Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 28–36
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Computerized Forward Reconstruction for Analysis in Diachronic Phonology, and Latin to French Reflex Prediction

Clayton Marr, David Mortensen
Carnegie Mellon University, Carnegie Mellon University
Pittsburgh PA 15213 USA, Pittsburgh PA 15213 USA
cmarr@andrew.cmu.edu, dmortens@cs.cmu.edu

Abstract

Traditionally, historical phonologists have relied on tedious manual derivations to calibrate the sequences of sound changes that shaped the phonological evolution of languages. However, humans are prone to errors, and cannot track thousands of parallel word derivations in any efficient manner. We propose to instead automatically derive each lexical item in parallel, and we demonstrate *forward reconstruction* as both a computational task with metrics to optimize, and as an empirical tool for inquiry. For this end we present DiaSim, a user-facing application that simulates "cascades" of diachronic developments over a language's lexicon and provides diagnostics for "debugging" those cascades. We test our methodology on a Latin-to-French reflex prediction task, using a newly compiled dataset *FLLex* with 1368 paired Latin/French forms. We also present, *FLLAPS*, which maps 310 Latin reflexes through five stages until Modern French, derived from Pope (1934)'s sound tables. Our publicly available rule cascades include the baselines *BaseCLEF* and *BaseCLEF**, representing the received view of Latin to French development, and *DiaCLEF*, build by incremental corrections to *BaseCLEF* aided by DiaSim's diagnostics. DiaCLEF vastly outperforms the baselines, improving final accuracy on FLLex from 3.2%to 84.9%, and similar improvements across FLLAPS' stages. .

Keywords: diachronic phonology, computerized forward simulation, regular sound change, Romance linguistics, French, Latin, DiaSim

1. Introduction

When reconstructing the phonological history of a language, linguists usually 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 (excluding non-phonologically motivated phenomena such as analogy, homophony avoidance, hyper correction, et cetera). This paradigm operationalizes sound change as a classical function: an input maps to a unique output. Aggregated, the ordered sequence ("cascade") of these sound change functions forms an algorithm. Such an algorithmic phenomenon naturally lends itself to automated simulation. There are ample theoretical underpinnings for using simulations – or *computerized forward reconstruction* (Sims-Williams, 2018) (*CFR*) – to test the accuracy of the cascade implied by any given understanding of a language's phonological history. However, for reasons discussed in depth in 1.2., it failed to achieve widespread usage. Instead, current work 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 or memorize the effects of thousands of rules operating over millennia, mapping the forms of thousands of reflexes. No wonder, then, that current work prefers to "zoom in" on one phenomenon.

These typological discussions are greatly useful, but must remain grounded by understanding the histories of the languages in question. The phonological histories of the majority of the world's languages, which likely will not survive the next century, remain mysterious, and work on them would certainly be more efficient if aided by computers. While it could take months for a human to map thousands of etyma across millennia, a computer can do so in seconds. CFR furthermore greatly facilitates thorough coverage of the lexicon. Building on the example of earlier now abandoned projects discussed in section 1.2., we present *DiaSim*, a generalizable transparent forward reconstruction application which offers various diagnostic capabilities, hoping to improve the present situation.

We present our work in using *DiaSim* to "debug" the received understanding of French phonological history, as represented by Pope (1934). We additionally present our newly compiled datasets *FLLex* and *FLLAPs* (described in 5.), with which we demonstrate the utility of CFR using DiaSim. We present results on the measured performance of baseline (derived from Pope (1934)) rule cascades *BaseCLEF* and *BaseCLEF**, and the "debugged" cascade *DiaCLEF*. While the baseline model was 3.2%accuracy (without "uninteresting errors", 30.3%), the corrected ruleset achieved 84.9%accuracy, with the biggest improvement observed in the (largely unattested) Gallo-Roman stage, as discussed at length in section 7..

All of these resources are made publicly available for use, at the DiaSim github repo.

1.1. Related Work

1.1.1. French Phonological History

Romance philology is typically considered founded by François-Juste Raynouard (Posner, 1996, p. 3), and formalized by Diefenbach (1831) and Diez (1836), followed by a work on work on French propelled by Neogrammarianism (Thurot, 1881; Meyer-Lübke, 1899; Suchier, 1893; Marchot, 1901; Nyrop, 1914); foundational early 20th century work includes Fouché (1961), Martinet (1970), Brunot and Charlier (1927), and, of course, Pope (1934). Of the extensive subsequent work, we specifically note Adams (2007)'s work on regional ("Popular") Latin inscriptions, work on French historical sociolinguistics (Lodge, 2013; Lodge and others, 2004; Lusignan, 1986), French orthographical history, and "protofrançais" (Noske, 2011; Banniard, 2001).

Traditional methodology balanced Neogrammarian inquiry with the principle that, as Pope (1934) describes it, "the history of a language should be related as closely as possible to the study of texts". Such methodology often involved tracing changes in spelling as it represented certain sounds and morphemes ("flexion") and taking the remarks of historical writers (especially grammarians) as objective evidence. We, like other recent researchers (Posner, 2011; Fouché, 1961), take a more sceptical look at these writings, viewing them not as descriptions of reality but rather prescriptions for how French subjectively *should* be pronounced. We offer an alternative to relying on these voices: the empirical methodology described in section 3..

Since the beginning, work in French diachronic phonology has functioned more or less to calibrate what is in effect the diachronic *cascade* of French, with Pope's meticulous 1934 opus still considered the "invaluable" (Posner and others, 1997, p. 3) baseline against which new theories in French are being presented as improving upon (Short, 2013). Our aim in this work is twofold. Alongside the goal of demonstrating the power of CFR, we also aim to, like Pope before us, provide a holistic account of French diachrony. Ultimately, our vision is a publicly available cascade for every language of interest that may be improved upon whenever a correction becomes accepted in the field.

