widespread usage. However, this is not because of flaws in the methodology, but rather that prior systems were not intended to be generalized as empirical methods, and were not considered “serious” work (Lowe & Mazaudon 1994), despite their potential to improve scientific inquiry.

Human working memory is unreliable and cripplingly inefficient for the millions of calculations necessary to holistically test their understanding of a language’s phonological history.

On the other hand, sequentially mapping thousands of functions over thousands of tokens is trivial for a computer. Information attained in this much more efficient and rigorous manner can then be leveraged to improve our understanding of the phonological histories in question. We can find new sound laws and analogical patterns, refining existing ones, and revealing new reflexes and cognates. Such systems can better inform ongoing debates in historical phonology, as information obtained from transparent mass simulation is representative and holistic. Transparent mass simulation provides a method to ensure that sampling of test cases is done in a principled and comprehensive fashion. This improved efficiency in rigor could be crucial for advancing our critical understanding for less well studied and especially endangered language families — especially where phylogeny, which often relies on diachronic phonology, is concerned. It has also been argued that the more efficient rigor of transparent mass diachronic simulation will help inform theoretical debates and larger structural phenomena (Sims-Williams 2018).

This paper contributes the following:

- DiaSim, a tool which addresses these issues by transparently simulating the realization of a rule cascade over a lexicon

---

<sup>6</sup>There is nonetheless promising recent work that uses methods that essentially implement CFR, such as Pyysalo (2017)’s generative approach to Indo-European etymology.

- FLLex, a dataset consisting of 1368 Classical Latin etyma paired with their inherited Modern French reflexes.
- FLLAPS, an analogous dataset consisting of 306 Classical Latin etyma associated with their reflexes in a series of stages including Late Latin, two stages of Old French, Middle French and Modern French, from Pope (1934)’s periodic tables and phonological values.
- A set of metrics for evaluating a simulated derivation against a gold standard set, and a set of diagnostics for pinpointing sources of errors
- BaseCLEF, a cascade of sound changes from Classical Latin to Modern French as per the received account provided by Pope (1934).
- DiaCLEF, an improved cascade of sound changes developed using DiaSim
- An empirical approach for using CFR for scientific inquiry (§4)

#### 4 Iterative Refinement of an Analysis using DiaSim

DiaSim, like any tool, calls for a methodology.<sup>7</sup> Unlike certain completely data-driven machine learning approaches to linguistic problems, it does not presuppose starting only with empirical observations. Instead, it should be primed by a cascade of historical sound changes representing an initial hypothesis, which can then be iteratively refined. This cascade should be based on the established view in the relevant literature. On each iteration, the linguist first uses DiaSim to isolate the source of an error in the working cascade; it is often preferable to seek the source of error using automated, often statistical, methods. Then, they apply all plausible solutions (re-ordering rules, reformulating rules, and so on) and select the solution (or solutions) that optimize along the DiaSim metrics.

If two or more solutions result in the same increase in accuracy, the linguist favors the solution that is most parsimonious (that is, that requires the fewest assumptions or stipulations). Generalized and predictive solutions using a constrained formal apparatus are more parsimonious than multitudes of lexically idiosyncratic explanations.<sup>8</sup> A solution, however, may be formally elegant or boost performance on narrowly defined metrics but be linguistically implausible. The next step is to test the solution against linguistic knowledge in order to evaluate its plausibility. If a new sound change is proposed, it ideally should have some basis in articulation, perception, sociolinguistic factors, or—at the very least—belong to a cross-linguistically attested class of changes. If a reordering of sound changes is proposed, the linguist should ensure that (to the degree possible) the resulting order accords with what can be determined on philological and other grounds (but see §8.4.2).

---

<sup>7</sup>The methodology described here is not specific to DiaSim and should be applied to any comparable tool.

<sup>8</sup>“Generalized and predictive solutions” need not be solely phonological, but can also include principled appeals to analogical, sound-symbolic, orthographic and sociolinguistic factors.

This evaluation process works best, in our experience, when it is grounded in the history of scholarship. Indeed, on multiple occasions, we have proposed solutions to problems in the historical phonology of French, only to find that the same solutions were already proposed in earlier literature, providing some confirmation of the solution (see, e.g., §8.2.1 and §8.3). Contrariwise, existing scholarship may provide good reasons for doubting a solution that seems attractive on the basis of DiaSim’s metrics alone. In other words, we do not advocate DiaSim as a *replacement* or *displacement* of traditional methods, but as a tool that can make these methods both more efficient and more rigorous.

The last step in each iteration is to choose a solution, from the competing set, based on the evaluation step, and apply it to the working cascade. The process then repeats until it “converges”: until successive iterations do not produce any improvement in performance. We do not claim that such a “converged” cascade is *optimized* in any ultimate sense, but only that it embodies a coherent and defensible hypothesis about the development of a language. In other words, we do not see this methodology as a sausage grinder that can mechanically turn out the right analysis but as a fine scalpel meant to be wielded by a skilled historical linguist in the dissection of diachronic beasts.

## 5 DiaSim

### 5.1 Transparent mass simulation

DiaSim simulates a specified rule cascade for every lexeme in the data set simultaneously.

At minimum, the user must supply (1) a lexicon file, and (2) a cascade (if either is different from the ones we provide). The lexicon file includes the input forms to forward reconstruction, and optionally gold standard reflex forms for the final or intermediate results of computerized forward reconstruction (CFR). Each rule in the cascade is written in the conventional SPE format (Chomsky & Halle 1968). DiaSim implements the subset of the SPE rule formalism that Johnson (1972/2019) and Kaplan & Kay (1981) showed to be formally equivalent to finite state transducers (*FSTs*), while also allowing for explicit surface-level modeling of phonological features.<sup>9</sup> DiaSim rule cascades are able to capture any and all regular relations between strings in the specified symbol alphabet, using either the IPA default supplied as part of the DiaSim package, or any other one supplied by the user.

In between any two rules, the user may flag a stage. In this way, the simulation state at that stage can be stored and retrieved, and compared to a stage-wise gold standard set if supplied.

Being able to observe the iterative realization of cascade *transparently* (i.e., the results of each rule can be retrieved at will), is itself quite useful for better understanding the processes involved as it reveals how the preconditions for later rules may fall into place or be perturbed. For this end of *transparency*, DiaSim can output the derivation specifying each instance an etymon was changed, its new form and by what rule, for any etymon. This capability is illustrated in Figure 1, which points to where the critical divergence from the correct path that would have led to the (accurate) modern form /mənas/ rather than the observed form which has /a/ instead of /a/ (note, however, /a/

---

<sup>9</sup>As opposed to implicitly doing so while working directly with FSTs. Note: Johnson’s demonstration requires that rules do not immediately apply to their own outputs.

```

#m,əŋ'at'sə# | R534 : t's' > t's
#m,əŋ'at'sə# | R628 : [+syl,-front] > [+nas] / __ [+nas,-syl]
#m,əŋ'asə# | R648 : [+delrel] > [+cont]
#m,əŋ'asə# | R653 : {ə;,ə} > {ə;,ə}
#m,əŋ'a:sə# | R706 : [-round,+syl] > [+long] / __ s ə #
#m,əŋ'a:sə# | R708 : ə > [-syl] / __ #
#m,əŋ'a:sə# | R715 : [+lo,+long] > [+back]
#m,əŋ'a:s# | R736 : ə > ø
#məŋa:s# | R753 : [+stres] > [-stres]
#məŋas# | R754 : [+syl,+long] > [-long]

```

Figure 1: The printed derivation of *menace* (< Latin MINACIA).

and /a/ have now, as of the 21st century, largely merged in standard French). Flagged stages may also be used as “pivots”, at which the current forms of lexemes can be retrieved and compared against correct forms for that state in time if provided. This is quite useful for untangling the interaction of long-term patterns and deducing when error is introduced.

### 5.1.1 Performance metrics

For either the entire lexicon or a specific subset, when comparing to reference forms DiaSim can supply the word-wise Accuracy (proportion accurate), the Accuracy within a certain number of phones, the word-wise average Levenshtein distance between result and gold standard form (normalized for gold-standard length), and the word-wise average length-normalized feature edit distance (*FED*) (Kondrak 2003; Mortensen et al. 2016) between result and gold standard forms.

These different measures offer different information. Accuracy indicates what percent of the lexicon the rule cascade renders correct. On the other hand, Levenshtein distance measures how wrong we are overall if we consider phones as discrete tokens. Finally, feature edit distance (*FED*) approximates phonetic distance between result and gold standard forms by measuring the average feature vector distance over the the entire lexicon — i.e., where there is a wrong phone, just how different is it from the correct one? We also advocate that future work should incorporate some measure of “overall cascade complexity” that should be kept relatively low to avoid overfitting.

## 5.2 Diagnostics

Aside from failure to consider how the rule cascade could affect every word in the lexicon, significant sources of error could be missed, especially where rules interact, given the multiplicity of all the factors at play. Additionally, what is actually observed as one relatively acute error could in fact be a sign of a much larger pattern of errors. These factors are likely more problematic in less studied languages as we inevitably have less source information that could help inform us on the relative dating of phenomena. To facilitate empiricism, DiaSim comes equipped with a suite of diagnostics to help users identify what is going wrong.

By default, DiaSim runs the simulation and creates the relevant output files as specified at command line. If interactive mode is flagged at command line, at the end of the simulation, and also any specified gold stage flagged by the user, DiaSim halts, gives basic performance metrics, and queries the user if they’d like to run any of DiaSim’s various diagnostic options. These options include diagnostics in terms of phone errors, phone sequence-wise diagnostics, printing metrics for defined subsets of the lexicon, correlating error with any saved stage, and testing what the effect



```

Success: now making subsample with filter 'a [+ant,+strid,-cont] ə
(Pivot moment name: pivot@R633)
Filter seq : 'a [+ant,+strid,-cont] ə
Size of subset : 7;
0.507% of whole
Accuracy on subset with sequence 'a [+ant,+strid,-cont] ə in pivot@R633 : 0.0%
Percent of errors included in subset: 3.431372549019608%

```

Figure 2: A context autopsy, one of the diagnostics offered by DiaSim. In this case, we see that the error is likely conditioned on a following  $\widehat{ts/}$ .

would be of a proposed change to the cascade, among others. The diagnostics are enumerated and described in the diagnostics README file in the GitHub repository.

Where information concerning phone-wise errors is concerned, an alignment algorithm based on minimizing feature edit distance (Mortensen et al. 2016) is used in calculation of phone-wise error to ensure maximally correct mappings between corresponding forms. By allowing very specific stipulations on analysis, DiaSim aims to help pinpoint where in the sequence of realized shifts the critical error occurred. For example, the final stage error correlated with a particular phone measures how much error arises from failure to properly generate it or its effects on neighbors. The same statistic observed for an earlier pivot stage would instead indicate how much inaccuracy comes from errant handling of its future reflexes and their behavior. Meanwhile, error correlated with the resulting phone for an earlier “pivot” stage could instead reveal the degree of error propagation caused by errant generation of the said phone at the pivot stage. Likewise, when analyzing specific errors between the gold standard and the result, DiaSim can pinpoint for the user if the type of error happens to be particularly common in certain contexts.

DiaSim’s diagnostics can be useful for identifying the regularity of the contexts of a phenomenon that may have otherwise appeared sporadic or inexplicable. Given that DiaSim, unlike previous models, is explicitly modeled using phonological features, it is well-equipped to identify phonological regularity that humans could easily miss. We discuss a number of these in our later sections, illustrating how DiaSim can capture diachronic patterns that humans miss, even for languages with extensive academic study spanning over a century and a half, such as French. A number of screenshots of DiaSim’s diagnostic capabilities, such as the context autopsy, as used to correct a specific error are showcased in §8.1; all diagnostic functions are described in greater detail in DiaSim’s publicly available documentation.

### 5.3 Consistency with longstanding theory

DiaSim was built with the explicit aim of faithfulness to longstanding theory in diachronic phonology while also being flexible enough to accommodate different frameworks. It assumes words consist of token instances of a bounded set of segments (in addition to *juncture phonemes*: word and morpheme boundaries), and that segments are uniquely defined by values given for each of a constant set of features (Chomsky & Halle 1968; Hall 2007; Hock 2009). Each feature can take one of three values : *positive* (+), *negative* (-) or *unspecified* (0). Which features from the global set are relevant varies with the language being analyzed, and sometimes there are different feature paradigms to choose from, so DiaSim empowers the user to make a custom set features (and phones)

if they wish, while also providing them with a default set.<sup>10</sup>

## 6 Datasets

### 6.1 FLLex

The dataset **French from Latin Lexicon** (*FLLex*) consists of 1368 pairs of Classical Latin etyma and their modern French reflexes, made for simulating regular continuous derivation from Latin to French. These 1368 samples include all 1061 inherited etyma used in Pope (1934) except for some verb forms, as well as 307 additional etyma recruited from Rey (2013) and the TLFi.<sup>11</sup>

For inclusion, lexemes had to have been in continuous usage throughout the entire modeled period. Words affected by non-phonologically motivated phenomena such as analogy were excluded, but words with apparent irregularity that could not be attributed to such processes (cases of sporadic metathesis, etc) *were* included. As *FLLex* is a *French from Latin Lexicon*, loanwords entering the language after the regional Latin developed into Gallo-Romance are excluded.<sup>12</sup>

Each entry was checked with the multiple sources (Pope 1934, Rey 2013, the TLFi) to ensure it developed *continuously* from Latin to modern French, *sans* interference from non-phonological factors. When the Popular Latin form a French words hails from is not attested in Classical Latin, we consult the listed resources for the appropriate form. Latin to French *inherited etyma* include loans predating the divergence of Gallo-Romance branched off, and excludes loans afterwards.

