Towards a computational model of long-term diachronic change: simulating the development of Classical Latin to Modern French

Clayton Marr
Vassar College

Follow this and additional works at: http://digitalwindow.vassar.edu/senior_capstone

Recommended Citation
http://digitalwindow.vassar.edu/senior_capstone/710

This Open Access is brought to you for free and open access by Digital Window @ Vassar. It has been accepted for inclusion in Senior Capstone Projects by an authorized administrator of Digital Window @ Vassar. For more information, please contact library_thesis@vassar.edu.
Towards A Computational Model of Long-Term Diachronic Change:

Simulating the development of Classical Latin to Modern French

Clayton Marr

Vassar College

May 1, 2017

Senior thesis submitted for fulfillment of the requirements for the Independent Major in Computational Linguistics

First reader: Jan Andrews

Second reader: Nancy Ide

Table of Contents:

2 ... 1. Abstract
3... 2. Introduction: Historical Phonology
4... 3. Major controversies in Historical Phonology
5... 3a. How does sound change? Neogrammarian regularism confronts lexical diffusionism
11... 3b. Why does sound change? Strict internalism confronts language contact effects
14... 4. Introduction: The French Language and Historical Phonology
16... 4a. Diachronic sound change and Neogrammarian regularity in French
17... 4b. Historical dialect and register relations in French
21... 4c. Language contact effects in French
28... 5. Computational models of diachronic change: reconstruction and simulation
30... 6. Description of this language simulation package
31... 6a. Representation of phonological/phonetic units
33... 6b. A class hierarchy of Phone objects
33... 6c. Modeling of phonological features within Phone objects
39... 6d. Representation of the lexeme
39... 6e. Alteration – an object representation of diachronic sound change
42... 6f. Parsers for extracting phonological structure form String objects
43... 6g. The Lexicon object
44... 7. An automated simulation of the development of Latin into French
46... 7a. Current performance of the simulation system: successes and remaining issues
56... 8. Where to go from here—possible routes of expansion and improvement
56... 8a. Incorporation of non-phonological information into lexemes
58... 8b. Testing etymologies for French words
60... 8c. Large-scale statistical analysis of the French simulation’s results
62... 8d. Analogous automated simulation for Latin loans in Albanian
65... 8e. Application of Machine Learning algorithms
68... 9. Bibliography

1. Abstract
For the culmination and conclusion of my independent major thesis in linguistic computation, I present a package of computational architectures intended for use in the study of Historical Phonology. Included in this package are (1) a class hierarchy for representing phone units, (2) a class for representing lexical phonologies which allows the future possibility of also implementing representation of morphological, syntactical and semantic information, (3) a class representing the lexicon of a language at a given time, including all words in that language being simulated, (4) parser classes that can convert text to phonological structures, (5) an automated simulation of the phonological developments undergone by words in the French language as it developed from Gallo-Romance to its present form, organized by conventional periodization of the language, with methods to implement those periodic changes upon a lexicon object that is given as input, and (6) testers and a demonstrational class that prints the development of words at each stage. It is my belief that this project can be expanded in multiple fruitful directions that may ultimately yield interesting insights on various topics in Historical Phonology.
1. Introduction: Historical Phonology

Within linguistics, the study of language and languages, lies phonology, the study of the sounds employed by language(s) – how are words pronounced, which sounds and sound contrasts function as units, which sounds are present in a language’s phonological inventory, and so on. Historical Linguistics, meanwhile, is the study of how language changes over time. Implied in this definition is the inclusion of both small scale changes in grammar and phonology, but also large scale analyses of how one language may have developed from another, and how multiple languages may have arisen divergently from the same source, as with the various Romance languages from Latin. At the intersection of the two disciplines of Phonology and Historical Linguistics naturally lies Historical Phonology—the study of how the arsenals, roles and distributions of sounds change (or may have changed) over time as a language evolves.

Historical phonology first emerged in Germany in the 19th century, with the writings of a school of historical phonologists known as the Neogrammarians. They argued that linguistics was a science, and thus driven by mathematical principles. Their writings engendered an explosion of linguistic discovery, and historical phonology was arguably the field of their most transformative successes. The idea that historical sound change was driven by scientific and functional principles which operated according to rules defined by the nature of a sound and its context was essentially the foundation of historical phonology. With this, we became able to trace what changes transformed Latin into French (or Sanscrit into Hindi, and so on) and when they occurred. We also learned how to use these sound shift functions in reverse: put the function in reverse to reconstruct the unknown, unattested parent of a living language (this is called a “proto-language”). This method is used at identifying what is the parent of a living language—as is the case with Basque and the ancient language Aquitanian, which has now been found to resemble “Proto-Basque” – and has also been used to help determine whether distant languages are related, by seeing if their proto-languages are similar. It is also used to pinpoint when
certain developments occurred, by reconstructing the proto-language of a given language for different periods of time. The conclusions drawn from these analyses have led to much larger scale discoveries in Historical Linguistics, such as the discovery of a wide spanning Indo-European family with branches of Romance, Germanic, Albanian, Celtic, Indo-Iranian, Anatolian, Balto-Slavic and so on, spanning from Iceland to Sri Lanka.