1.2. Computerized Forward Reconstruction

Not long after the mid-20th century emergence (Dunn, 2015) of computational historical linguistics (Jäger, 2019) with the works of scholars like Swadesh and Gleason (Swadesh, 1952; Gleason, 1959), the first published CFR (coarsely) derived 650 Russian words from Proto-Indo-European (Smith, 1969); the next derived Old French from Latin in 1976 (Burton-Hunter, 1976). Others looked at Medieval Ibero-Romance (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, and used orthography rather than underlying phones (Piwowarczyk, 2016), having "no notion of phonology" (Kondrak, 2002). Generalizable rule formalisms have in fact been presented in related topics, such as learning *syn*chronic sound rules (Gildea and Jurafsky, 1995).

Phono, a phonologically-motivated and phoneme-mediated forward reconstruction, appeared in 1996 and was applied to Spanish and Shawnee (Hartman, 2003; Muzaffar, 1997), but as far as we know, no further work using Phono was published. Despite computational modeling seeing an "explosion"[1] in other diachronic fields (Dunn, 2015) alongside rapid improvements in computing, CFR fell out of fashion by the late 20th century (Lowe and Mazaudon, 1994), and old CFR systems are now incompatible with modern computers (Kondrak, 2002), Reasons for this decline are varied, including dissent Neogrammarianism, and an unfortunate association with supposedly "unprofessional" enterprises (Sims-Williams, 2018).

2. Contributions

We aim to show that a sufficiently generalizable CFR system is a useful and professional research tool for diachronic phonology. It is recognized (Sims-Williams, 2018) that human cognition simply has insufficient working memory to track all the (likely millions of) implied calculations while mapping sound rule functions spanning centuries or millennia across a language's entire inherited lexicon. Ensuring the accuracy of the tedious human calculations in this scenario is itself extremely onerous and error-prone. On the other hand, the task is trivial for a computer. Information attained in this much more efficient and rigorous manner can then be leveraged to improve our diachronic understanding of the languages in question, revealing new sound laws and analogical patterns, refining existing ones, and revealing new reflexes and cognates, all while ensuring holistic coverage rather than cherry-picking for validation. This improved efficiency and 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.

This paper contributes the following:

- DiaSim, an application that performs *transparent*[2] CFR for rule cascades over any lexicon, offering accuracy metrics and a diagnostics for analysis
- FLLex, a dataset pairing 1368 Latin etyma with their known ("gold") inherited French reflexes.
- FLLAPS, a dataset mapping gold reflexes of 310 Latin etyma across five attested stages
- Two cascades based on the received understanding of Latin > French sound change, and a "debugged" cascade built using DiaSim with PATCH
- PATCH, a guideline for using CFR for inquiry

3. PATCH

We recommend PATCH as an empirically sound way to utilize CFR for scientific inquiry in "debugging" rule cascades. PATCH is described in the following prose, and summarized in figure 1.

The baseline cascade ideally should reflect the latest available, but conservative, "least common denominator" for which there is consensus. For French, such a baseline is easily identifiable — and explicitly used as such still in current research (Short, 2013) — as Pope (1934). In this way, our inquiry can independently support or challenge findings in subsequent literature.

PATCH is then performed on the "working cascade", which starts out as a copy of the baseline before it is progressively

[1] Including analogous work in closely related topics, such as learning FST-based *syn*chronic sound rules (Gildea and Jurafsky, 1995)

[2] See section 4.1.

Figure 1: The PATCH process, summarized.

1. *"Debug" the working cascade*[3] *by repeating the following steps:*
   (a) *(P)inpoint – Isolate a source of error*
   (b) *(A)mend – Try various solutions; choose the one with the best accuracy, preferring simplicity where there are statistical ties*
   (c) *(T)est – is the selection justifiable?*
      i. *If a new sound change is being added, preferably ensure that it can be motivated typologically/theoretically*
      ii. *Ensure there are no adverse side effects*
      iii. *Consult any relevant existing work, and relevant data as appropriate: philology, dialectology, loans, etc.*
   (d) ***CH**oose – If the proposal remains plausible, commit it to the working cascade. Otherwise recalibrate it, or redact it entirely.*

modified. We hold that when using CFR, a linguist should initially make fixes based solely on Neogrammarian empiricism, not prior knowledge (neither topical nor typological). Thus the *Pinpoint* stage is performed "blind" regarding any information not drawn from CFR results. Automated statistical assistance such as DiaSim's diagnostics is often useful to pinpoint the source of error.

One likewise performs the second stage (*Amend*) "blinded" of outside info: the researcher comes up with all reasonable possible solutions to the problem identified in *Pinpoint*, implements them on the working cascade, and records the effects on performance. Of these, (s)he chooses the one with the best performance; in cases where there is no significant difference in performance, choose the fix that is the "simplest". By "simplicity", we do not necessarily mean "the least rules possible and the least specifications on each rule", although in practice the two are often similar. Instead, "simplicity" here refers to the simplest possible way to explain the data. These are different, because leaving numerous lexemes with plausibly related developments unexplained by any single rule is to be considered simpler *only* if we have a "simple" and ideally *single* explanation, such as systematic analogy, interference, or identifiable sociolinguistic effects. On the other hand, leaving them with no explanation at all implies a "default" that they each have lexically explanations – which is the exact opposite of "simplicity", and to be avoided[4]. Then, implement the chosen "fix" by amending the cascade at the proper point.

It is only in the third stage, *Test*, that outside info is weighed against other factors, before a binding decision is made in the final stage *Choose*, to either enshrine the solution in the working cascade, enshrine a modified version, or redact it entirely. Then, to find more fixes, the linguist iteratively repeats this process.