### 6.2 FLLAPS

The period-informed dataset *French from Latin Lexicon by Attested Period Sublexica* (FLLAPS) is recruited from Pope (1934)’s “Sound Tables” section. This set has an intentional degree of balance in phonological coverage, as the set compiled by Pope was assembled explicitly so it that had at least one word to demonstrate every notable sound change. In comparison, *FLLex* is more representative of the overall diachronic phonemic frequencies of the French language, contrasting with the relatively even phonological distribution present in this set. Whereas the global set has a gold standard set only for the final result, this set offers gold standard forms derived from Pope (1934)’s philological work for four intermediate stages, including Late/Popular Latin (postdated to circa 400 CE), Old French I (circa 1100 CE), Old French II (circa 1325 CE), and Middle French (circa 1550

---

<sup>10</sup>DiaSim, by default, operates in the traditional paradigm of exclusively phonetically motivated rules, but it is also able to store and access information on any word’s lexical paradigm, semantic domains, and lexical class, so factors operating across these attributes can be implemented with minimal change to the code.

<sup>11</sup>An online resource, the *Trésor de la Langue Française informatisé* (Dendien & Pierrel 2003) of the public research laboratory group ATILF (*Analyse et Traitement Informatique de la Langue Française*) and the agglomeration CNRTL (*Centre National de Ressources Textuelles et Lexicales*), maintained by specialists from the Académie Française and various French universities. Note that, for brevity, citations of the TLFi are indicated as “ATILF, TLFi:” and then the specific lexeme.

<sup>12</sup>An anonymous reviewer notes the importance of Germanic etyma in tracing various developments in French; we agree and would note also the importance of words from other Romance languages, especially Occitan. Future progress could use a cascade with multiple chronological inputs to incorporate such etyma into consideration

CE).<sup>13</sup> These stages were chosen by Pope for her table, but they neatly align so that we have a fairly large philological corpus at each point except for Late Latin. A few corrections were made in order to adapt the set for this task. For example, as Pope could not have foreseen the computational use of her set, for segments that were not of interest to the specific sound changes being demonstrated by the table in question, she sometimes omits finer distinctions (such as lax/tense distinctions). In these cases, the sound changes described elsewhere in her chart for the period in question were regularly applied and consistency enforced.

Because the periodic dataset aims to be phonologically balanced, if one of the words in Pope (1934)’s diachronic period tables was found to have grounds for exclusion (elaborated in the file *ExclusionsFromPeriodData.txt* on the GitHub repository linked further above), it was replaced with an acceptable word that also had the sound pattern that Pope indicated the word represented.

## 7 Rule cascades

In order to demonstrate both how DiaSim simulates long-term and holistic diachronic Neogrammarian sound change, we make use of Pope (1934), which enumerated all the major developments between Latin and modern French and remains used as a baseline for current work in the field (Short 2013). From this work, we derive the *BaseCLEF* dataset, while the *DiaCLEF* dataset was built by exhaustively correcting non-sporadic errors detected using DiaSim’s simulation and evaluation functionalities.

Although Pope (1934) is comprehensive, it was not without errors and omissions. Much of the later work in French historical phonology has built off Pope’s account and revised it in various ways (Buckley 2003; Mazzola 2013; Rochet 2015; Wernicke-Heinrichs 1996). In addition to demonstrating how a Pope-derived ruleset works (or any analog for another diachronic scenario), our methodology can also be used to test the proposed revisions to it, revealing remaining omissions and misorderings that even trained phonologists can miss when arduously combing through data, as seen in the superior accuracy of the *DiaCLEF* rule cascade built using DiaSim.

### 7.1 BaseCLEF

The **Baseline Classical Latin Etyma to French** (*BaseCLEF*) cascade includes all regular sound changes posited in Pope (1934) in the order specified. Where Pope’s writing is ambiguous, especially where relative ordering is concerned, the benefit of the doubt is given as a general policy (that is, the interpretation that results in the outcome matching forms that Pope cited is assumed). Additionally, there are a number of cases where literal interpretation of Pope (1934) engenders errors that are not “interesting”, as they arise from omissions that at the time of writing were not essential, perhaps because Pope didn’t foresee her work being converted into an explicit rule cascade.

<sup>14</sup> These trivial omissions are corrected in the cascade *BaseCLEF\**; results for BaseCLEF\* are

<sup>13</sup>These assigned dates mark the point at which the changes assigned to the period are considered by Pope (1934) to have *resolved*. Although some may in fact fall outside of what philologists consider the span of that period to be, this allows us to test the accuracy of Pope (1934) within her own framework.

<sup>14</sup>For example, Pope states that modern French lacks any phonemic length difference, but never states when it was lost. Another example concerns the fate of /r/: Pope notes that it becomes uvular, but never explicitly states that it goes

presented separately.

## 7.2 DiaCLEF

The corrected **DiaSim**-informed **Classical Latin Etyma to French** (*DiaCLEF*) rule cascade was built by using DiaSim with our empirical procedure (see §4) to iteratively modify, split, merge, delete, add, and reorder rules to optimize performance over the dataset.

To illustrate how rules were corrected to make DiaCLEF, we present the correction to Pope (1934)’s rule §187iia, which holds that when intervocalic, Latin *v* (then [ɣ<sup>w</sup>] per Pope (1934))<sup>15</sup> is deleted before a round vowel. The rule as formulated by Pope (1934)<sup>16</sup> is thus:

$$(2) \text{ } \gamma^w > / [+syl] \text{ } \text{---} \begin{bmatrix} +syl \\ +round \end{bmatrix}$$

However, it turns out that we observe a pattern by which final *-f* (the usual result of *v* in the coda) is erroneously absent in the observed results, and using the lexical derivations, we can trace the origin of the error consistently to this rule. The apparent exceptions to this rule include *vīVUM* /ɣ<sup>w</sup>i:ɣ<sup>w</sup>um/ “alive” > · · · > ⟨*vif*⟩ “lively”, *NĀTĪVUM* “original, by birth” /n̩ɑ:t̪i:ɣ<sup>w</sup>um/ > · · · > ⟨*naif*⟩ “gullible”, and *ōVUM* /o:ɣ<sup>w</sup>um/ “egg” > · · · > ⟨*œuf*⟩ “egg”. One could explain this as three independent different cases of “restorative analogy” rather than a regular phonological rule, noting *vĪVERE* /ɣ<sup>w</sup>i:ɣ<sup>w</sup>ere/ “to live” for *vĪVUM*, *ōVIS* /o:ɣ<sup>w</sup>is/ “sheep” for *ōVUM* (and etc) wherein formally identical stems were not regularly affected by the rule. However, one must also entertain a regular phonological explanation. In fact, every word with a sequence of /-ɑɣ<sup>w</sup>-/ is unproblematic where this rule is concerned, with *AVUNCULUM* /ɑɣ<sup>w</sup>ʊŋklum/ “uncle”, *FLAVUM* /fl̩ɑɣ<sup>w</sup>um/ “blond”, *CLĀVUM* /kl̩ɑ:ɣ<sup>w</sup>um/ “nail”, *PĀVŌNEM* /p̩ɑ:ɣ<sup>w</sup>o:nem/ “peacock” and behaving exactly as they should by default. Thus, one can render a result where those words for which the rule is problematic are *regularly* excluded, by specifying that it only operates on prior context of [+syl,+lo], rather than just [+syl].

Corrections like the one above were made using DiaSim’s diagnostic functions, and in conjunction all of these corrections formed the DiaCLEF dataset. Many of these corrections are of interest to scholarship on French historical phonology; we discuss a subset of these in §8. We found that many of these independently supported the findings of scholarly work subsequent to Pope (1934), even though DiaCLEF was made before consulting this literature.

## 8 Results and discussion

As seen in Table 1, the increase in accuracy from “debugging” the cascade using DiaSim is striking, with raw accuracy going from 3.2% to 84.9%. The improvement in average feature edit distance,

from a trill to a fricative. Yet she does cite modern reflexes as having a fricative, not a trill, in her periodic tables. If taken literally, the base ruleset would automatically be wrong in its final result form for any word with an *r*.

<sup>15</sup>§109ia (p56): “a fricative bi-labial with velar modification...”

<sup>16</sup>§187 iia (p91) operates “(a) when intervocalic before the labial vowels *o* and *u*, (b)...”. Here and elsewhere, we follow Pope’s lead in treating Latin *v* as originally representing the *fricative* [ɣ<sup>w</sup>]. Although other authors may not be in agreement, this matter is immaterial for the accuracy of the final outcome.

Table 1: Performance on FLLex

Metric	BaseCLEF	BaseCLEF*	DiaCLEF
Accuracy	3.2%	30.3%	84.9%
Accuracy within 1 phone	26.3%	55.7%	94.8%
Accuracy within 2 phones	56.7%	79.9%	99.1%
Avg Normalized Levenshtein Edit Distance	0.518	0.380	0.056
Avg Normalized Feature Edit Distance	0.673	0.392	0.061

Table 2: “Accuracy” is the percentage of forms in the set hypothesized by DiaSim that exactly match the reference set. Accuracy within one or two phones allows matches that differ by at most one or two phones respectively. LED and FED are defined in §5.1.1.

a decrease from 0.518 to 0.056, is also quite large. This difference is still large even when we consider the baseline to be BaseCLEF\*, with its accuracy of 30.3% and mean feature edit distance of 0.380. In the difference between BaseCLEF and BaseCLEF\*, one can see the magnitude of the effect of correcting errors that are “uninteresting” in as much as they do not reflect significant revisions to Pope (1934)’s understanding of French historical phonology. This is less than the difference between BaseCLEF\* and DiaCLEF for all metrics; this difference represents the magnitude of quantified improvement by making the “interesting” corrections. BaseCLEF\*’s improvements over BaseCLEF in FED are more modest than the improvement in accuracy and also slightly larger than its improvement over phone edit distance (PED). This makes sense, considering that fixes in BaseCLEF\* typically involved changes of single features (e.g., whether /r/ is a trill or a fricative — a difference of the single feature [±son], or differences in vowel length) — which meant the relative magnitude of corrections was greater at the segment level (as measured by accuracy and PED) than the feature level (as measured by FED). Overall, mean feature edit distance may be the most indicative measure with respect to gold standard and result pairs that share the same number of segments. However, otherwise, this is likely not true, as it punishes insertions and deletions much more than substitutions, considering them to be “changes in each feature” — a policy that is hard to justify theoretically especially when speaking of the quite widespread deletion and insertion of segments such as reduced vowels or word-final voiceless stops (an observation quite relevant given that French has repeatedly and regularly experienced both (Martinet 1952)). One could expect the highest sensitivity to this issue to be located in Gallo-Roman, the stage during which the most extensive deletion and insertion of phonemes occurred. Future work might enhance FED by taking into account typologically likely insertions and deletions when calculating distance between phonetic sequences.

In Figure 3, we see a breakdown of the sorts of corrections that were done for DiaCLEF (excluding those also handled in BaseCLEF\*). Cases of *rule deletion* are those where our empirical procedure with DiaSim led to the deletion of a rule stipulated in Pope (1934) as that was deemed a better fix than modifying the rule. Cases where modification led to better performance are counted under *amend existing rules*. Counted separately from any modification on the same rule, any changes in the relative chronology, including the insertion of rules that were mentioned but whose time of operation were left ambiguous in Pope (1934), are counted under *re-orderings* (the internal breakdown

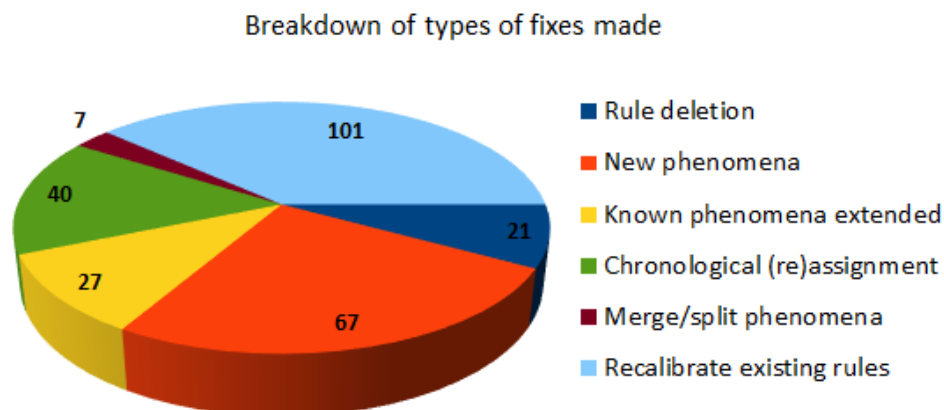


Figure 3: Breakdown of all corrections made in the creation of DiaCLEF by type. Note that whether phenomena are “new” is relative to Pope (1934), so many of our “new” rules in fact ended up confirming prior scholarship.

for these between forward movement, backward movement, and new chronological assignment, also differentiated by period, is depicted in Figure 5). The relatively minuscule category of corrections that involved either merging or splitting phenomena are counted separately. Meanwhile, the results of cases where the best solution was deemed to be the creation of new rules are broken into two categories here: those that *extended known phenomena* (such as a later application of a known intervocalic voicing rule), and those that stipulated entirely *new phenomena*, the latter being the more radical change. As one sees in Figure 3, this most radical sort of correction constituted about a quarter of all corrections. Together, the less radical sorts of corrections, those being re-calibration, mergers/splits, and extension of known phenomena, constituted 135 of the 263, or 51.3%. This leaves just a little less than half of the corrections in the relatively radical types of new phenomena insertion, reordering, and wholesale rule deletion.

However, there were notable periodic differences with regard to where changes fundamentally challenge Pope (1934)’s understanding of French diachrony led to meaningful improvements. As seen in Figure 4, the biggest volume of changes occur in the Gallo-Roman and Old French periods. This is also true of re-orderings, which are broken down by period and subtype in Figure 5. On the other hand, few changes were necessary for the transition from Classical Latin to Late Latin, and even fewer were necessary for early modern French. On the surface, it would appear that the same is true for the Middle French period. However, while it is true that few changes were made within Middle French, it should be noted that a number of highly significant phenomena were reassigned from Middle French to Old French, and thus counted in both Figures 5 and 4 for Old French because that is where they were inserted, as was our consistent policy for counting re-orderings. Overall, it is nevertheless correct to say that fixes were disproportionately necessary for Gallo-Roman and Old French.