3. Major controversies in Historical Phonology:

Despite the massive predictive power that the philosophy of the Neogrammarians gave to Historical Phonologists, their account of how sound change occurs has been controversial for the last few decades, with Neogrammarian views of *exceptionless* sound change being challenged by the theory of *lexical diffusion*, as well as various alternative and compromise theories. Another controversy concerns the reasons sound change may occur: does sound change within a language only occur due to reasons *internal* to that language (its grammar, the distribution of sounds, the arsenal of sounds, et cetera), or can it also rise due to *external* factors, namely influence (*contact*) from other languages?

If the Neogrammarian assumption that sound change is rule-driven and algorithmic is true, then naturally it makes sense to craft an algorithm to represent it. Thus, my goal in this thesis project has been to begin the crafting of a large-scale computer program to do just that, for the sake of testing different theories about Historical Phonology. This is an ambitious task, to say the least. Before discussing what progress has been made toward constructing such a system, it is important to discuss in greater detail both of the major relevant controversies.

3a. How does sound change? Neogrammarian regularism confronts lexical diffusionism

As discussed earlier, the early Neogrammarians, who mostly worked in the German city of Leipzig, based their Historical Phonology program on the belief that sound change affected specific segments
often in specific contexts and operated at a certain time, leaving no such segment unaffected in the relevant context in the language at that time, as part of sound change rules that could be established through Historical Phonological research. Despite initial skepticism, most of the early opposition to the Neogrammarians was converted.

Perhaps the archetypical example of the Neogrammarian sound shift is Grimm’s Law [Campbell 2004 : 49], representing a historic shift in proto-Germanic that differentiated the Germanic languages from their other Indo-European cousins, described in each of its aspects below (with sounds represented in IPA, International Phonetic Alphabet):

- Voiceless stops (p,t,k,kʷ) became voiceless fricatives (φ>f, θ>h, x>xʷ>f)

- voiced stops (b,d,g,gʷ) became voiceless stops (p,t,k,kʷ)

- aspirated voiced stops (bʰ, dʰ, gʰ, gʷʰ) became normal voiced stops (b,d,g,gʷ)

This led to sound correspondences between Germanic and non-Germanic Indo-European languages, where Germanic languages have /f/ where non-Germanic languages typically have /p/: for example for English foot there is French pied and for English father there is French père.

Neogrammarians held that sound change should be regular and exceptionless, spontaneously affecting all relevant segments in relevant contexts. If exceptions were found, ultimately there were typically new, regular, sound changes that were discovered to explain them—as happened with Verner’s Law, which was discovered to explain every remaining exception to Grimm’s Law [Millar 2015:209]. There were also other explanations for the scattered cases of irregularity, all of which were reducible to the effects of some nonphonological interference. Loanwords, of course, don’t follow the rules if they haven’t been in the language at the times a sound shift operated: hence Standard German Butike would seem to violate the High German consonant shift (it should have been *Busiche, or perhaps *Büsiche by
ümlaut) but this is only because the word was loaned from French. Similarly, loans between dialects can occur: Standard German *Ratte* violates a rule that geminate (double) stops (like tt) became affricates (like ts), but this is because the word *Ratte* was borrowed from the Low Saxon dialect of German that had not been affected. Another source of irregularity is morphological analogy: the English plural for *cow* was once *kine*, but it became *cows* under the influence of the singular. Other common sources of irregularity are avoidance of homophony, hypercorrection (i.e. the spelling pronunciation of English *salmon* and *falcon* with an /l/) and the influence of onomatopoeia. Substrate and superstrate influences can also cause irregularity – hence the corruption of Latin *ranunculus* (tadpole), which would regularly render *renoncle*, to French *grenouille* (frog) under the influence of Gaulish *craxaulios* (diminutive on frog, lit. “froggie”), which may be responsible the for insertion of the initial velar stop and the palatalization of the /l/ to -ille /j/.