We tried our best to follow PATCH building DiaCLEF. However, we do not advocate brittle literal adherence to PATCH, but rather suggest it as a guideline; we additionally suggest some specific exceptions to its use. Firstly, at the end of the *Choose* stage, if other fixes become clear with the synthesis of data from the simulation and from other sources (such as dated attested forms), they can also be fixed at the time, as long as there is (a) robust corroboration in coverage, and (b) no adverse side effects when checked with the entire dataset. Secondly, fixing baseline rules so that they obtain their stated intended effects when otherwise they clearly do not [5] is exempt from PATCH. Lastly, fixing rules that have already been changed (or moved), or have been created anew by prior iterations of PATCH can be done without the entire process, because this is really a revision of the re-calibration aspect of *Choose*.

4. DiaSim

4.1. Transparent Large-Scale Cascade Simulation

DiaSim transparently simulates a specified rule cascade for every lexeme in parallel. The user must input at minimum (1) a lexicon file, and (2) a cascade. The lexicon file includes the input forms to forward reconstruction, and optionally gold reflex forms for the final or intermediate results of CFR. Each rule in the cascade is written in the conventional SPE format (Chomsky and Halle, 1968). DiaSim implements the subset of the SPE rule formalism that (Johnson, 1972) and (Kaplan and Kay, 1981) showed to be formally equivalent to finite state transducers (FSTs), while enabling users to explicitly modify sound laws in terms of conventional notation rather than computer code[6].

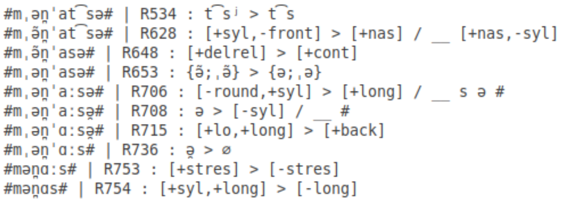
DiaSim can capture any and all regular relations between strings in the specified symbol alphabet, whether that alphabet is the provided IPA default, or another supplied by the user. In between rules, the user may flag a stage, at which the simulation state can be stored and retrieved. Flagged stages may also be used as *pivots* during evaluation to help detect long-distance interactions between rules.

Being able to observe the iterative realization of cascade *transparently* (effects of each rule being "visible") is quite useful for illuminating relationships between involved processes. One can see how the preconditions for later rules may emerge, or be perturbed, or how they fail to do so when expected. For such "transparency", DiaSim can retrieve each time an etyma was changed (shown in figure 2), its new form and by what rule, or all effects of any rule.

[4] We except from this cases that are known to be predictably lexically specific: homophony avoidance, onomatopoeia, and spelling pronunciations.

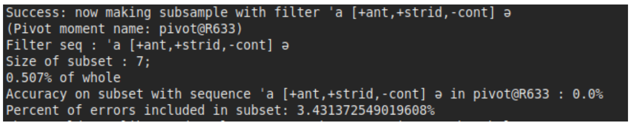
[5] I.e. the baseline source states one outcome but the rule formalism does not produce it. When using DiaSim, a quick way to check this is to check the printouts of etymon-wise *transparent* mutations for the sound change in question.

[6] DiaSim's sound rule "grammar" handles all IPA except for clicks and tones, can support all SPE notations including complicated alpha functions, disjunction, and nested parenthetical optional segments, and adds "@" for "any single phone" (anything but the word bound #).



```
#mˌən̪ˈat͡sə# | R534 : t͡sʲ > t͡s
#mˌə̃n̪ˈat͡sə# | R628 : [+syl,-front] > [+nas] / __ [+nas,-syl]
#mˌə̃n̪ˈasə# | R648 : [+delrel] > [+cont]
#mˌən̪ˈasə# | R653 : {ə̃;ˌə̃} > {ə;ˌə}
#mˌən̪ˈaːsə# | R706 : [-round,+syl] > [+long] / __ s ə #
#mˌən̪ˈaːsə̯# | R708 : ə > [-syl] / __ #
#mˌən̪ˈɑːsə̯# | R715 : [+lo,+long] > [+back]
#mˌən̪ˈɑːs# | R736 : ə̯ > ∅
#mən̪ɑːs# | R753 : [+stres] > [-stres]
#mən̪ɑs# | R754 : [+syl,+long] > [-long]
```