This should come as no surprise. The Gallo-Roman period (except in its very latest stages) is by far the least well-attested – and therefore, the most like what we would be dealing with if we were working with an understudied indigenous language. Given that Pope (1934)’s philological

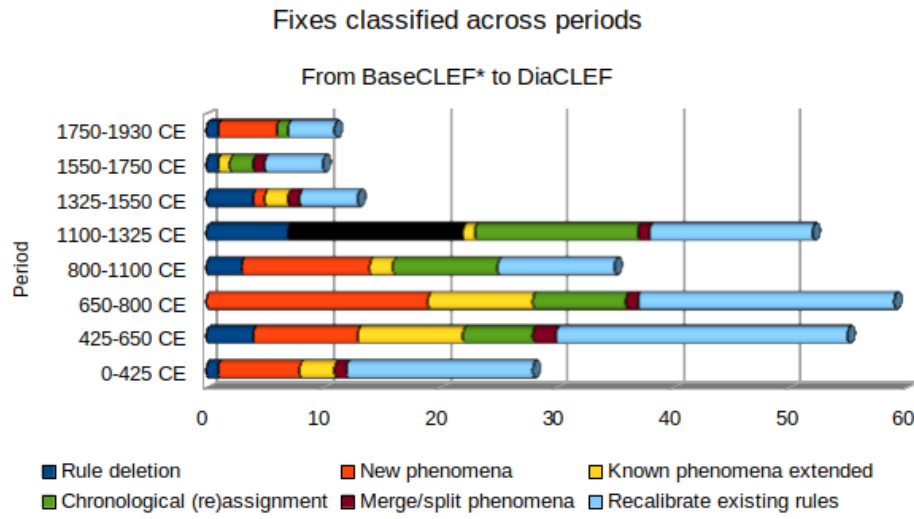


Figure 4: Differences between different periods in number and type of edits made to the cascade, with changes included in BaseCLEFstar not counted.

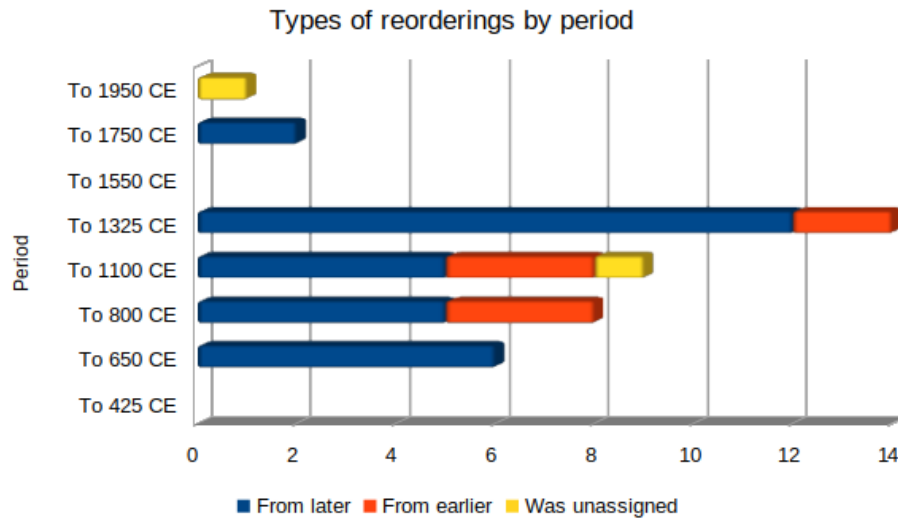


Figure 5: Corrections of ordering by period.

```

Result phones most associated with error:
0: /k/ with rate 2.3333333333333335, Rate present in mismatches : 24.13%
1: /p/ with rate 1.4444444444444444, Rate present in mismatches : 14.94%
2: /s/ with rate 0.5645161290322581, Rate present in mismatches : 40.22%
3: /ə/ with rate 0.5, Rate present in mismatches : 1.1494252873563218
Gold phones most associated with error:
0: /p/ with rate 1.5555555555555556, Rate present in mismatches : 16.09%
1: /k/ with rate 1.4444444444444444, Rate present in mismatches : 14.94%
Focus point phones most associated with error:
0: /a/ with rate Infinity, Rate present in mismatches : 1.1494252873563218
1: /tʰ/ with rate 1.0, Rate present in mismatches : 1.1494252873563218
---
Most common distortions:
----
Distortion 1: k for g
% of errant words with this distortion : 8.0459%
Most common predictors of this distortion:
No constant features for pre prior
No particularly common pre prior phones.
Percent word bound for prior: 100.0
posterior phone constant features: -syl -nas -sg -cg -lab -hi -lo -front -l
Most common posterior phones: /ʊ/ (85.7%)
post posterior phone constant features: -cons -lat -nas -strid -sg -cg -anl
Most common post posterior phones: /a/ (71.4%)
----
Distortion 2: e for ε
% of errant words with this distortion : 6.8965%

```

Figure 6: DiaSim’s Confusion Diagnosis in a scenario where the lack of representation of velar onset voicing in the cascade is the most problematic remaining error.

approach leaned heavily on corpus data and meta-linguistic commentary by historical writers, it is only natural that accuracy would suffer the most in the period where these were least available.

The raw numbers of edits made to the cascade do not scale to their impact. In fact, in terms of effect size, there was a disparity. The majority of rules were dwarfed in impact by a set of about twenty rules which each, alone, changed the accuracy by around one percent or more. Most of these were either re-orderings, the deletion of errant rules, or the proposal of very simple rules that usually lacked any fine conditioning. On the other hand, fixes that served to extend or amend known phenomena tended to have low individual effect sizes, but were more numerous overall.<sup>17</sup> There were also many rules that had a very large effect on the accuracy of intermediate stages, but not on the accuracy of the final outcome. In the following sections, we examine some of the relevant phenomena in greater detail.

## 8.1 A regular account of “sporadic” k-voicing

In this section, we show how CFR helped us discover a regular account for the voicing of Old French velar stop onsets, a phenomenon for which — in a century of work — no regular explanation had been offered, with a myriad of supposedly independent individualized explanations for each relevant word, “missing the forest for the trees”.

<sup>17</sup>Because these were much less meaningful departures from the baseline, their number is also inflated by a somewhat higher tolerance to rules with small effect sizes for rules in this category on our part; we much more zealously enforced the idea of being sure the effects were significant for phenomena that were not at all present in the baseline (“new phenomena” — many of these having been independently produced by scholarship subsequent to Pope (1934)).



```

6
What results would you like to check? Please enter the appropriate number:
| 0 : Print stats (at evaluation point) (for subset lexicon if specified)~~~~~|
| 1 : Print all corresponding forms (init(,focus),res,gold) (for subset if specified)|
| 2 : Print all mismatched forms at evaluation point (for subset if specified)|
| 9 : Exit this menu._____|
2
Printing all mismatched etyma for filter # k @ [+lo] at In
Res : Gold
/kle/ : /glev/
/kla/ : /gla/
/klef/ : /kle/
/kwas/ : /gwas/
/kwas/ : /gwas/
/kwes/ : /gwes/
/kwa/ : /gwa/
/kwa/ : /gwa/
/kwaj/ : /gwij/
/kavvuv/ : /kavfuv/
/ke/ : /ket/

```

Figure 7: All mismatched etyma with the filter sequence at the input stage (i.e., the Classical Latin forms start with CLA- or CRA-)

First, we consult DiaSim’s *Confusion Diagnosis* (Figure 6). In the top of the figure, the left column shows the phones with the highest ratio of occurrence in error cases to correct cases, and the right column shows the percent of all error cases having the phone. Below this, the *distortions* that were responsible for the most errors are presented.<sup>18</sup> Here, [k] is the phone most correlated with error, and the most problematic distortion is /k/:/g/, observed [k] for what should be /g/, accounting for 8% of all errors. This strongly suggests that a regular rule is at work, with its conditioning environment evident from the contextual statistics: 100% of /k/:/g/ errors occur just after the word onset, 86% are before the uvular fricative /ɣ/, and 71% have /a/ two positions later.

This suggests that an onset velar voicing happened at some point, but we don’t know when. Now we want to find *when* this occurred. We try to filter out “noise” by using a subset of the data defined with the aid of the observed statistics. In DiaSim, we set a *focus point*<sup>19</sup> as the input form from Classical Latin, and use a *filter sequence* of onset k with a low vowel two positions afterward.<sup>20</sup>

The resulting subset’s error cases (Figure 7) all have /k/ between the word onset and a consonant. All the consonants in question are sonorants, but we continue to condition our rule broadly on consonants, since never observe a non-sonorant consonant at all. If we add a low vowel afterward as a new constraint, we now perfectly predict the /k/:/g/ distortion, with one unrelated exception.<sup>21</sup>

Filtering the data to only etyma with the Latin sequence “# k [+cons] [+lo]”, the data subset has

<sup>18</sup>These calculations are done using an alignment algorithm that minimizes *Feature Edit Distance* (Mortensen et al. 2016) between aligned phones. FED Alignments are less accurate when pairs are so perturbed that the error is opaque, or in cases of long distance conditioning, but these are rare. The goal of CFR is to quickly find the simplest possible fixes for the current misunderstandings that affect the largest number of etyma, so these sorts of complex fixes are not the sort our methodology is aiming for.

<sup>19</sup>The *focus point* is the time step at which the dataset is filtered using the *filter sequence*.

<sup>20</sup>In DiaSim’s fairly accessible syntax, this is “# k @ [+lo]”. In fact, “@”, meaning “any single phone” is the only symbol in use for this function, which can also support complicated SPE notations including alpha functions, disjunction, and parenthetical optional segments.

<sup>21</sup>The <clef>/<clé> doublet, reflexes of CLĀVEM, with a low vowel. One can’t be accurate for both *clef* and *clé*.

```

k l 'a: r a m
#kl'aram# | Rule 58 : [+syl,+long] > [-long,-splng]
#kl'ara# | Rule 74 : [+nas,+cons] > ø / [-stres] ___ #
#kl'a:ra# | Rule 116 : [+prim] > [+long] / ___ [+cons] [-cons]
#kl'a:ra# | Rule 205 : [+lo] > [+front,-back]
#kl'aəra# | Rule 420 : {'a:;'e:;'o:;'ε:} > {'a ə;'e j;'o w;'i ə}
#kl'aəra# | Rule 447 : a > ə / [+syl] ( [-syl] )*
#kl'e:rə# | Rule 554 : {a ə;'a ə;'a ə} > {'e:;'e:;'e:}
#kl'ε:rə# | Rule 612 : {'e:;'e} > {'ε:;'ε} / ___ [+cons] [+syl]

```

Figure 8: Derivation of CLĀRAM >...>⟨*claire*⟩.

```

In: delete & filter by input
Out: delete & filter at current output
Gold: delete & filter by current gold
U: delete and also delete filter
R#: right before rule with index number <#>(you can find rule indices with option 3
Please enter the appropriate indicator.
R461
On rule number 0
On rule number 100
On rule number 200
On rule number 300
On rule number 400
Size of subset : 7;
0.508% of whole
Accuracy on subset with sequence # k [+cons] [+lo] in pivot@R461 : 0.0%

```

Figure 9: In DiaSim’s Error Analysis, we isolate our error by setting our focus point to be the instant after the last rule *bleeding* the observed velar onset voicing phenomenon.

well under 50% accuracy. All of the non-error cases in this subset have also changed Latin A into a non-low vowel, and in all of these,<sup>22</sup> the A had primary stress and was in an open syllable. The same is true of only one error case.<sup>23</sup>

Having determined the conditioning of the development, we now want to find when it occurred in relation to other rules, to amend the error by placing a new rule in the cascade. To see if it was indeed *bled* (Kiparsky & Good 1968) by vocalic changes, we examine the derivations of these exceptions. Figure 8 shows the derivation for CLĀRAM >...><sup>24</sup> ⟨*claire*⟩. We see that the critical point of bleeding is rule 554, as the low vowel is lost (/aə/ > /e), hence why we CLARA retained onset /k/ (and likewise for CLĀRUM and CLĀVEM). The derivation of CLĀVUM >...>⟨*clou*⟩ reveals a similar bleeding effect: /aw/ passed to /ɔw/. As these both bleed the new rule, it must be placed after them.

Now that we have a proposed rule, proposed conditioning for that rule, and an idea of when it occurred relative to other rules, we must test it. Is this what the data supports? As seen in Figure 9, we set our focus point to the point after rule 461 to filter out the words affected by those two

<sup>22</sup>In FLLex, these consist of CLĀRAM, CLĀRUM, CLĀVEM and CLĀVUM

<sup>23</sup>Namely, ⟨*glaiue*⟩, which should not be the basis of any argument, given the uncertainty about its etymology (ATILF, *TLFi*: *glaiue*, (Fouché 1952/1966:602), (Rheinfelder 1975:283), von Wartburg t.4:144–146).

<sup>24</sup>Here and henceforth we use “>...>” to indicate an opaque long term diachronic equivalence which has a multitude of intermediate steps, in line with Janda & Joseph (2003)

bleeding rules. Now we have a data subset defined by a single and identified source of error and an idea of the timing for our corrective rule. As expected, our accuracy on that subset is zero. Given that we have now isolated the source of error, we know what rule must be placed to fix it: rule (3). Our proposal will be validated if, after doing so, accuracy dramatically improves.

(3)  $k > g / \# \text{ \_\_\_ } [+cons] \text{ \_\_\_ } [+lo]$

Surely enough, after placing this rule after the last rule bleeding it (i.e.,  $a\epsilon > e:$ ), we achieve perfect accuracy for all etyma in the subset except one.<sup>25</sup>

We now consider our proposal in light of relevant scholarship, which until this point was not considered. Pope (1934:69) is likely correct that there was at one point a *synchronic* tendency of such a form, but she seems to have missed that a *diachronic* effect became phonologized later. It is easy to see this phenomenon in the context of earlier lenition processes in French, which parallel those in most Western Romance and British Celtic languages, whereby stops that were either intervocalic or in an intervocalic stop + sonorant cluster were voiced, often as a precursor to spirantization (Martinet 1952). While some attributed the process to contact from Celtic (Martinet 1952), there is also an internal motivation proposed and possible shared developments in Sardinian and Neapolitan (Cravens 2002); in Western Romance, such internal factors may have been boosted by contact with Celtic. Whatever the case, attested data suggests lenition was indeed a solely phonetic process at first where it occurred within spoken Latin, operating without regard to word boundaries (Adams 2013:183).