While various attempts to debunk the Neogrammarians’ functionalist assumption failed markedly for about a century, in the second half of the 20th century a more serious challenge emerged with the “diffusionist” theory, that held that rather than affecting all similar segments simultaneously, sound changes “diffused” across phonologically similar words in a language until they had affected all relevant sound segments. This theory originated in 1969 with a study of Teochew, a Sinitic language [see Wang 1969]. Although Wang’s original study of Teochew is now widely criticized as vastly flawed [Egerod 1982, Mazaudon 1994] and was not accepted by prominent Sinitic phonologists [see Pulleyblank 1977], the theory caught on. It acquired supporters and purported evidence, such as field data on the supposedly ongoing propagation of American a-tensing [see Labov 1994], the emergence of diatonic verb/noun pairs [see Sherman 1973] and the foot-strut split [see Kiparsky 1994]. Essentially, this theory of lexical diffusion suggested that rather than being exceptionless, sound change instead spread word-
by-word across a language’s lexicon\(^1\). Furthermore, diffusionists propose that apparent exceptions often represent cases where the sound change halted before it had affected the whole lexicon. Lastly, the core diffusionist claim, at least in its “strong” version as articulated by Phillipps [2006], was that lexical diffusion is in fact the default form of change and cases that appear to be exceptionless are merely cases where lexical diffusion proceeded to affect the entire lexicon.

As a result, in the late 20\(^{th}\) century, one might have thought Historical Phonology as a discipline was in crisis. If one accepts diffusionist theory, then the credibility of a large bulk of the historical linguistic work, especially in language reconstruction (which had been based on Neogrammarian assumptions), instantly becomes jeopardized. And yet, the predictive power of the assumption of regularity has been lately reinforced by new discoveries. For example, in the early 20\(^{th}\) century, based solely on the Neogrammarian assumptions, Bloomfield predicted the existence of a fifth fricative voiceless velar cluster in Proto-Central-Algonquian, even though all its known descendants at the time had only four such clusters. Many decades later, his Neogrammarian prediction was vindicated: Swampy Cree was discovered, as a Central Algonquian language that preserved a fifth voiceless velar cluster matching his predictions. Because of remarkable discoveries like this, the utility of Neogrammarian regularity as a methodology, especially in reconstruction, remains widely defended [see Fox, Anthony, *Phonological Reconstruction* in Honeybone 2015]. Most phonologists still adhere to the Neogrammarian model, alternatingly disputing and accommodating diffusionist positions, with many attempts to formulate compromises between diffusionist and Neogrammarian theory [Labov 1994, Kiparsky 1988, McMahon 1994].

---

\(^1\) Spreading word-to-word should be distinguished from spreading speaker-to-speaker, which is widely agreed to occur by all historical phonologists, regularists and diffusionists alike.
Today, although lexical diffusion is maintained by a committed core of supporters and a much larger number of linguists who partially accept it, Neogrammarian ideas are making a comeback, and cases of supposed lexical diffusion are being disproven, though some stubbornly remain. A common theme in these refutations is that what was assumed to be lexical diffusion turns out to be, just as Leipzig’s Neogrammarians would predict, non-phonological influence. Wang’s claimed “tone-split” in Teochew [see Wang 1969] turned out to be a case of borrowing from a prestige register [see Egerod 1982], while his and Chen’s claim of diffusion in the emergence of English diatonic stress pairs [see Chen & Wang 1975, also Kreidler 1987] look like a classic Neogrammarian-friendly case of analogy\(^2\) to even Phillipps [2006], a maximal diffusionist. Some proposed cases of lexical diffusion also appear better explained as stable variation\(^3\) [Abrahamowicz 2007] while Kiparsky in 1995 claimed could all be considered variants of analogy. Abrahamowicz’ paper is particularly notable since it counterexplains the “frequency effects” that form a major part of diffusionist theory.

Cases where supposed violations of Neogrammarianism turn out to have originated from sociolinguistic interference are particularly interesting. Janda and Joseph [2003] present a unified “Big Bang model” of sound change, where sound change is initially regular and solely phonetically concerned, but may be later changed so as to lose its regularity in accommodation of non-phonological factors, typically sociolinguistic, often ultimately causing irregularity. Mazaudon [1994], meanwhile, makes a persuasive case, citing a study on Czech [Andersen 1973], that what may appear to be cases of diffusion are in fact cases of sound changes where speakers tried to reverse sound changes for various sociolinguistic reasons. Just like Egerod observed with register borrowing [1982], this non-phonological process may indeed result in lexical splits and seeming irregularity, but it only does so because it is a

\(^2\) i.e. the influence exerted by the presence of predominant morphological patterns

\(^3\) “Stable variation” meaning there is in fact no diachronic change present, but rather constant and unchanging sociolinguistic variation such as between the historic and current uses of -\textit{in}’ for -\textit{ing} segments
sociolinguistic rather than phonological process. However, today, the “Neogrammarian controversy” (as Labov calls it) still rages, in part because the interference theories posited by linguists like Mazaudon, Egerod, and Janda have not (yet) explained all the purported evidence for lexical diffusion. However, it will be interesting to see what in the future new computational models may contribute to this debate.