Figure 2: Derivation of *menace* (< Latin MINACIA).



```
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 3: A context autopsy, one of DiaSim's diagnostics. Here the error is likely related to following /t͡s/.

4.2. Performance Metrics

For either the entire lexicon or a chosen subset, DiaSim can supply the word-wise accuracy, the accuracy within one or two phones, the word-wise average Levenshtein distance between result and gold form (normalized for gold length, hence forth *mPED*[7]), and the word-wise average length-normalized *feature edit distance* (Mortensen et al., 2016; Kondrak, 2003) *(mFED)* between result and gold forms. Future work should incorporate a measure of *implied complexity*[8].

These metrics offer different information. Accuracy indicates how much of the lexicon the rule cascade renders correct. On the other hand, mPED gives how wrong we are if we treat phones as discrete tokens, whereas mFED indicates mean phonetic result/gold distance between in terms of phone-wise feature vector distance — on average, how different is each wrong phone from the correct one?

4.3. 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 actually be a sign of a much larger pattern of errors. To help overcome these factors, DiaSim offers a suite of diagnostics.

If interactive mode is flagged at command line, at the end of the simulation, and also any flagged gold stage, DiaSim halts, gives basic performance metrics, and queries if the user would like to run any diagnostic options. These diagnostics, including correlation of error with the presence of segments at the same or different stages (the "context autopsy" diagnostic presented in 3 being an example), identification of particularly common correspondences between errant and gold phones, among others, are enumerated in more detail in the diagnostics README contained in the package.

Wherever phone-wise errors is involved, an alignment algorithm based on minimizing feature edit distance (Mortensen et al., 2016) measures phone-wise error. DiaSim's diagnostics aims to help pinpoint where in the sequence of realized shifts the critical error occurred. For example, the final stage error correlated to 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 and the result, DiaSim can pinpoint for the user if the type of error happens to be particularly common in certain contexts.

These sorts of 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. For example, the traditional paradigm for French (Pope, 1934) holds that the voicing of Latin initial /k/ to Gallo-Roman /g/ was simply sporadic, but as we demonstrate in section 7.1., we were able to detect a plausible new regular rule to explain them collectively.

4.4. Theoretical Grounding

DiaSim was constructed to be faithful to longstanding theory while maintaining flexibility. It is built on the premise that words consist of token instances of a bounded set of phone types (alongside juncture phonemes), and that phones are uniquely defined by value vectors for each of a constant feature set (Chomsky and Halle, 1968; Hall, 2007; Hock, 2009). Each feature can be assigned one of three values : positive (+), negative (-) or unspecified (0). Which features are relevant for phonemic distinctions vary by language. DiaSim allows the user to use a custom set of feature-defined phones and/or of phone-defining features, while providing holistic default sets for each.

5. Datasets

The dataset FLLex[9] consists of 1368 Latin etyma paired with their inherited modern French reflexes. These include all 1061 inherited etyma in French (excluding some verb forms) that are used in Pope (1934), as well as 307 etyma recruited from Rey (2013) and from the online French philological resource, *Trésor de la Langue Française informatisé* (TLFi) ATILF (2019a).

For inclusion, lexemes had to have been in continuous usage throughout the relevant sixteen centuries. Words affected by non-phonologically motivated phenomena such as analogy, folk etymology, etc were excluded, but words with apparent irregularity that could not be attributed to such processes (such as cases of sporadic metathesis) remained included.

Each entry was checked with multiple sources (Pope, 1934; Rey, 2013; ATILF, 2019a) to ensure it was indeed an etymon

[7] (m)ean (P)honeme (E)dit (D)istance

[8] Considering the explicit cascade and the "implicit" complexity of exception cases made for words considered *non-regular* and thus excluded from calculation of all (other) provided metrics

[9] (F)rench from (L)atin (Lex)icon

with continuous usage from Latin to French, unaffected by non-phonologically motivated interference.

The period-indexed dataset FLLAPS[10] is recruited from Pope (1934)'s sound tables. FLLAPS has an intentional degree of balance in phonological coverage, as Pope designed her sound tables to have at least one etymon that was affected by each notable sound change (FLLex, meanwhile, is more proportionally representative of the overall phonemic frequencies of the French language). FLLAPS offers gold forms derived from Pope's philological work for each of the four intermediate stages, including Late Latin in Gallia (dated to circa 400 CE), Old French I (EOFdate), Old French II (circa 1325 CE), and Middle French (circa 1550 CE). A few corrections were made in order to adapt the set for this task. For example, as Pope did not foresee this use of her work, she sometimes omits finer distinctions (such as lax/tense distinctions). When these concern segments that are not of interest to the specific sound changes being demonstrated, the sound changes described elsewhere in her work for the period in question were regularly applied and consistency enforced.

6. Rule cascades

In order to demonstrate both how DiaSim simulates long-term and holistic Neogrammarian sound change, we designed our baseline cascade, **BaseCLEF** [11] to include all regular sound changes posited in (Pope, 1934), which represents the received view of French phonological history, and remains the "indispensible"(Short, 2013) work that others in the field build off of. The **DiaCLEF**[12] cascade was then built from a copy of BaseCLEF by exhaustively correcting non-sporadic errors detected using DiaSim's simulation and evaluation functionalities.

We built BaseCLEF to include all regular sound changes posited in (Pope, 1934), in the order specified. Where Pope's writing is ambiguous, the benefit of the doubt is given as a general policy (that is, we assume the reading that gives the correct output). There are numerous cases where literal interpretation of Pope's treatise leads to "non-interesting" errors, mere omissions and the like because at the time of writing were not essential, perhaps because Pope didn't foresee her work being converted into an explicit rule cascade. For example, Pope states that modern French lacks any phonemic length differences, but never states when it was lost. To handle this, we made an additional ruleset, *BaseCLEF**, where these trivial omissions are corrected.

7. Results

As seen in table 1, the increase in accuracy obtained by "debugging" via DiaSim is striking, with raw accuracy going from 3.2%to 84.9%. The improvement in average feature edit distance, a decrease from 0.518to 0.056, is also large, even when we consider the baseline to be BaseCLEFstar (with "uninteresting" errors already corrected as discussed in section 6.), with 30.3%accuracy and 0.380mean FED.

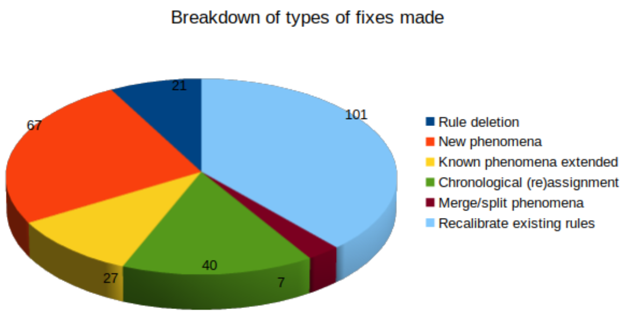


Figure 4: Breakdown of "fixes" in DiaCLEF

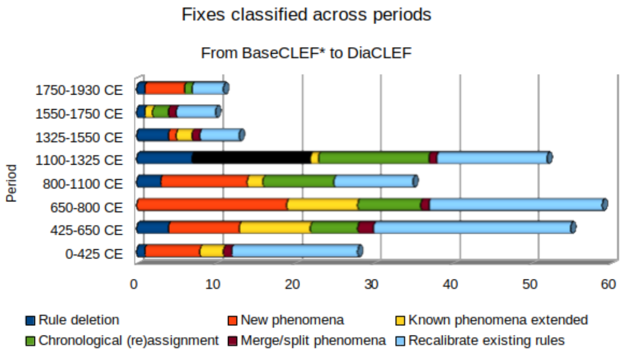


Figure 5: Differences between different periods in number and type of edits made to the cascade

In table 4, we see a breakdown of the sorts the corrections that were done for DiaCLEF (excluding those also handled in BaseCLEF*). The more radical sorts of changes include *rule deletion*, *rule creation*, and re-orderings, constituted 48.7% of changes, leaving the rest to less radical amendments such as extension of acknowledged phenomena, recalibration of rule contexts, and mergers and splits of existing rules.

As displayed in figure 5, the biggest volume of changes occur in the Gallo-Roman and Old French periods. There were notable differences with regard to where changes that fundamentally challenge Pope's understanding of French diachronology led to meaningful improvements. This is also true of re-orderings, which are broken down by period and type in figure 6. 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.

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.

Many of these new insights are discussed at length in (Marr and Mortensen, 2020); we present just one here at length in section 7.1. to demonstrate the empirical use of CFR with PATCH.

[10] (F)rench from (L)atin (L)exicon by (A)ttested (P)eriod (S)ublexica

[11] **Base**line **C**lassical **L**atin **E**tyma to **F**rench

[12] **Dia**Sim-informed **C**lassical **L**atin **E**tyma to **F**rench

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

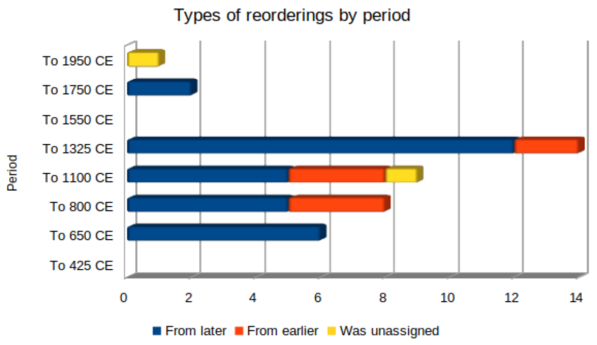
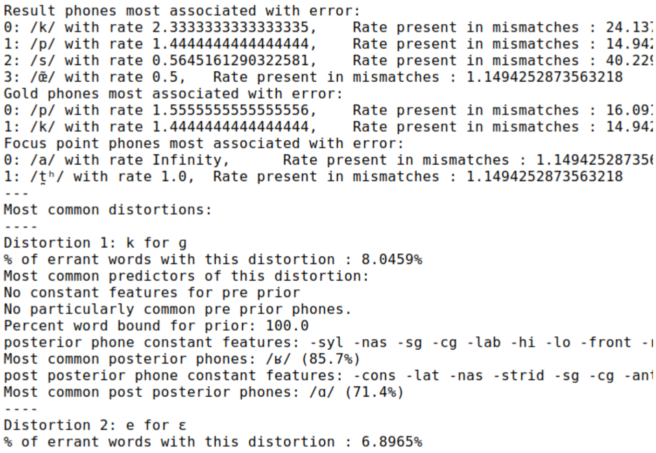


Figure 6: Corrections of ordering by period.