“Lenition” was phonologized intervocalically relatively early, but vestiges of it continue to operate across word boundaries even today in both Ibero-Romance and Insular Celtic (Martinet 1952). The general lack of “lexical autonomy” engendering (among other things) porous word boundaries that French continues to share with Insular Celtic and Ibero-Romance is well known, and manifests in sandhi phenomena such as liaison in French still today (Cerquiglini 2018; Wehr 2001). Thus, a vestige of lenition quite plausibly remained active where it had not been phonologized: at the word onset.

The specificity of the context (posterior sonorant + low vowel) for the diachronic results need not be an issue: recall the evidence that such a process in the onset remained productive synchronically more broadly (Pope 1934:p. 96), in light of the theoretical grounding for synchronic processes leading to diachronic outcomes wherein their original phonetic motivations are made opaque through changes in scope (Janda & Joseph 2003). There is further typological support for a change in scope manifesting specifically this way within Romance. Lenition processes targeting stops in Romance disproportionately effect velars (Recasens 2002), and this is especially so for voicing at the onset (Figge 1966), as is the case here. While the lack of voicing in initial  $/\#ka/$  segments may appear surprising at first, it is less so when one considers that by this point, they had all regularly become  $/tʃ/$  through Gallo-Romance palatalization. The role of following  $/a/$  likewise must be considered in the context of known analogous preferences for lower vowels as conditioners of modern lenition phenomena in Spanish and Catalan (File-Muriel & Brown 2011; Simonet et al. 2012). This also

<sup>25</sup>The exception is  $CR\bar{A}T\bar{I}CULAM > \langle grille \rangle$ , due to irregular hiatus behavior after the loss of  $/\delta/$  (from intervocalic  $\tau$ ). This suggests something else to fix, not that our otherwise well corroborated proposal is wrong; indeed, as we shall see later, the Old French corpus shows that  $CR\bar{A}T\bar{I}CVLAM$  was *graille* at this point, and voiced in parallel with the other etyma here.

means the process became *phonologized* after both this change in scope and late enough to be bled by the raising of former /a/ in words like <clou> (< CLAVUM) and <clore> (< CLAUDERE), as well as those like <claire> (< CLĀRAM, <clef> (CLAVEM), and so forth.

Curiously, at the same time, the voicing rule may plausibly be dated right around the beginning of final consonant deletion in French<sup>26</sup>, meaning that many onset clusters would newly become intervocalic where previously they hadn't been.<sup>27</sup>

Our proposed rule 3 is also quite robustly supported by Old French data attesting it mostly in the 12th century — which happens to also be when the relevant final consonant deletions began. This is seen in the FEW (von Wartburg 1922–2002:t. 2,16) for a slew of etyma (Marr, in preparation), including seven more inherited etyma not in FLLEX, and many derived words.<sup>28</sup> There is further support from Germanic etyma which had entered the language early enough.<sup>29</sup> Exceptions are largely limited to two most probably analogized lexical classes,<sup>30</sup> or otherwise explicable through onomatopoeia,<sup>31</sup> leaving only two problematic cases which are also possible though less certain cases of onomatopoeia<sup>32</sup>

However, the traditional view has considered this at most a sporadic and synchronic Gallo-Roman era “tendency”,<sup>33</sup> where each diachronized case is explained independently. (Bourciez & Bourciez 1967:p. 142) and (Nyrop 1899:p. 375) both note the frequency of pre-liquid initial velar voicing but leave it unexplained except for chalking *gras* up to analogy from *gros*. The case of *glas*

<sup>26</sup>Final consonants /l/, /s/, /z/, /x/, /w/, and /θ/ were deleted for various reasons but a result was a pattern of more and more vowel final words.

<sup>27</sup>Specific lexemes that tend to precede nouns are relevant here as they first became pronounced with a vowel-final coda during this period: the conjunction ET (<e θ/), the prepositions <à> (< /aθ/), the articles <ce> (< <cel> < ECCE ILLUM), <du> (originally with -l), <ceci> and <ci> (< /tsix/ < ECCE HĪC), and <cela> (the latter part from là < /lax/ < ILLAC).

<sup>28</sup>CRĀT-IS > · · · >grate, CRĀDALEM > · · · >graal, CRĀTĪCULAM > · · · >gradille > grille's related but separately inherited gradil > graīl > gril, \*CLAREA > · · · >glaire, and CRAXA- (< Gaulish) > · · · >graisset, plus their derivations. \*CRE:DENTARE > · · · >greanter, graanter, following the same lowering seen in words like REDEMPTIŌNEM > · · · >raançon (> rançon), likewise seen also in its derivations (creantement > grantement, creant > grant, etc); lastly, though slightly less robustly attested, is CREĒPANTĀRE > · · · >cravanter (11th century) > gravanter (1213). Derivations for these before the late 12th century are largely coined in *cr-* but those after 1200 are formed in *gr-*, even when the original phonetic motive is gone, suggesting it is now phonologized.

<sup>29</sup>Old Frankish \*krappo > · · · >crape (11th cent) > grappe (1121), \*kratt-ôn > · · · >gratter, \*kraw-jan > · · · >gravir, \*krawa > · · · >grau, groe > Old Picard grauwet, corresponding to later attested French grevet; Old Norse kraf-la > grafi(g)ner, with the same change in how derived words are formed as well.

<sup>30</sup>One: words related to CLĀRVUM, itself regularly *cler* via the known bleeding effect on open A: OF *clarté*, *cleron*, *clarion*, *claret*, *clairret*, and *claror*, all loans or later coinages from before 1100. Although \*CLAREA was also derived from CLĀRVUM, its semantics (“egg white”, “mucus”) diverged enough to leave the “family”. Two: those related to CLAVEM and CLAVUM, both also regularly still voiceless due to aforementioned bleeding phenomena: *clacielle*, *clacelier*, *clavel* > *claveau*, *clavedure* > *claveure*, *clavon*, and *clavette*.

<sup>31</sup>*crachier* “to clank”, *cras* “the cry of a crow”, *claque* “a smack by the hand”, *clap* “a slap, a hit”, and *clapier* (originally “stone heap”, before tabooistically becoming “brothel”, but cf Old Bourguignon *glapier*). Reflexes of Old Frankish \*kramp “spasm, cramp, twist” remain voiceless in French but widely become voiced in the regional *oïl* varieties.

<sup>32</sup>These are CLAMŌREM “a shout” > · · · >clamour > clameur “loud noise, outcry” and CLAMĀRE “to cry out” > clamer, “id.” Aside from onomatopoeia, influence from German *klagen* has also been proposed.

<sup>33</sup>(Pope 1934:p. 96), for example, considers it to be non-regular, and exemplifies the “tendency” with the substitution of “Glodoveus” for “Chlodoveus”.

alone has been attributed, to assimilation to the following voiced sonorant by one source (ATILF, *TLFi*: glas), but other cases have been explained individually and independently of this. Sometimes these individualized explanations are peculiar: voicing in \*CLAREA > · · · > *glaire* (“mucus”, “egg white”) is explained by influence from the word for “gravel” (FEW t2 p738b). What is more likely: *each* of these independent, unfalsifiable and sometimes contrived lexical solutions being true, or the “trees” being part of a forest with a falsifiable explanation?

An anonymous reviewer worries about the possibility of “abracadabra” rules that improve downstream accuracy without necessarily matching the true historical developments – “overfitting”, essentially. We acknowledge this risk, which is why we tested our approach using CFR on French, a language whose history is among the most well studied. Here, we feel vindicated: a new rule is found that boasts support in the corpus and a quite plausible explanation in typology, with notable support in closely related languages. On the other hand, the hyper-individualized explanations for each affected word are *each* possibly coincidental. And since analogy and the like can easily work in tandem with regular sound laws to produce the same results, in order to actually make a proposed regular rule untenable, *each* individualized explanation must be correct, even the weakest links like the supposed influence of “gravel” upon *glaire* “egg white, mucus”. Of course, each of these is actually unfalsifiable, while a regular rule’s placement in the cascade can be falsified if it fails to produce the desired results without errant side effects. Still, “overfitting” remains a danger; thus, future work should incorporate a measure of *overall complexity* as discussed in §5.1.1. Even so, we maintain that while traditional approaches have turned a blind eye on the proliferation of lexically specific explanations at the expense of “seeing the forest”, adopting CFR alongside traditional methods will lead to simpler and more robust explanations.

## 8.2 Major re-orderings

A large share of changes made in the construction of DiaCLEF involved reordering existing rules — suggesting that the baseline derived from the work of Pope (1934) had mostly the right ideas about what happened, and often also the right general idea about the dates of the sound changes postulated, but not the right ordering of sound changes. One aspect of such re-orderings is that distinctions are lost earlier, or they are retained later (such as in §8.3).

### 8.2.1 Alveolar deaffrication counterfeeding vowel lengthening

In many cases, the solution we arrived at had already been proposed in past works. Such is the case with regards to the timing of the merger of Old French  $\sqrt{ts}$  and /s/ in relation to the lengthening of /a/ in the suffix /asə/. The initial problem in this case was a pattern of errors among words containing the observed output phone /a/ is visible in Table 3. There, we see tabulated various diachronic suffix correspondences among etyma for which the modern output’s final syllable has /a/ in the nucleus. While in most cases shown in Table 3, the correct (gold) output (in the second column) matches the observed output (third column), this is not the case for those grouped in the second row. In the second row, we see the /a/ corresponding to a correct /a/.

All of these /a:/a/ error cases happen before an /s#/, but as the third row of Table 3 shows, there is also a corresponding case of etyma with observed outputs ending in /as/ which are *correct*.

Latin	Gold standard output	Observed output	Examples
AETĀTICUM	/aʒ/	/aʒ/	⟨âge⟩
-ACIA(M) -ACIAT -ACHIA(M) -ATTEA(M)	/-as/	/-as/ (ERROR)	⟨fasse, {f,gl,men}ace⟩ ⟨fasse⟩ ⟨brace⟩ ⟨{m,pl} ace⟩
-ASSA(M) -ASSĀS	/-as/	/-as/	⟨{b,gr,l,m,n} asse⟩ ⟨grasses⟩
-AQUILAM -ĀLIA(M) -ALIA(M) -ALEA(M) -ACULA(M)	/-aj/	/-aj/	⟨aille⟩ ⟨{au,bat,m,mur,piét} aille⟩ ⟨entrailles⟩ ⟨paille⟩ ⟨{m,ten} aille⟩
-ASTRA(M) -ASTER -ASTOR -ASTRUM	/-atʁ/	/-atʁ/	⟨marâtre⟩ ⟨pâtre⟩ ⟨saumâtre⟩ ⟨parâtre⟩

Table 3: Observed accuracy patterns for rhymes with /a/ as (early 20th century) output, before fixing the issue with the group starting with -ACIAM.

```
#m,əŋ'at'sə# | R534 : t's' > t's
#m,əŋ'at'sə# | R628 : [+syl,-front] > [+nas] / __ [+nas,-syl]
#m,əŋ'asə# | R648 : [+delrel] > [+cont]
#m,əŋ'asə# | R653 : {ə; ,ə} > {ə; ,ə}
#m,əŋ'a:sə# | R706 : [-round,+syl] > [+long] / __ s ə #
#m,əŋ'a:sə# | R708 : ə > [-syl] / __ #
#m,əŋ'a:sə# | R715 : [+lo,+long] > [+back]
#m,əŋ'a:s# | R736 : ə > ø
#məŋa:s# | R753 : [+stres] > [-stres]
#məŋas# | R754 : [+syl,+long] > [-long]
```

Figure 10: The (errant) by-word derivation of MINACIA > · · · > ⟨menace⟩.

As we shall see, although opaque in the present day, this difference in outcomes is predicted by a difference in context at an earlier stage. The /s/ seen in both the -ACIAM and -ASSAM groups is in fact the result of a merger of two formerly separate phonemes. For the most part, the orthography of French still preserves a fossil of this former phonemic contrast: the modern suffix *-asse* usually represents an earlier */-asə/*, while *-ace* usually indicates */-atsə/* at that time.

Before discussing our conclusion in more detail, and how it corroborates views that in this case already existed in the literature, let us first establish how we got here. As elsewhere, we performed our empirical procedure with DiaSim by considering all plausible possibilities, without any constraint from background knowledge or consultation with existing literature. We do, however, immediately rule out the possibility of a rule conditioned on being before a *modern* /s/, because that net would erroneously catch many correctly derived etyma as well.

We then look at the by-word derivation files on the error cases. We see the case of *menace* in Figure 10.

2 before	1 before	FOCUS	1 after	2 after
<0.3 thresh	<0.3 thresh	XXXXX	/t̥s/ : 0.60	<0.3 thresh
<0.3 thresh	<0.3 thresh	XXXXX	+strid : 0.36	<0.3 thresh
<0.3 thresh	<0.3 thresh	XXXXX	<0.3 thresh	<0.3 thresh
<0.3 thresh	<0.3 thresh	XXXXX	<0.3 thresh	<0.3 thresh

Figure 11: A context autopsy at a point in Later Old French reveals that the /a:/a/ error is predicted by  $\widehat{ts}$ , which is soon to vanish before its (correct) time. Displayed in each column are the error rates, in decimal form, for words with the targeted segment (in this case, /a/) in terms of what the surrounding sounds are. The segment operated on (*focus*) is marked with Xs, while the surrounding columns indicate the probability with which phonological features occur in errant etyma. Probabilities less than the threshold (an error rate of 0.3, or 30%) are omitted to not distract the user. In this case, the factors above this threshold are if the next sound is either a strident, or specifically the voiceless alveolar affricate /ts/.

```
Success: now making subsample with filter 'a [+ant,+strid,-cont] ə
(Pivot moment name: pivot@R633)
Filter seq : 'a [+ant,+strid,-cont] ə
Size of subset : 7;
0.507% of whole
Accuracy on subset with sequence 'a [+ant,+strid,-cont] ə in pivot@R633 : 0.0%
Percent of errors included in subset: 3.431372549019608%
```

Figure 12: Old French /atsə/ has a 100% output error rate (shown as a 0% rate of output accuracy).

Here, the relationship between /s/ and the backing of /a/ to /a/ becomes clear: a following /s/ is indeed the conditioning context for a rule that feeds the one of interest: instances of /a/ before /s/ become long /a:/ (rule 706), which would later retract to /a:/ (rule 715), before the ultimate loss of the distinctive length (rule 754). But we also see that in the case of *MINĀCIA* to *menace*, the /s/ in question had originally been a different phoneme, an affricate  $\widehat{ts}$ . Exploring the derivation and others further, one would find that this in turn was formerly palatalized (as  $\widehat{ts}^j$ ), and ultimately reflects a Latin *c* or *t* that was palatalized by a following palatal-articulated segment.<sup>34</sup>

Could it be that the merger of  $\widehat{ts}$  with /s/ (rule 648 in Figure 10 is feeding the lengthening of /a/ where it shouldn't be? To quickly determine this, we can place a *pivot point* in Later Old French, just before that merger. The results of context autopsy at this pivot point, seen in Figure 11, illustrate how it is in fact Later Old French /ts/ that is a very strong predictor of later errors in the reflexes of Old French /a/.

The sequence /atsə#/ in Old French has a significant frequency and a 100% output error rate for lexemes with it (Figure 12). This matches what we know of the data: words like *face*, *brace*, and others in the second row of Figure 3 have not /a/ but /a/ in 20th century French citation forms, including those provided by Pope (1934) and Rey (2013). This sets them apart from words like *grasse*, *nasse*, and the other examples in the third row of Figure 3, for which both the reference forms and the observed outputs agree in exhibiting /a/<sup>35</sup>.

<sup>34</sup>I.e., it is a result of the First Romance Palatalization (Posner 1996:113)

<sup>35</sup>However, the sound has now merged in metropolitan France, even though citation forms and conservative speech may maintain the difference

One explanation is that French retained the  $\widehat{ts}/s/$  distinction at least long enough to counterfeed the later /a/-lengthening. Or, alternatively, the relevant lengthening process happened early enough to counterfeed the merger of /s/ and /ts/. The simplest accuracy-optimal way to implement this is to keep Pope's late 13th century dating for deaffrication<sup>36</sup>, but move the dating for the lengthening process back to just before it.

As it turns out, that we arrived at this modification corroborates findings in the existing literature. We know Early Old French had two alveolar voiceless sibilant sounds, which are typically considered to be /s/ and  $\widehat{ts}/$ . The former would be written *ss* between vowels *s* elsewhere, whereas the latter was  $\langle c \rangle$  before  $\langle e \rangle$  or  $\langle i \rangle$  and  $\langle \zeta \rangle$  elsewhere<sup>37</sup>.

At some point these two merged into /s/. This is traditionally held to have occurred in Later Old French (Anderson & Creore 2018:279), with Pope (1934:93) among others<sup>38</sup> specifying the change as occurring "in the course of the thirteenth century". Pope (1934), meanwhile, dates the /s/-conditioned lengthening of /a/ to Middle French, but as we saw, her assumption that /a/-lengthening came after the aforementioned merger causes errors.

Our result is consistent with literature after Pope (1934), such as Joos (1952) and Martinet (1970). Joos shared our conclusion that the difference in reflexes of /a/ before  $\widehat{ts}/$  and /s/ should imply the distinction lasted long enough to condition a split in the reflexes of /a/, and also provides further corroboration: loans from Later Old French in Middle High German for words with  $\langle ce \rangle$  or  $\langle \zeta \rangle$  end up with German  $\langle z \rangle$  (/ts/) while those with /s/ render German /s/.

The view of Joos (1952) and Martinet (1970) is not the universal consensus on the matter, however. Adams (1975), on the other hand, agrees that the distinction survived, but argues the affricate itself did not survive until the late 13th century as the traditional view holds. Instead, he argues, the distinction came to be the place of articulation, rather than the manner. Adams grounds his reasoning on borrowings from French to Middle English, where Later Old French  $\langle ce \rangle$  and  $\langle \zeta \rangle$  are reflected as /s/, while Later Old French  $\langle ss \rangle$  may be reflected in English as either /s/ or /ʃ/. Adams takes these reflexes to mean that Later Old French distinguished not between an affricate and a fricative voiceless alveolar sibilant, but rather between a "prelaminal" and a "retracted" voiceless sibilant. Regarding German, Penzl (1968) also challenged the views of Joos (1952) and Martinet (1970), and much has been written on the issue (Esau 1976; Fought 1979; Voyles 1972). Fought (1979:854) challenges Adams (1975)'s account, noting that inferences based on French loans into English should be viewed with suspicion, especially given that they came from the Norman-Picard dialect group, which he argues often had retracted "shibilants" corresponding to Central French sibilants. In our view, the details of this dispute are less important than the tacit consensus toward which our procedure with DiaSim independently led us as well: /s/ and /ts/ retained some form of distinction late enough to make divergent outcomes for a preceding /a/. In addition to favoring Joos (1952) and Martinet (1970) over Pope (1934), we do not find the view of Adams (1975) to be necessary from an internal perspective.

Ultimately it appears, as Adams (1975) concedes, that the views of Joos (1952) and Martinet

<sup>36</sup>See Pope (1934) §172, §194

<sup>37</sup>This is a simplification. In contexts that are not important for this specific investigation,  $\widehat{ts}/$  was also written as  $\langle z \rangle$  or  $\langle s \rangle$ , such as  $\langle filz \rangle$  "son")

<sup>38</sup>For example, Delattre (1946)



1 before	FOCUS	1 after	2 after	3 after	4 after
<0.3 thresh	XXXXX	+lab : 0.91	-long : 0.42	<0.3 thresh	<0.3 thresh
<0.3 thresh	XXXXX	+round : 0.83	-lo : 0.42	<0.3 thresh	<0.3 thresh
<0.3 thresh	XXXXX	-long : 0.42	-splng : 0.40	<0.3 thresh	<0.3 thresh
<0.3 thresh	XXXXX	+cont : 0.40	-rtr : 0.40	<0.3 thresh	<0.3 thresh

Figure 13: Context autopsy for words with v at the input.

```
Setting filter sequence to define lexicon subsample.
[Filtering from In]
Enter the phoneme sequence filter, delimiting phones with ' '
w [+lab]
Success: now making subsample with filter w [+lab]
(Pivot moment name: In)
Filter seq : w [+lab]
Size of subset : 33;
2.394% of whole
Accuracy on subset with sequence w [+lab] in In : 30.%
Percent of errors included in subset: 10.043668122270741%
```

Figure 14: Statistics for the data subset with filter sequence “w [+lab]”.

(1970) appear more prevalent. It is further corroboration, then, that we reached the same conclusion independently using CFR as an empirical tool.

### 8.3 Retention of Latin b/v distinction into Gallo-Roman

One fairly widespread traditional view is that Classical Latin  $v$ <sup>39</sup> and intervocalic  $b$  merged into a single sound, thought to be /β/ which in the majority of Romance languages ultimately moved to /v/. This development in French is equated by (Pope 1934:p.91) with that early *pan-Romance* shift, but our investigation suggests that it may have instead happened centuries later in French.

The traditional dating of this shift is quite early (Adamik 2017b), dated to the 1st century CE (Allen 1989:40–42). This timing is grounded on attested evidence, such as errant spelling of the name of the emperor Nerva.<sup>40</sup> However, a closer look suggests two apparently later-dated phenomena which imply Latin  $b$  and  $v$  remained distinct later than previously thought.

We set our focus point to the input state, and filter for words containing this phoneme to perform the context autopsy (Figure 13). As we see in the autopsy, the most predictive context is a following phone that is labial and round — these must be round vowels, because no input forms have geminate /ww/ and we have no rounded consonants, treating /kw/ as two separate phonemes in sequence rather than /k<sup>w</sup>/ just as Pope (1934) does.

We confirm our suspicions by resetting the filter to the sequence “w [+lab]”, to obtain the statis-

<sup>39</sup>Often rendered as /w/ in textbooks on Latin, but Pope considers it to have been a rounded “fricative” with the same place of articulation – /ɣ<sup>w</sup>/.

<sup>40</sup>As noted in Pope (1934) §186 as *Nerβa*. For an indepth discussion on some of this evidence – and its relevance for relative measures of Greek influence – consult Adamik (2017b).

```

,a d w o k 'a : t u m
#,aḍɣʷok'a:tum# | 4 : {j;w} > {j;ɣʷ} / __ [+syl]
#,aḍɣʷok'a:tum# | 12 : [+syl,-lo,-long] > [-tense,-cons,-lat,+cont,θdelrel]
#,aḍβok'a:tum# | 47 : ɣʷ > β
#,aḍβok'a:tum# | 55 : [+syl,+long] > [-long,-splng]
#,aḍβok'a:tum# | 65 : [+hi,-tense] > [-hi,+tense,-cons,-lat,+cont,θdelrel]
#,aḍβok'a:tum# | 70 : [+nas,+cons] > ø / [-stres] __ #
Late Latin stage form : #,aḍβok'a:tum#
#,aḍβok'a:tum# | 109 : [+prim] > [+long] / __ [+cons] [-cons]
#,aḍok'a:tum# | 126 : β > ø / __ [+lab,+syl]
#,aḍog'a:tum# | 132 : k > g / [-cons] __ [+syl]
#,aḍog'a:tum# | 134 : g > ɣ / [-cons] __ [-cons]
#,aḍo'a:tum# | 178 : ɣ > ø / [+round,-cons] __ [+lo]
#,aḍo'a:tum# | 193 : [+lo] > [+front,-back]
#,aḍo'a:tum# | 241 : [-delrel,+voi] > [+cont,θdelrel] / [-cons] __ [-cons]
#,aḍo'a:tə# | 262 : [-stres,-lo,+syl] > ə / __ #
#,aḍo'a:tə# | 269 : ə > ɐ
#,aḍo'a:t# | 292 : ɐ > ø
#,aḍo'aet# | 405 : {'a:; 'e:; 'o:; 'ε:} > {'a ɛ; 'e j; 'o w; 'i ɛ}
#,aḍo'aet# | 432 : [-delrel] > [+cont,θdelrel] / [-cons] __ #
#,aḍo'e:θ# | 444 : {a ɛ; a ɛ; 'a ɛ} > {e:; e:; 'e:}
Old French I stage form : #,aḍo'e:θ#

```

Figure 15: The errant derivation of ADVOCĀTUM to <avoué>

tics (Figure 14), where we see that the subset consists of 33 etyma, on which we start out with 30% accuracy. Furthermore that this subset contains over 10% of all the errors overall. Clearly we have found something meaningful pertaining to this error. To go deeper, we look at a lexical derivation to see what is going wrong.

In Figure 15, we see the (currently errant) derivation of ADVOCATUM, which should ultimately arrive at <avoué> /avwe/ but instead ends up as in /aḍo'e:θ/ by Old French, and after the cutoff in the figure, ultimately develops in the errant model into the (typologically shocking) sequence /aœ/. What went wrong here?

Looking at rule 126 in Figure 15 ("β > / \_\_ [+lab,+syl]"), we locate the spot at which the /β/, which developed into /v/ elsewhere, is effaced. Given that the /k/ or its reflex at some later point being effaced is consistent between the correct result /avwe/, we suspect that the error must be connected to rule 126.

However, this does not mean that rule 126 is itself the problem. Consider the words affected by rule 126 in Table 4. Close observation reveals a strongly predictive pattern: those etyma where /β/ reflects a Latin *b* are unproblematic. On the other hand, the larger category of affected etyma for which the /β/ just before the operation of rule 126 reflects not /b/ but rather /w/<sup>41</sup> are, strikingly, all incorrect. This strongly suggests that the error lies in the merger of Latin *v* and *b* into the posited Gallo-Roman /β/. In Figure 15, we locate this merger of /ɣʷ/ at rule 47.

<sup>41</sup>which in all three cascades is quickly replaced by ɣʷ for the sake of consistency with Pope when non-problematically possible)

Latin	Pivot at rule 126	Result output	Gold standard output
ADVOCĀTUM	/ᵛᵃḍβᵃk¹a:to/	/aʁe/	<i>avoué</i> /avwe/
CERVUM	/c¹erβo/	/sɛʁ/	<i>cerf</i> /sɛʁ/
NĀTĪVUM	/ᵛᵃt¹i:βo/	/ᵛaiw/	<i>naïf</i> /ᵛaif/
NERVUM	/ᵛ¹erβo/	/ᵛjɛʁ/	<i>nerf</i> /ᵛɛʁf/
OWUM	/¹o:βo/	/jø/	<i>œuf</i> /œf/
PARABOLAM	/p¹ar¹a:βola/	/paʁol/	<i>parole</i> /paʁol/
SALVUM	/s¹al¹vβo/	/sø/	<i>sauf</i> /sof/
SERVŌS	/s¹erβos/	/sjɛʁ/	<i>sers</i> /sɛʁ/
SERVUM	/s¹erβo/	/sjɛʁ/	<i>serf</i> /sɛʁ/
TABŌNEM	/t¹aβ¹o:ᵛɛ/	/tā/	<i>taon</i> /tā/

Table 4: The forms of words that are valid with the sequence “β [+lab,+round]”, which triggers rule 126, just before its operation. Note that outputs are correct when /β/ reflects Latin ʁ rather than v. The first column of shows the Latin forms, the second shows the generated forms right before the operation of rule 126, and the third and fourth columns show the observed and correct modern French outputs respectively.

The simplest fix to this issue is to delay the merger of ʁ and v, and the simplest way to realize this is to maintain rule 47 as the point at which Latin v became primarily labial in articulation, but make it different from /β/. We choose our “dummy phoneme” in this case <sup>42</sup> to be /β<sup>w</sup>/, and merge it into /β/ later than rule 126. As it turns out, this fix alone increases overall accuracy on FLLEX by 1.2%.