```
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: /œ̃/ 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 : 1.14942528735
1: /t̪ʰ/ 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 -
Most common posterior phones: /ʁ/ (85.7%)
post posterior phone constant features: -cons -lat -nas -strid -sg -cg -ant
Most common post posterior phones: /ɑ/ (71.4%)
----
Distortion 2: e for ɛ
% of errant words with this distortion : 6.8965%
```

Figure 7: DiaSim's Confusion Prognosis

7.1. Regular Explanation for "Sporadic" Onset /k/ Voicing

We use the simple yet striking example of the plausible regularity of Early Old French initial velar stop voicing to demonstrate the use of CFR with PATCH to propose and validate new rules. In this case, we are unable to find any work in the past century and a half of research that treats this plausible regularity as a unified phenomenon, instead giving a number of unrelated explanations for affected etyma.

We begin our investigation (*Find* in PATCH) with DiaSim's *Confusion Prognosis* (figure 7). In the top left, we see the phones which have the highest ratio of occurrence in error cases to correct cases, and in top right we see the overall prevalence in error cases. In the bottom part of the Confusion Prognosis, the most significant correspondences between specific errant and gold phones ("distortions") are displayed.[13]

Here, the most problematic distortion is /k/:/g/, where we find /k/ for what should be /g/, comprising 8% of all errors. Furthermore, /k/ is the phone most correlated with error. 100% of /k/:/g/ distortions occur immediately after the word onset, 86% of cases have the uvular fricative /ʁ/ immediately after, and for 71% of cases, the next phone is /ɑ/. This suggests to the linguist that behind this error, a regular rule may be hiding, and those statistics give an idea of what its conditioning context likely is.

Clearly we are dealing with a case of onset voicing. French fricative /ʁ/ reflects historic sonorant /r/, which is significant, as French lenition likewise happened regularly in Gallo-Roman intervocalic consonant + sonorant clusters. However, because we are consciously choosing to ignore what we think we know about French (per PATCH), we ignore this fact at this point so as not to bias our search, and as seen we will end condition our rule not on specifically sonorant consonants but instead simply on consonants.

This suggests that an onset velar voicing happened at some point in the history of French, but we don't know when. We next aim to isolate the problem by filtering out "noise", identified with the help of our statistics, to get a "noise"-less subset. In our case, we set the *focus point*[14] as the input form from Classical Latin, and use a *filter sequence* "# k @ [+lo]"[15].

The user can then access a list of the resulting subset's errors, which include (with correct forms second) /klе/:/glɛv/, /klɑ/:/glɑ/, /kʁɑs/:/gʁɑs/, /kʁaj/:/gʁij/, and so on. Viewing this list, it is apparent that /k/ in all the error cases lies between the word onset and a consonant. We no longer have to rely on prior knowledge because all the words which end up with uvular /ʁ/ still have alveolar /r/ at our *focus point*. However, because we never observe a non-sonorant consonant having a different effect, we continue to condition our rule on consonants, not sonorants, because we seek the least specific rule possible. If we assert a low vowel after the onset cluster, we perfectly predict the /k/:/g/ distortion, with one exception[16].

The subset of data filtered for etyma with the Latin sequence "# k [+cons] [+lo]" has well under 50% accuracy. Examining the specific non-error cases among this subset, they all have changed the original A into a non-low vowel, and in all of these cases, the A had primary stress and was in an open syllable. The same is true of only one of the error cases[17].

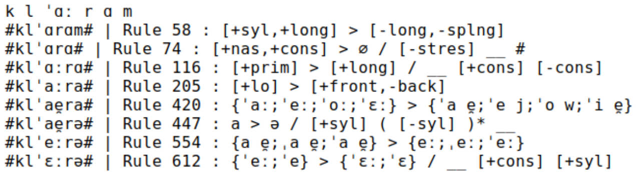
[13] These calculations are done on the back of an alignment algorithm that aligns phones so as to minimize *Feature Edit Distance* (Mortensen et al., 2016).

[14] The time step at which a subset is made using the *filter sequence*

[15] onset k, any single phone ("@"), then a low vowel

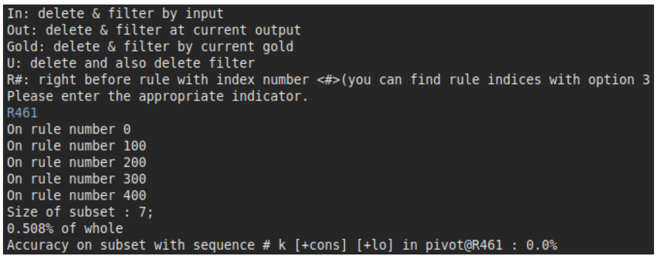
[16] The ⟨*clef*⟩/⟨*clé*⟩ doublet, reflexes of CLĀVEM, with a low vowel

[17] namely, ⟨*glaive*⟩, whose exact history is unclear