Our solution is consistent with recent scholarship. Surveys of epigraphic evidence have revealed a relative scarcity of inscriptions confusing the ʁ and v in both Northern and Southern Gaul compared to the rest of the Empire (Adamik 2017a), and a near absence of the confusion during the period it was supposed to have happened, with it only appearing in the later imperial period, as attested in inscriptions from the 4th century onward (Adamik 2017b; Barbarino 2018). Even when it does appear in the later imperial period in Gaul, the confusion between the two letters appears in more restricted contexts than other regions, compared to regions in Italy, Iberia and Dalmatia (Herman 1965). Of all regions, only Gaul and Britain show such a lack of b/v confusion (Adams 2007:662), consistent with the long-running scholarly characterization of the regional Latin of even the later empire in both Britain (Jackson 1953:107) and Northern Gaul (Pope 1934:98) as hypercorrect and strongly shaped by the teaching in schools. Thus, in summary, we have again independently reached conclusions that are supported by literature subsequent to Pope.

#### 8.4 Classical French grammarians as reliable primary sources?

A disproportionate share of the fixes (some with quite large impact) concern cases where Pope relies on the word of grammarians who lived during the epochs of interest to determine when and

<sup>42</sup>It doesn’t really matter which one we pick, as it will ultimately be merged into /β/. We do not assert the value of this phoneme was specifically /β<sup>w</sup>/, merely that it had not yet merged with /β/ (< intervocalic Latin ʁ), and the simplest way to do this is to have \*ʁ<sup>w</sup> retain its roundedness, since \*ʁ is a separate phoneme at this stage, reflecting a lenited intervocalic Latin c or g. Essentially, this is a \*β<sub>2</sub>.

how changes took place. These fixes display particular patterns as well, in that they often involve the dating of phenomena being moved, typically earlier (and some many centuries earlier). Another trend is that they also very often directly concern effects on historical vowel length. We suggest these phenomena warrant re-examination. Notable examples include:

1. the dating of the degemination of /rr/ (§8.4.1)
2. when and if there was a meaningfully diachronic shift of /ɛ/ to /a/ before /r/ (§8.4.2)
3. the dating of lengthening effects before final /sə/ relative to deaffrication (§8.2.1)
4. the dating of /je/ to /i.e/ “diaeresis” (cf Pope (1934: §511) ; we moved it a century earlier)
5. Pope (1934 : §171.3) posits intervocalic voicing in Later Old French affecting /s/ and /f/, but this only seems to cause problems for /s/, and is unnecessary for /f/, at least in our data.

The writings of such grammarians do provide valuable insight into a language’s development when they are available. However, there are many risks in uncritically reading pre-modern grammarians as objective and representative sources for then-synchronic descriptive phonological “truth”, rather than for their subjective prescriptive preferences, often stylistically deemed useful for specific language uses (in the case of Old French, poetry). Even for a reliably descriptive source, though, the “extreme” philology-reliant position remains dangerous, because, as Posner (1997) notes, it merely “recreates the language of a single text, or the **idiolect** of an individual writer”. Posner was writing about French, whose historical grammarians, all aristocratic and overwhelmingly male, certainly did not lack for prescriptivist reporting that could not faithfully represent most French speech<sup>43</sup>. Thus, while we do not dispute the utility of comparing remarks of historical grammarians from different eras, we are wary of *relying* on grammarians as representative and objective evidence as a method for identifying and dating diachronic phenomena.

This also reflects differences in how we conceive of sound change. Whereas a philology-reliant approach carries implicitly assumes that the written forms represent the ancestor of the modern variety, we make no assumptions against or in favor of this position, instead viewing it as an empirical question that can be tackled using CFR. After all, in the case of French, we do know that the Vulgar Latin ancestor of French is *not* equivalent to its Classical Latin contemporary.

#### 8.4.1 Dating of /rr/ degemination

One case where using CFR leads to conclusions that challenge Pope’s philologically grounded dating is the degemination of intervocalic /rr/. Pope (1934: §366) assigns it to Middle French, but moving it back to the 12th century (i.e., Old French) produced an increase in final accuracy of over one percent. If /rr/ was preserved as (the only) intervocalic geminate for centuries until it was simplified in Middle French “usually with compensatory [prior] vowel lengthening” (Pope 1934:p. 147), then we end up with any /a/ beforehand becoming /a / and thus later /a/. However, this

---

<sup>43</sup>Indeed, quotes from grammarians cited by Pope rarely seem to be objective and dispassionate. They are usually praising proper poetic diction, or denouncing some pronunciation as an affront to their language committed by the youth or the “vulgar classes”.

is not what we observe for the relevant words: ⟨*barre*⟩ /baʁ/ (< Latin *BARRAM*), ⟨*charrue*⟩ /ʃaʁy/ (< Latin *CARRŪCAM*), and so forth all have /a/, not /ɑ/. Either a separate rule shifted /ɑ/ to /a/ before /r/ *but nowhere else*, or whatever variety of French it was that experienced /arr/ > /a:r/ > /a:r/, it wasn't the ancestor of standard modern French. The latter is the simpler solution.

Like many other high impact fixes, there is scholarship which is consistent with our findings here. Fouché (1952/1966) agrees with Pope that a true /r/ vs. /rr/ distinction continued to exist, but others (de la Chaussée 1974; Nyrop 1899) dispute this view, and Rheinfelder (1975) argues the merger occurs yet earlier than we came to conservatively conclude. Martinet (1952:205), consistent with our analysis, argues that as a whole “Northern Gallo-Romance” is less conservative than the rest of Romance in preserving the intervocalic geminate and was the first among Western Romance varieties to degeminate intervocalic /rr/.

The degemination of /rr/ need not mean that Old French did not have a geminate /rr/ at all, because it could have arisen anew – namely, such a sequence may have developed from /ðr/ (from earlier /tr/, /dr/). Scheer (2014) proposes /ðr/ > /rr/ specifically after an unstressed vowel.<sup>44</sup> Scheer (2014)'s proposal seems to reconcile what we found with the orthography: as long as we also abandon Pope's assertion, based on classical grammarians, that primary stressed vowels lengthened before /rr/, this would also fix the errors we observed. The reasoning of Scheer (2014) also makes sense, but the existing literature aside from Pope (1934) was not considered when making the adjustment for DiaCLEF so that it could be replicated (or not) independently, and losing the distinction in the twelfth century achieves the same improvement without the complexity. Of course, this may be biased by our modern reference point, since later phenomena might have erased a distinction that could have made Scheer's interpretation necessary. Either way, however, this is clearly a case where we have been led to a conclusion that challenges the reliability of historical grammarians as “sources” for diachronic phonology.

#### 8.4.2 Pre-rhotic lowering: Prescriptivist miracle or prescriptivist error?

As seen in Figure 16, one particularly damaging error pairs a correct final /ɛ/ with an erroneous /a/. This confusion had multiple causes, but it does not take long to find the most prominent one.

At the state when we took the screenshot (Figure 16), 83% of these errors occur before /ʁ/. Examining the derivations of affected etyma, one encounters a certain baseline cascade rule both early and often: the shift of /ɛ/ to /a/ before /r/. Setting a filter of /ɛ r/ at the point just before this captures a very large and very erroneous subset. As seen in Figure 17, we have only 24% accuracy on this subset of 89 etyma (less than 7% of the data), which contains nearly 27% of all errors.

The 89 etyma involved include 21 for which the shift was at least superficially “correct” in the sense that the correct output forms did indeed have /a/, not /ɛ/. On the other hand, the errors for the other 69 words involved caused a disturbance that was significant for the score over the entire lexicon.

<sup>44</sup> Martinet (1952) suggests that Parisian French preserved a “tongue/tip” distinction between reflexes of /r/ and /rr/ (< /ðr /); /rr/'s reflex would become /ʁ/ before the two merged through hypercorrection after “confusion for two generations”. Posner (1997:289) meanwhile attributes the appearance of ‘guttural R’ in Paris to dialectal influences that became fashionable for a time, though we cannot necessarily assume that the literary dialect of Paris is (or is not) what is being modeled by forward reconstruction.

```

Result phones most associated with error:
0: /a/ with rate 0.3065326633165829, Rate present in mismatches : 48.60557768924303
1: /ə/ with rate 0.2857142857142857, Rate present in mismatches : 0.796812749003984
2: /v/ with rate 0.26011560693641617, Rate present in mismatches : 17.92828685258964
3: /q/ with rate 0.2571428571428571, Rate present in mismatches : 3.585657370517928
Gold phones most associated with error:
---
Most common confusions:
----
Confusion 1: a for ε
% of errant words with this confusion : 31.075%
Most common predictors of this confusion :
Percent word bound for pre prior: 78.2051282051282
pre prior phone constant features: -lat -strid -sg -cg -rtr -long -splng
No particularly common pre prior phones.
prior phone constant features: -syl -sg -cg -hi -lo -round -rtr -long -splng
Most common prior phones: /s/ (24.3%)
posterior phone constant features: -syl -lat -sg -cg -lab -lo -round -rtr -long -splng
Most common posterior phones: /ʌ/ (83.3%)
Percent word bound for post posterior: 33.333333333333336
post posterior phone constant features: -sg -cg -lo -round -rtr -long -splng
No particularly common post posterior phones.
----
Confusion 2: w for ø

```

Figure 16: It appears that /a:/ε/ is a particularly problematic error. Why could this be?

```

Enter the phoneme sequence filter, delimiting phones with ' '
[-round,-tense] r
Success: now making subsample with filter [-round,-tense] r
(Pivot moment name: pivot@R659)
Filter seq : [-round,-tense] r
Size of subset : 89;
6.463% of whole
Accuracy on subset with sequence [-round,-tense] r in pivot@R659 : 24.%
Percent of errors included in subset: 26.8%

```

Figure 17: The rule lowering /ε/ > /a/ before /r/ affects only 6.5% of etyma but causes nearly 27% of errors.

The picture becomes more damning when one examines whether the cases where the rule produced the correct outcome needed the rule to do so. Twelve of these are words with the sequence ⟨*oir*⟩ /wɛr/; since /wɛ/ regularly became /wa/ later anyway, they remain correct without this rule.

That leaves the eight cases where this otherwise deeply problematic rule is the source of a correct outcome ⟨*dartre*⟩ “tetter” (a medieval skin disease), ⟨*sarcelle*⟩ “oldsquaw”,<sup>45</sup> ⟨*larme*⟩ “teardrop”, ⟨*marchant*⟩ “merchant”, ⟨*marché*⟩ “market”, ⟨*par*⟩ “by”, ⟨*parchemin*⟩ “parchment”, and ⟨*parpaing*⟩ “perpend stone” (a type of building block). For ⟨*marché*⟩ “market” (< Latin *MERCĀTUM*), the English loan *market* must date before the effacement of the final /t > θ/, yet it already shows effects of the pre-rhotic lowering. This is surprising, given that the rule causing all these errors is four centuries too late for that, in Middle French. This is important, but before we bring in other factors to the discussion, we first make a decision of what to do to fix the errors caused by this rule. Analysis on possible patterning at the point of this rule’s operation (in our case using DiaSim’s context autopsy diagnostic) revealed no persuasive patterning that could be used to restrict the effects of the rule so that they preserve its desirable effects while avoiding the problems it caused. Since there is no clear way to fix the existing rule, the sensible thing to do here is to remove the rule entirely, which immediately causes a gain in accuracy of 4%.

Perhaps, however, given that the rule did after all have the right effect for eight etyma, we should worry about throwing out the baby with the bathwater. We return to our simulation and filter it for words that had the same /ɛr/ ([-round,-tense] r) sequence at the point when the rule once operated, a subset for which we have now improved our accuracy from 24% to 78%. Unfortunately, there does not seem to be any uniting pattern which we can use to make a rule that salvages those eight etyma – as we shall soon see however, it was the timing that was wrong.

Now we must ask, “where did this baseline rule come from?” It was part of the baseline as it was a sound change stipulated by Pope (1934: §496), who supported it with reference to the words of Middle French authors Villon (15th century) and Henri Estienne (16th century) who discussed the “widespread” (and supposedly abhorrent) propagation of the tendency to replace /ɛ/ with /a/ before /r/. Pope does not explicitly speak in terms of the shift’s “regularity”, but she does present this in the same way as she does other shifts which were indeed diachronically regular, such as the similar lowering of /e/ to /ɛ/ (Pope 1934: §494) which occurred around the same time before laterals. In presenting the phenomenon, she takes a sociolinguistic approach, and notes the “two opposing tendencies” of the development’s spread from the eastern part of the French speaking community into the lower class of Paris, and the upper class reaction to it (Pope 1934:188). Pope (1934: §498) notes that the ultimate effect of the “vacillation” between the “opposing” forces of the lowering and the upper class “reaction” to it ultimately ended up in an equilibrium wherein the “vulgar” pronunciation remained established for words of the “colloquial” type, in technical terms and “words of unknown origin”,<sup>46</sup> whereas those of “more elevated diction” and/or infrequent usage<sup>47</sup> experienced the hypercorrection of /a/ to /ɛ/.

<sup>45</sup> A specific kind of salt water duck that inhabits coastal regions of Northern Europe

<sup>46</sup> These words are now variously attributed to Germanic (⟨*boulevard*⟩ “boulevard”, ⟨*écharpe*⟩ “scarf”), Celtic (⟨*dartre*⟩, “tetter”), and Provençal (⟨*farouche*⟩ “shy of humans”); some were actually inherited (⟨*larme*⟩ “teardrop” < LACRIMA, ⟨*sarcelle*⟩ “type of duck” < QUERQUEDULA).

<sup>47</sup> Examples: ⟨*asperge*⟩ “asparagus”, ⟨*cercueil*⟩ “coffin”, ⟨*chair*⟩ “meat”, ⟨*épervier*⟩ “sparrow hawk”, ⟨*gercer*⟩ “to become chapped”, ⟨*serpe*⟩ “sickle”, ⟨*ergot*⟩ “spur, ergot”, ⟨*gerbe*⟩ “sheaf of wheat”, and ⟨*guérir*⟩ “to heal”. While

Per Janda & Joseph (2003), we initially assume the original rule to be phonetically motivated and regular, but vulnerable to later distortion by irregular non-phonetically motivated phenomena.<sup>48</sup> As such, we consider the initial lowering shift to be the regular rule to be placed in the cascade, and the “corrections” to be irregular non-phonological phenomena. This is the same policy we held consistent for all similar cases in Pope (1934) when building BaseCLEF. The sheer weight of the error challenges baseline assumption that an originally regular sound change lowering /ɛ/ to /a/ before /r/ was later perturbed by an aristocratic “reaction” leading to restoration of prior forms and occasional overcorrection.

However, regarding the eight correct cases, there is more to the picture here. Pope (1934: §496) *also* notes another relevant phenomenon. In her view, there were analogous “isolated instances” where /ɛ/ lowered to /a/ before /r/, and she lists the two cases of ⟨*per*⟩ and (Early Old French) ⟨*marchié*⟩. Both of these are among the rule’s eight “success cases”, and we might begin to wonder if there is actually something systematic here.

The Old French corpus reveals that indeed there appears to be something going on in late Gallo-Roman and Early Old French, rather than Middle French. We see that ⟨*per*⟩ was already ⟨*par*⟩ by 1050, as it is in *St Alexis* and *Roland* (ATILF, TLFi: *par*). *MERCĀTUM* becomes ⟨*marched*⟩ already in 1000 and ⟨*marchié*⟩ by 1150 (ATILF, TLFi: *marché*); we see similar timing for *parchemin* (1050) and *marchand* (1140) (ATILF, TLFi: *parchemin*; ATILF, TLFi: *marchand*). However, *LACRIMA* >...>⟨*larme*⟩ shows a less linear trajectory,<sup>49</sup> perhaps due to abortive hypercorrection, and *sarcelle* and *dartre* behave more in line with Pope’s account.<sup>50</sup>

We place our focus point between the effacement of preconsonantal /z/, which is known to have occurred just before (Pope 1934: §377) the Norman conquest of England (1066), and the raising of /o/ to /u/ which occurred in the 11th and 12th centuries (Pope 1934: §184). Here, we find that there is indeed a pattern separating the “correct cases” from the rest. The countertonic /ɛ/ lowers to /a/ when it is after a labial consonant, and when the /r/ precedes either /p/, /tʃ/, or the word boundary. That /p/ and /tʃ/ pattern together may be surprising at first; however, French /ʃ/, which reflects Old French /tʃ/, is articulated with lip rounding. If this is also the case in Old French, /p/ and /tʃ/ are then the only two voiceless occlusives with rounded articulations. We place this rule at the appropriate 11th century point, and observe that of our prior 89 word subset, only three errors remain: ⟨*dartre*⟩, ⟨*sarcelle*⟩, and ⟨*larme*⟩.

While Pope considers each etymon alone, we look at the inherited lexicon en masse. We find it hard to justify viewing a shift that is correct only for two etyma (⟨*larme*⟩, ⟨*sarcelle*⟩) and causes errors for at least 69 as being the regular shift. We are not convinced that prescriptivism among a subset of the literate and aristocratic hypercorrection alone could accomplish the feat of near-total

---

some of these items are associated with monastery (asparagus) or aristocratic (sparrow hawks in falconry) life, we are not sure about some of the others here: ⟨*chair*⟩ “meat” would be quite frequently used, while some of the others here pertain to agricultural life that is likely to be unaffected by the prescriptivism of aristocrats; indeed ⟨*ergot*⟩ “spur” is an agricultural substrate word would normally be assumed to be most *resistant* to trends associated with aristocracy.

<sup>48</sup>Pope notes attestations of irregular hypercorrection, including overgeneralizing effects in “*Perys*” for Paris and “*mery*” for ⟨*mari*⟩ (“husband”), as noted by Geoffrey Tory (1521).

<sup>49</sup>⟨*lerme*⟩ from 1050 to 1200, replaced by ⟨*larme*⟩ in 1200, but ⟨*lerme*⟩ resurfaces for a bit around 1500.

<sup>50</sup>We first see lowered ⟨*sarcelle*⟩ in 1564, though it alternates with ⟨*cercelle*⟩ continuously for two centuries afterward; ⟨*dertre*⟩ is ousted by ⟨*dartre*⟩ in 1478 (ATILF, TLFi: *dartre*).



reversal of a regular sound change for most of the lexicon. This would be especially dubious given that 16th century France was an overwhelmingly illiterate and agricultural society.<sup>51</sup> This is more plausible after the advent of mass education, in which French prescriptivism has long been strong indeed.

Whereas the baseline assumes that the 11th century phenomenon was sporadic and the 15th century phenomenon was regular, we propose the reverse: the former was regular but the latter was not. This leaves only three relatively unproblematic error cases in our dataset: *<sarcelle>*, a salt water duck likely unknown to most French people except sailors, *<dartre>*, a rustic word for a host of unrelated skin diseases, and *<larme>*, for which we can attest oscillation back and forth lasting two centuries. As such, it is much easier to defend than making *<sarcelle>* and maybe *<larme>* regular at the cost of failing to account for almost everything else.

## 9 Conclusion

We have demonstrated the utility of computerized forward reconstruction (CFR) for refining diachronic rule cascades. In fact, the magnitude of improvement, from a baseline accuracy of 3.2% up to an improved accuracy of 84.9%, was far better than we expected. Equally important however is that applying our methodology with CFR not only reproduces conclusions in literature coming after Pope (1934), but also contributes new insights even for a language as well-studied as French. That the epoch with, by far, the highest density of corrections was Gallo-Roman demonstrates the utility of our method for less well-studied languages, because our corpus throughout all Gallo-Roman centuries is either quite sparse or nonexistent.

We strongly advocate the adoption of transparent CFR, for the clear advantages it offers in efficiency, accuracy, accountability, and coverage. Furthermore, for less well-studied languages, which constitute the overwhelming majority of the world’s languages at the moment, our method offers a way to accelerate research, and help us understand the relationships between these languages, before they become extinct.

## Acknowledgements

We thank the three anonymous reviewers and the editors at *Diachronica*, as well as the following. For help with checking the FLLex dataset: Teven Le Scao, Raphael Olivier, and Jean-Baptiste Lamare. For help with the German abstract: Rahel Ringger, Mario Kuçi, and Björn Köhnlein. For general useful input at various stages of the project: Jan Andrews, Brian Joseph, Becca Morley, and Björn Köhnlein.

---

<sup>51</sup>It wasn’t until over a century later in the Age of Enlightenment that literacy reached the then-impressive rate of 30% for males (Van Horn Melton 2003), while the female population – which still played the critical role in the development of speech in children – remained almost entirely illiterate.

## References

- Adamik, Béla. 2017a. On the Vulgar Latin merger of /b/ and /w/ and its correlation with the loss of intervocalic /w/: Dialectological evidence from inscriptions. *Pallas. Revue d'études antiques* 103. 25–36.
- Adamik, Béla. 2017b. Potential Greek influence on the Vulgar Latin sound change [b]>[β]: Dialectological evidence from inscriptions. *Acta Antiqua Academiae Scientiarum Hungaricae* 57(1). 11–33.
- Adams, Douglas Q. 1975. The distribution of retracted sibilants in medieval Europe. *Language* 282–292.
- Adams, James Noel. 2007. *The regional diversification of Latin 200 BC-AD 600*. Cambridge: Cambridge University Press.
- Adams, James Noel. 2013. *Social variation and the Latin language*. Cambridge: Cambridge University Press.
- Allen, William Sidney. 1989. *Vox Latina*. Cambridge: Cambridge University Press.
- Anderson, James M & Jo Ann Creore. 2018. *Readings in Romance linguistics*. Berlin: De Gruyter.
- ATILF. 1998-2019a. glas. <https://www.cnrtl.fr/definition/glas/>. Accessed August 29, 2019.
- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2018. *Trésor de la langue français informatisé*. Université de Lorraine, CNRS. At <http://atilf.atilf.fr/tlfi>.
- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2019b. dartre. In *Trésor de la langue française informatisé*, Université de Lorraine, CNRS. At <https://www.cnrtl.fr/definition/dartre/>. Accessed November 16, 2019.
- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2019c. glaive. In *Trésor de la langue française informatisé*, Université de Lorraine , CNRS. At <https://www.cnrtl.fr/definition/glaive/>. Accessed August 29, 2019.
- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2019d. marchand. In *Trésor de la langue française informatisé*, Université de Lorraine , CNRS. At <https://www.cnrtl.fr/definition/marchand/>. Accessed November 16, 2019.
- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2019e. marché. In *Trésor de la langue française informatisé*, Université de Lorraine , CNRS. At <https://www.cnrtl.fr/definition/marché/>. Accessed November 16, 2019.

- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2019f. par. In *Trésor de la langue française informatisé*, Université de Lorraine , CNRS. At <https://www.cnrtl.fr/definition/par/>. Accessed November 16, 2019.
- ATILF, (Analyse et Traitement Informatique de la Langue Française). 1998-2019g. parchemin. In *Tlfi (trésor de la Langue Française) informatisé*, Université de Lorraine , CNRS. At <https://www.cnrtl.fr/definition/parchemin/>. Accessed November 16, 2019.
- Banniard, Michel. 2001. Causes et rythmes du changement langagier en Occident Latin (IIIe-VIIIe s.). *Travaux Neuchatelois de Linguistique (Tranel)* 34(35). 85–99.
- Barbarino, Joseph L. 2018. *The evolution of the Latin /b/-/w/ merger: A quantitative and comparative analysis of the B-V alternation in Latin inscriptions*. <https://muse.jhu.edu/book/58278>. Accessed September 3, 2019.
- Blom, Alderik. 2009. Lingua gallica, lingua celtica: Gaulish, Gallo-Latin, or Gallo-Romance? *Keltische Forschungen* 4.
- Borin, Lars. 1988. A computer model of sound change: An example from Old Church Slavic. *Literary and Linguistic Computing* 3(2). 105–108.
- Bourciez, Édouard & Jean Bourciez. 1967. *Phonétique française: étude historique*. Tradition de l'humanisme. Paris: Klincksieck. <https://books.google.com/books?id=t3pcAAAAAMAAJ>.
- Bourciez, Édouard Eugène Joseph. 1889. *Précis historique de phonétique française: ou exposé des loi qui régissent la transformation des mots latins en français*. Paris: Klincksieck.
- Brunot, Ferdinand & Gustave Charlier. 1927. Histoire de la langue française des origines à 1900, t. vii. la propagation du français en France jusqu'à la fin de l'Ancien Régime. *Revue belge de Philologie et d'Histoire* 6(1). 326–330.
- Buckley, Eugene. 2003. The phonetic origin and phonological extension of Gallo-Roman palatalization. In *First north-american phonology conference*, vol. 1, Montreal: Concordia University.
- Burton-Hunter, Sarah K. 1976. Romance etymology: A computerized model. *Computers and the Humanities* 10(4). 217–220.
- Cerquiglini, Bernard. 2018. *Une langue orpheline*. Paris: Minuit.
- de la Chaussée, François. 1974. *Initiation à la phonétique historique de l'ancien français*. Paris: Klincksieck.
- Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper& Row.
- Cravens, Thomas. 2002. *Comparative historical phonology: Italo-Romance clues to Ibero-Romance sound change*. Amsterdam: John Benjamins Publishing.

- Delattre, Pierre. 1946. Stages of Old French phonetic changes observed in Modern Spanish. *Publications of the Modern Language Association of America* 7–41.
- Dendien, Jacques & Jean-Marie Pierrel. 2003. Le Trésor de la Langue Française informatisé: un exemple d’informatisation d’un dictionnaire de langue de référence. *Traitement automatique des langues* 44(2). 11–37.
- Diefenbach, Lorenz. 1831. *Ueber die jetzigen romanischen schriftsprachen, die spanische, portugiesische, rhätoromanische, in der schweiz, französische, italiaänische und dakoromaische, in mehreren ländern des östlichen Europa, mit vorbemerkungen über Entstehung, Verwandtschaft usw dieses, Sprachstammes*. Frankfurt am Main: J. Ricker.
- Diez, Friedrich. 1836. Grammatik der romanischen Sprachen, 3 vols. *Bonn: Weber* (3rd ed. 1870–1872).
- Dunn, Michael. 2015. Language phylogenies. In *The Routledge handbook of historical linguistics*, 208–229. New York, NY: Routledge.
- Eastlack, Charles L. 1977. Iberochange: a program to simulate systematic sound change in Ibero-Romance. *Computers and the Humanities* 11(2). 81–88.
- Esau, Helmut. 1976. The medieval German sibilants /s/ and /z/. *The Journal of English and Germanic Philology* 75(1/2). 188–197. <http://www.jstor.org/stable/27707993>.
- Figge, Udo. 1966. *Die romanische anlautsonorisation*. Bonn: Romanisches Seminar der Universität Bonn.
- File-Muriel, Richard J & Earl K Brown. 2011. The gradient nature of s-lenition in Caleño Spanish. *Language Variation and Change* 23(2). 223–243. Accessed February 28, 2021.
- Fouché, Pierre. 1952/1966. *Phonétique historique du français*. Paris: Klincksieck.
- Fought, John. 1979. The ‘medieval sibilants’ of the Eulalia-Ludwigslied manuscript and their development in early Old French. *Language* 842–858.
- Gleason, Henry A. 1959. Counting and calculating for historical reconstruction. *Anthropological Linguistics* 22–32.
- Grimes, Joseph E & Frederick B Agard. 1959. Linguistic divergence in Romance. *Language* 35(4). 598–604.
- Hall, Tracy Alan. 2007. Segmental features. *The Cambridge handbook of phonology* 311–334.
- Hartman, Lee. 2003. Phono (version 4.0): Software for modeling regular historical sound change. In *Actas: Viii simposio internacional de comunicación social: Santiago de cuba*, 20–24.
- Herman, J. 1965. *Aspects de la différenciation territoriale du latin sous l’empire*. Paris: Klincksieck. <https://books.google.com/books?id=jRKstAEACAAJ>.