```
k l ˈɑː r ɑ m
#klˈɑrɑm# | Rule 58 : [+syl,+long] > [-long,-splng]
#klˈɑrɑ# | Rule 74 : [+nas,+cons] > ∅ / [-stres] __ #
#klˈɑːrɑ# | Rule 116 : [+prim] > [+long] / __ [+cons] [-cons]
#klˈaːra# | Rule 205 : [+lo] > [+front,-back]
#klˈae̯ra# | Rule 420 : {ˈaː;ˈeː;ˈoː;ˈɛː} > {ˈa e̯;ˈe j;ˈo w;ˈi e̯}
#klˈae̯rə# | Rule 447 : a > ə / [+syl] ( [-syl] )* __
#klˈeːrə# | Rule 554 : {a e̯;ˌa e̯;ˈa e̯} > {eː;ˌeː;ˈeː}
#klˈɛːrə# | Rule 612 : {ˈeː;ˈe} > {ˈɛː;ˈɛ} / __ [+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: We isolate our error by setting our focus point right after the last *bleeding* rule, to find a subset with zero accuracy.

This pattern points us toward our next objective — to propose a solution (*Amend* in PATCH). Now that we have determined our rule's conditioning, we want to pin down where it should be placed in the cascade. To locate when the vocalic changes that *bled* (Kiparsky and Good, 1968) our proposed rule occurred, we examine the derivations of affected cases. In the derivation for CLĀRAM >· · ·>⟨*claire*⟩ (figure 8) we see the bleeding rule at rule 554: /ae̯/ > /eː. This explains not only why we have ⟨*claire*⟩ and not ⟨*glaire*⟩, but also the cases of of CLĀRUM and CLĀVEM. The printout derivation of CLĀVUM >· · ·>⟨*clou*⟩ likewise reveals an earlier bleeding effect as /aw/ passed to / w/. Our proposed rule must thus be placed after these bleeding rules.

Now that we have a proposed rule, its conditioning, and its relative date, we must next justify it (*Test* in PATCH). First, we want to make sure that this is really what the data supports.

As demonstrated in figure 9, in DiaSim we do this by setting our focus point to time step 555, to exclude the words affected by bleeding rules. As expected, our accuracy on that subset is zero. Now that we have zeroed in on the source of error, and inserted a corrective rule (figure 2) at a specified time, the proposal will be validated if our accuracy dramatically improves.

(2) k > g / # __ [+cons] [+lo]

Surely enough, we achieve perfect accuracy for all etyma in the subset except one.[18]

Since we have added a new rule, per PATCH we also justify it. It is easy to see this phenomenon in the context of earlier lenition processes in French, as well as 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. Although in French, the process ceased being productive without diachronic affects on onset consonants, in both Ibero-Romance and Insular Celtic, it continues to operate across word boundaries (Martinet, 1952); the general tendency toward weak word boundaries is known in French is well known, and is realized in sandhi phenomena such as liaison (Cerquiglini, 2018). At the same time, our proposed rule is dated right around the time that the deletion of final consonants was beginning, meaning that many onset clusters would newly become intervocalic where previously they weren't.[19] There is evidence suggesting a related synchronic phenomenon that was once broader in coverage, such as attested k > g substitution in initial /klo-/ (Pope, 1934, p. 96).

It is at this point that one consults other relevant lexical data to corroborate their simulation-guided proposal. In this case, we are supported by philological data from the Old French corpus. Replacement of initial ⟨*c*⟩ with ⟨*g*⟩ in these effected words, is first attested in early 12th century Old French, which is after both bleeding effects on stressed /a/.[20]

Despite this evidence from the early 12th century, the traditional view in the literature has been that such voicing was only a sporadic "tendency" that occurred at the *Gallo-Roman* stage (Pope, 1934, p. 96). Meanwhile, the involved words have been assigned a number of unrelated and often rather convoluted explanations by the scholarship: ⟨*glas*⟩ *alone* is said to be affected by "*assimilation du c initial à la consonne sonore suivante*" (ATILF, 2019c), while analogy is proposed for ⟨*gras*⟩ (ATILF, 2019g), which supposedly cascaded onto ⟨*graisse*⟩ (ATILF, 2019d). The explanation of ⟨*glaive*⟩ relies on both of two proposed language contact effects holding true (ATILF, 2019b), while the voicing in the case of *grille* is not explained at all. Bourciez (1971, p. 146) in fact notes a large subset of our filtered set and includes ⟨*gratter*⟩, from Frankish ⟨*kratton*⟩, a relevant lexeme that agrees with our analysis but was outside our dataset. But, tantalizingly, he does not investigate an explanation using regular sound change, instead attributing the case of *gras* to analogy from *gros*, and leaving the others unexplained.