- Hock, Hans Henrich. 2009. *Principles of historical linguistics*. Berlin: De Gruyter.
- Hombert, Jean-Marie, Médard Mouele & Lai-Won Seo. 1991. Outils informatiques pour la linguistique historique bantou. *Pholia* 131.
- Jackson, Kenneth Hurlstone. 1953. *Language and history in early Britain: a chronological survey of the Brittonic languages, first to twelfth century ad* 4. Cambridge: Harvard University Press.
- Janda, Richard D & Brian D Joseph. 2003. Reconsidering the canons of sound-change. In *Historical linguistics 2001: Selected papers from the 15th International Conference on Historical Linguistics, Melbourne, 13-17 August 2001*, vol. 237, 205. Amsterdam: John Benjamins Publishing.
- Johnson, C Douglas. 1972/2019. Formal aspects of phonological description. In *Formal aspects of phonological description*, Berlin: De Gruyter Mouton.
- Joos, Martin. 1952. The medieval sibilants. *Language* 28(2). 222–231.
- Kaplan, Ronald M & Martin Kay. 1981. Phonological rules and finite-state transducers. In *Linguistic Society of America meeting handbook, fifty-sixth annual meeting*, 27–30.
- Kiparsky, Paul & Jeff Good. 1968. Linguistic universals and language change. *Universals in linguistic theory* 170–202.
- Kondrak, Grzegorz. 2002. *Algorithms for language reconstruction*: University of Toronto dissertation.
- Kondrak, Grzegorz. 2003. Phonetic alignment and similarity. *Computers and the Humanities* 37(3). 273–291.
- Lodge, R Anthony. 2004. *A sociolinguistic history of Parisian French*. Cambridge: Cambridge University Press.
- Lodge, R Anthony. 2013. *French: From dialect to standard*. New York, NY: Routledge.
- Lowe, John B. & Martine Mazaudon. 1994. The reconstruction engine: a computer implementation of the comparative method. *Computational Linguistics* 20(3). 381–417.
- Lusignan, Serge. 1986. *Parler vulgairement: les intellectuels et la langue française aux XIIIe et XIVe siècles*, vol. 1. Librairie philosophique J. Vrin; Montréal: Presses de l'Université de Montréal.
- Machonis, Peter A. 1990. *Histoire de la langue: du latin à l'ancien français*. Lanham, MD: University Press of America.
- Maniet, A. 1985. Un programme de phonologie diachronique: de l'«indo-européen» au latin par ordinateur; version définitive. *Cahiers de l'Institut de linguistique de Louvain* 11(1-2). 203–243.

- Marchot, Paul. 1901. *Petite phonétique du française pré littéraire (VIe-Xe siècles)*. Freiburg: B. Veith.
- Marr, Clayton & David R Mortensen. 2020. Computerized forward reconstruction for analysis in diachronic phonology, and Latin to French reflex prediction. In *Proceedings of LT4HALA 2020-1st workshop on language technologies for historical and ancient languages*, 28–36.
- Martinet, André. 1952. Celtic lenition and Western Romance consonants. *Language* 28(2). 192–217.
- Martinet, André. 1970. *Economie des changements phonétiques*. Berne: A. Francke.
- Mazzola, Michael L. 2013. Analogy among French sounds. In *Research on old french: The state of the art*, 149–165. Springer.
- Meyer-Lübke, Wilhelm. 1908. *Historische grammatik der französischen sprache*. Heidelberg: C. Winter.
- Morin, Yves-Charles. 2009. Histoire des systèmes phonique et graphique du français. *Romanische Sprachgeschichte/Histoire linguistique de la Romania* 3. 2907–2926.
- Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer & Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers*, 3475–3484.
- Muzaffar, Towhid Bin. 1997. *Computer simulation of Shawnee historical phonology*: Memorial University of Newfoundland dissertation.
- Noske, Roland. 2011. L’accent en proto-français: arguments factuels et typologiques contre l’influence du francique. In *Congrès Mondial de Linguistique Française 2008*, 307–320. Institut de Linguistique Française, Paris.
- Nyrop, Kristoffer. 1899. Grammaire historique de la langue française vol. 3. *Copenhagen/Paris: Gyldendal/Nordisk Forlag*.
- Pei, Mario A. 1949. A new methodology for Romance classification. *Word* 5(2). 135–146.
- Penzl, Herbert. 1968. Die mittelhochdeutschen sibilanten und ihre Weiterentwicklung. *Word* 24(1-3). 340–349.
- Piowarczyk, Dariusz. 2016. Abstract: A computational-linguistic approach to historical phonology. *New Developments in the Quantitative Study of Languages* 70.
- Pope, Mildred Katharine. 1934. *From Latin to Modern French with especial consideration of Anglo-Norman: Phonology and morphology*. Manchester University Press.

- Posner, Rebecca. 1994. Historical linguistics, language change and the history of French. *Journal of French Language Studies* 4(1). 75–97.
- Posner, Rebecca. 1996. *The Romance languages*. Cambridge: Cambridge University Press.
- Posner, Rebecca. 1997. *Linguistic change in French*. Oxford: Oxford University Press.
- Posner, Rebecca. 2011. Phonemic overlapping and repulsion revisited. *General and Theoretical Linguistics* 7. 235.
- Pyysalo, Jouna. 2017. Proto-Indo-European lexicon: The generative etymological dictionary of Indo-European languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, nodalida, 22-24 may 2017, gothenburg, sweden* 131, 259–262. Linköping: Linköping University Electronic Press.
- Recasens, Daniel. 2002. Weakening and strengthening in Romance revisited. *Italian Journal of Linguistics* 14. 327–374.
- Rey, Alain. 2013. *Dictionnaire historique de la langue française*. Paris: Le Robert.
- Rheinfelder, Hans. 1975. Altfranzösische grammatik, 2 bde. *I: Laut*.
- Richter, Elise. 1934. *Beiträge zur Geschichte der Romanismen, i: Chronologische Phonetik des Französischen*. Halle: Niemeyer.
- Rochet, Bernard L. 2015. *The formation and evolution of the French nasal vowels*, vol. 153. Berlin: De Gruyter.
- Scheer, Tobias. 2014. Muta cum liquida in the light of Tertenia Sardinian metathesis and compensatory lengthening Latin tr > Old French vrr. In *Variation within and across Romance languages: Selected papers from the 41st Linguistic Symposium on Romance Languages (lsrl), ottawa, 5-7 may 2011*, 77–100.
- Short, Ian R. 2013. *Manual of Anglo-Norman*, vol. 8. Manchester: Anglo-Norman Text Society.
- Simonet, Miquel, José I Hualde & Marianna Nadeu. 2012. Lenition of /d/ in spontaneous Spanish and Catalan. In *Thirteenth Annual Conference of the International Speech Communication Association*, .
- Sims-Williams, Patrick. 2018. Mechanising historical phonology. *Transactions of the Philological Society* 116(3). 555–573.
- Smith, Raoul N. 1969. A computer simulation of phonological change. *ITL-Tijdschrift voor Toegepaste Linguistiek* 5(1). 82–91.
- Straka, Georges. 1970. *L'évolution phonétique du latin au français sous l'effet de l'énergie et de la faiblesse articulatoires*. Centre de philologie et de littératures romanes.

- Suchier, Hermann. 1893. *Altfranzösische Grammatik*. Halle: M. Niemeyer.
- Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4). 452–463.
- Thurot, Charles. 1881. *De la prononciation française depuis le commencement du XVI<sup>e</sup> siècle: d'après les témoignages des grammairiens*, vol. 1. Paris: Impr. nationale.
- Van Horn Melton, James. 2003. *Absolutism and the eighteenth-century origins of compulsory schooling in Prussia and Austria*. Cambridge: Cambridge University Press.
- Voyles, Joseph B. 1972. The phonetic quality of OHG Z. *The Journal of English and Germanic Philology* 71(1). 47–55. <http://www.jstor.org/stable/27706155>.
- von Wartburg, Walther et al. 1922–2002. *Französisches Etymologisches Wörterbuch. eine darstellung des galloromanischen sprachschatzes*. Klopp/Winter/Teubner/Zbinden. 25 vols.
- Wehr, Barbara. 2001. Ein westlich-atlantischer sprachbund: Irisch, Französisch, Portugiesisch. In *Fremd und Eigen. untersuchungen zu Grammatik und Wortschatz des Uralischen und Indogermanischen in memoriam Hartmut Katz*, 253–78. Vienna: Edition Praesens.
- Wernicke-Heinrichs, Meike. 1996. *The evolution of French R: a phonological perspective*: Theses (Dept. of French)/Simon Fraser University dissertation.
- Zink, Gaston. 1986. *Phonétique historique du français*. Paris: Presses universitaires de France.

## Résumé

Pour ordonner des modifications phonétiques qui ont façonné l'évolution phonologique de chaque langue, la recherche en phonologie historique s'appuie traditionnellement sur de fastidieuses dérivations manuelles. Cependant, le cerveau humain est enclin à l'erreur et n'a pas la capacité suffisante pour suivre des milliers de dérivations en parallèle. Nous démontrons la reconstruction informatisée de chaque lexème en parallèle comme tâche de calcul avec des métriques d'optimalité, ainsi qu'un outil pour faciliter drastiquement l'empirisme. Dans ce but, nous présentons «DiaSim», une application qui simule des «cascades» de processus diachroniques sur tout le lexique, tout en proposant des diagnostics pour le «débogage» desdites cascades. Nous appliquons notre méthode à une tâche de prédiction réflexe par reconstruction en avant en utilisant *FLLex*, un nouvel ensemble de données que nous avons compilé et rendons accessibles au public, comprenant 1368 couples de formes latin-français. Nous avons également produit et publions un second ensemble de données, *FLLAPS*, qui associe à 310 racines latines leur évolution via cinq étapes intermédiaires jusqu'au français moderne; ces trajectoires ont été obtenues grâce aux tables de développement de Pope (1934). Nous présentons trois cascades accessibles au public : les cascades de références *BaseCLEF* et *BaseCLEF\**, dérivées de l'approche traditionnelle de la diachronologie française par Pope (1934); et




*DiaCLEF*, qui s'obtient en corrigeant *BaseCLEF* suivant les diagnostics de *DiaSim*. *DiaCLEF* surpasse largement les références, faisant passer la précision brute sur *FLLex* de 3.2% à 84.9% d'étymons avec des gains semblables pour chacune des étapes intermédiaires de *FLLAPS*. Étant donné que les modifications effectuées pour construire *DiaCLEF* furent appliquées sans se référer aux recherches passées, leur accord avec les conclusions de divers travaux antérieurs corrobore à la fois la méthode traditionnelle et notre amélioration de celle-ci; nous examinons en détail les implications de certains de nos résultats.

## Zusammenfassung

Um die Sequenzen von Lautänderungen zu ordnen, die die phonologische Entwicklung von Sprachen prägten, haben sich Phonologen bisher auf mühsam von Hand hergeleitete Ableitungen verlassen. Solche Ableitungen sind jedoch fehleranfällig und es ist schwierig, tausende davon in effizienter Weise parallel zu verfolgen. Wir zeigen, wie unser automatisches Verfahren, computerized forward reconstruction (CFR), jedes lexikalische Element automatisch ableitet, was sowohl für Rechenaufgaben mit optimierbaren Metriken als auch als Hilfsmittel für empirische Untersuchungen genutzt werden kann. Zu diesem Zweck stellen wir *DiaSim* vor, eine Applikation, die ‚Kaskaden‘ diachronischer Entwicklungen über das Lexikon einer Sprache simuliert und verschiedene Diagnosen zum ‚Debuggen‘ dieser Kaskaden bereitstellt. Wir testen unsere Methode anhand einer Reflex-Vorhersage von Latein nach Französisch unter Verwendung eines neu kompilierten, öffentlich verfügbaren Datensatzes *FLLex*, der aus 1368 gepaarten lateinischen und modernen französischen Formen besteht. Außerdem präsentieren wir einen zweiten Datensatz, *FLLAPS*, der 310 Reflexe aus dem Lateinischen über fünf attestierte Zwischenstufen bis ins moderne Französische abbildet und aus den periodischen Entwicklungstabellen von Pope (1934) abgeleitet ist. Weiter stellen wir öffentlich verfügbare Regelkaskaden vor: die Basiskaskaden *BaseCLEF* und *BaseCLEF\**, die auf Popes (1934) vielzitierte Annahme zur Entwicklung des Französischen basieren, und *DiaCLEF*, die aus inkrementellen Korrekturen an *BaseCLEF* mit diagnostischer Hilfe von *DiaSim* erstellt wurden. *DiaCLEF* übertrifft die Basislinien um ein Vielfaches und verbessert die Genauigkeit von *FLLex* von 3.2% zu 84.9% aller Etyma mit ähnlich großen Verbesserungen für jede *FLLAPS*-Periode. Veränderungen wurden vorgenommen, um *DiaClef* nur mithilfe eines Ausgangswerts und *DiaSims* diagnostischen Kriterien arbeiten zu lassen, aber oftmals reproduzierten sie unabhängig voneinander bestehende Forschungsergebnisse zur französischen Sprachgeschichte, was sowohl unsere Methoden als auch bestehende Analysen bestätigt. Wir diskutieren die Implikationen einiger unserer Resultate im Detail.

## Authors' Address

Clayton Marr  
300 Oxley Hall  
Department of Linguistics  
The Ohio State University  
1712 Neil Ave  
COLUMBUS, OH  
marr.54@osu.edu  
 <https://orcid.org/0000-0002-9139-3369>

## Co-author information

David Mortensen  
5407 Gates Hillman Complex  
Language Technologies Institute  
Carnegie Mellon University  
5000 Forbes Ave  
PITTSBURGH, PA 15213  
dmortens@cs.cmu.edu

## Publication history

Date received: 22 May 2020  
Date accepted: 13 January 2022  
Date published: November 11, 2023