However, the conditioning and timing we found perfectly divides affected words from all other words with an initial /k/ in Latin which were unaffected, except for CAVEŌLA > *geole* and Celtic *CAMBITU-, which are separately explained by Bourciez (1971, p. 134,142) anyways. Furthermore, our findings were supported by words outside our dataset, such as ⟨*grappe*⟩ and ⟨*gratter*⟩. Thus, for the *Choose* stage of

[18] The exception is CRĀTĪCULAM > ⟨*grille*⟩, due to irregular hiatus behavior after the loss of the interdental fricative /ð/, reflex of /t/. The only other words with EOF sequence /ˌaðˈi/ show different but also irregular behavior. See also CLADĒBON >· · ·>⟨*glaive*⟩ and TRĀDITOR >· · ·>⟨*traitre*⟩, which are similarly nearby a vanishing /ð/, and also display irregularity. These suggest there something *else* to fix, not that our otherwise well corroborated proposal is wrong.

[19] Specific lexemes that tend to precede nouns are especially relevant here: the conjunction ET (</eθ/), the prepositions ⟨*à*⟩ (< /aθ/), the articles ⟨*ce*⟩ (< ⟨*cel*⟩), ⟨*ceci*⟩ and ⟨*ci*⟩ (< /t͡six/), and ⟨*cela*⟩ (*ce* + *là* < /lax/).

[20] The reflex of Latin CRASSIA is still attested as ⟨*craisse*⟩ in 1100 but is attested as ⟨*graisse*⟩ in 1150 (ATILF, 2019d), thus falling into line with ⟨*grappe*⟩ (1121) (ATILF, 2019e; ATILF, 2019f), ⟨*glaive*⟩ (1121) (ATILF, 2019b), ⟨*glas*⟩ (1140) (ATILF, 2019c) and so forth.

PATCH, we uphold our proposed fix.

Pope (Pope, 1934, p. 69) is likely correct that there was at one point a *synchronic* tendency of such form, the *diachronic* effect became phonologized later, late enough to be bled by the loss of /a/ in both of our bleeding cases, hence why we nevertheless have ⟨*clou*⟩ (< CLAVUM), ⟨*clore*⟩ (< CLAUDERE), ⟨*claire*⟩ < CLĀRAM, and so forth.

A possible criticism is that we could in fact be "overfitting" specifications on a sound law to the data. One may note that there would be a double standard in the application of this critique, because the traditional view has enshrined into the academic canon a large number of highly specific sound laws, or even sets of sound laws that explain only a few words, in this case and others[21] To reply, we in turn ask, "what is more likely"? According to the current view in the French diachronic literature, each one of these words is explained by different, highly specific, and sometimes rather elaborate explanations. What is more likely, that each of these words was the result of a different obscure effect perhaps involving two stages of language contact, or that an easily explained shift that we demonstrate here that leaves no exceptions gives a single, simple, and unified explanation?

Nevertheless, it is also difficult to conclusively "disprove" this critique. We do agree that future work should incorporate a measure of *overall complexity* as discussed in section 4.2., but even without this, we maintain that our method actually favors the simplest and most likely explanation much more than the traditional method, because it focuses on finding new rules that correct large numbers of derivations simultaneously whereas the traditional method not only tolerates but turns a blind eye to the proliferation of lexically specific explanations. As such, we propose that adopting CFR alongside traditional methods would in fact work against "overfitting".

8. Conclusion

We maintain that we have clearly demonstrated the utility of computerized forward simulation (CFR) for calibrating diachronic rule cascades. 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 the PATCH 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 Gallo-Roman is the only era without a substantial attested corpus.

The next step for CFR with PATCH is to take it out of the lab and into the field. We strongly advise the adoption of transparent computerized forward reconstruction, for the clear advantages it offers in efficiency, accuracy, accountability, and coverage. Furthermore, for the overwhelming majority of the world's languages which remain vastly understudied, our method offers a way to speed up research into diachronic phonology and by extension phylogeny, allowing us to advance our knowledge further before the majority of them likely become moribund in the next century.

9. Bibliographical References

Adams, J. N. (2007). *The regional diversification of Latin 200 BC-AD 600*. Cambridge University Press.

ATILF. (2019a). http://atilf.atilf.fr/. Accessed December 16 2018.

ATILF. (2019b). glaive. https://www.cnrtl.fr/definition/glaive/. Accessed August 29, 2019.

ATILF. (2019c). glas. https://www.cnrtl.fr/definition/glas/. Accessed August 29, 2019.

ATILF. (2019d). graisse. https://www.cnrtl.fr/definition/graisse/. Accessed August 29, 2019.

ATILF. (2019e). grappe. https://www.cnrtl.fr/definition/grappe/. Accessed August 29, 2019.

ATILF. (2019f). Grappe : Attestation dans frantext. http://atilf.atilf.fr/scripts/dmfAAA.exe?LGERM_FORMES_LEMME;FLEMME=GRAPPE1;FRANTEXT=1;XMODE=STELLa;FERMER;ISIS=isis_dmf2015.txt;MENU=menu_dmf;OUVRIR_MENU=1;ONGLET=dmf2015;OO1=2;OO2=1;OO3=-1;s=s133936ac;LANGUE=FR;FERMER. Accessed December 23, 2019.

ATILF. (2019g). gras. https://www.cnrtl.fr/definition/gras/. Accessed August 29, 2019.

Banniard, M. (2001). Causes et rythmes du changement langagier en occident latin (iiie-viiie s.)(causes and rhythms of language change in the latin occident [3rd-8th centuries]). *Travaux Neuchatelois de Linguistique (Tranel)*, 34(35):85–99.

Borin, L. (1988). A computer model of sound change: An example from Old Church Slavic. *Literary and Linguistic Computing*, 3(2):105–108.

Bourciez, E. (1971). Phonétique française.

Brunot, F. and Charlier, G. (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.

Burton-Hunter, S. K. (1976). Romance etymology: A computerized model. *Computers and the Humanities*, 10(4):217–220.

Cerquiglini, B. (2018). *Une langue orpheline*. Minuit.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper& Row, New York.

Diefenbach, L. (1831). *Ueber die jetzigen romanischen Schriftsprachen, die spanische, portugiesische, rhätoromanische, in der Schweiz, französische, italiaänische nd dakoromaische, in mehren Ländern des östlichen Europa, mit Vorbemerkungen über Entstehung, Verwandtschaft usw dieses, Sprachstammes*. Ricker.

Diez, F. (1836). Grammatik der romanischen sprachen, 3 vols. *Bonn: Weber (3rd ed. 1870–1872)*.

[21] Indeed, ⟨*glas*⟩ is supposedly explained by a lexically specific rule that only affected other words indirectly through sporadic analogy, despite that rule working better as a broader and regular rule, as we have just demonstrated. This, plus all the other lexically specific explanations, is not in line with Occam's razor at all.

Dunn, M. (2015). Language phylogenies. In *The Routledge handbook of historical linguistics*, pages 208–229. Routledge.

Eastlack, C. L. (1977). Iberochange: a program to simulate systematic sound change in Ibero-Romance. *Computers and the Humanities*, 11(2):81–88.

Fouché, P. (1961). *Phonétique historique du français: Les consonnes et index général*, volume 3. Klincksieck.

Gildea, D. and Jurafsky, D. (1995). Automatic induction of finite state transducers for simple phonological rules. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 9–15. Association for Computational Linguistics.

Gleason, H. A. (1959). Counting and calculating for historical reconstruction. *Anthropological Linguistics*, pages 22–32.

Hall, T. A. (2007). Segmental features. *The Cambridge handbook of phonology*, pages 311–334.

Hartman, L. (2003). Phono (version 4.0): Software for modeling regular historical sound change. In *Actas: VIII Simposio Internacional de Comunicación Social: Santiago de Cuba*, pages 20–24.

Hock, H. H. (2009). *Principles of historical linguistics*. Walter de Gruyter.

Hombert, J.-M., Mouele, M., and Seo, L.-W. (1991). Outils informatiques pour la linguistique historique bantu. *Pholia*, page 131.

Jäger, G. (2019). Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182.

Johnson, C. D. (1972). *Formal aspects of phonological description*. Mouton & Co. NN.

Kaplan, R. M. and Kay, M. (1981). Phonological rules and finite-state transducers. In *Linguistic Society of America Meeting Handbook, Fifty-Sixth Annual Meeting*, pages 27–30.

Kiparsky, P. and Good, J. (1968). Linguistic universals and language change. *Universals in linguistic theory*, pages 170–202.

Kondrak, G. (2002). *Algorithms for language reconstruction*. Ph.D. thesis, University of Toronto.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, 37(3):273–291.

Lodge, R. A. et al. (2004). *A sociolinguistic history of Parisian French*. Cambridge University Press.

Lodge, R. A. (2013). *French: From dialect to standard*. Routledge.

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

Lusignan, S. (1986). *Parler vulgairement: les intellectuels et la langue française aux XIIIe et XIVe siècles*, volume 1. Librairie philosophique J. Vrin; Montréal: Presses de l'Université de Montréal.

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, P. (1901). *Petite phonétique du française prélittéraire (VIe-Xe siècles)*. B. Veith.

Marr, C. and Mortensen, D. (2020). Large-scale computerized forward reconstruction yields new perspective in french diachronic phonology. unpublished.

Martinet, A. (1952). Celtic lenition and Western Romance consonants. *Language*, 28(2):192–217.

Martinet, A. (1970). Economie des changements phonétiques.

Meyer-Lübke, W. (1899). *Grammatik der romanischen Sprachen*, volume 1. Georg Olms Verlag.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (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*, pages 3475–3484.

Muzaffar, T. B. (1997). *Computer simulation of Shawnee historical phonology*. Ph.D. thesis, Memorial University of Newfoundland.

Noske, R. (2011). L'accent en proto-français: arguments factuels et typologiques contre l'influence du francique. In *Congrès Mondial de Linguistique Française 2008*, pages 307–320. Institut de Linguistique Française, Paris.

Nyrop, K. (1914). *Grammaire historique de la langue française*, volume 1. Gyldendal.

Piwowarczyk, D. (2016). Abstract: A computational-linguistic approach to historical phonology. *New Developments in the Quantitative Study of Languages*, page 70.

Pope, M. K. (1934). *From Latin to Modern French with especial consideration of Anglo-Norman: Phonology and morphology*. Manchester University Press.

Posner, R. et al. (1997). *Linguistic change in French*. Oxford University Press.

Posner, R. (1996). *The Romance languages*. Cambridge University Press.

Posner, R. (2011). 'phonemic overlapping and repulsion revisited. *General and Theoretical Linguistics*, 7:235.

Rey, A. (2013). *Dictionnaire historique de la langue française*. Le Robert.

Short, I. R. (2013). *Manual of Anglo-Norman*, volume 8. Anglo-Norman Text Society.

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

Smith, R. N. (1969). A computer simulation of phonological change. *ITL-Tijdschrift voor Toegepaste Linguistiek*, 5(1):82–91.

Suchier, H. (1893). *Altfranzösische Grammatik*. M. Niemeyer.

Swadesh, M. (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, C. (1881). *De la prononciation française depuis le commencement du XVIe siècle: d'après les témoinages des grammairiens*, volume 1. Impr. nationale.