3b. Why does sound change? Strict internalism confronts language contact effects

The early Neogrammarians did not interest themselves in the possibility that diachronic changes in one language could be brought about from the influence of another. On the one hand, the idea that languages could influence each other was not alien. That words could be borrowed from one language to another is painfully obvious to all linguists and many non-linguists, and indeed many Neogrammarians such as Meyer-Lübke spent much time tracing such borrowing. German philologists from the same schools claimed that Romanian was “semi-Romance” and argued over whether French was more “Celtic” or “Germanic”, with massive non-Latin influence being accepted as given. However, while the early Neogrammarians believed that lexical items could be imported, they did not see any possibility for the similar causation of “structural changes” in phonology or syntax. These internalist assumptions were almost immediately challenged [Craddock 1969 : 18].

This challenge came from Graziadio Isaia Ascoli, an Italophone Jewish linguist from multiethnic Friulian town of Gorizia, where the speech one heard on the street was equally likely to be Slovene, Friulian, Venetian, German or Italian. Although Ascoli is now considered a Neogrammarian due to his endorsement and contribution to the theory of regular sound changes, he was critical of the internalist assumptions of his German colleagues. For Ascoli, these assumptions contradicted an obvious fact he experienced in everyday life: if one speaks multiple languages, their speech in one language may be influenced by the other. Furthermore, if large numbers of speakers of one language also speak another, over time influences of the latter language accumulate structurally as well as lexically in the latter. In
1881, Ascoli published his *Lettere glottologiche*, forming the scientific foundations substrate theory, and igniting a controversy in Historical Linguistics that continues today [Sala 2013: 192].

Ascoli divides what came to be called “language contact” effects into two categories. The first, arises in a scenario where a socioeconomically privileged language replaces a less advantaged one, but speech habits from the replaced language, called the *substratum*, are retained in the newly adopted language, and ultimately shape its local development, often leading to the emergence of new dialects and even languages. Along these lines, Ascoli argued that various Romance languages spoken in formerly Celtic areas, such as French as well as the Friulian of his homeland, had developments separating them from the rest of the Romance languages that originated from the importation and retention of Celtic speech habits. Ascoli also noted the reverse phenomenon, the influence of an elite language as its speakers switch to the language of the laypeople (as seen with the Germanic languages of invader elites in the post-Roman area), and called it a *superstratum*. A third category was added later – the *adstratum* – which accounted for cases of language contact arising from more equitable situations of multilingualism.

While it is now accepted by most scholars that change can arise from language contact, the majority of proposed cases remain controversial to varying degrees. In the most extremely generous cases, obscure words, sound shifts, and grammatical changes are attributed to the influence of hypothetical evanescent languages [Sala 2013: 194], of which we know nearly nothing. Internalists use these issues of proof to discredit theories of influence as “simplistic ideals of monogetic holism” that linguists should entertain only as a “last resort” [Mees 2003: 33], and structuralists propose alternative models of language development that rely only on internal factors. Supporters of substrate and superstrate theories, meanwhile, are likely to respond that a lack of proof does not equal conclusive negative evidence, and note the oft-neglected fact that the internal explanations are often also devoid of proof; indeed for many diachronic changes, it is essentially impossible to “prove” that any one
explanation is the real “cause” of a shift, due to the impossibility of applying the empirical method back in time.

4. The French Language and Historical Phonology

French is one of the most widely spoken languages in the world, and it is a member of a family of languages known as the “Romance languages”. Fellow members of the family include the state languages of Spanish, Portuguese, Catalan, Italian and Romanian, as well as a host of other less privileged languages such as Sardinian, Venetian, Dalmatian and many others. Within Historical Linguistics, it is common practice to group languages into “families” based on “descent” from a common “parent” language, which often engenders shared features – so Russian and Polish are Slavic, Hebrew and Arabic are Semitic, Mandarin and Cantonese are Sinitic, and so forth. Each Romance language is included because it is thought to have developed from their common parent, termed proto-Romance. Although the exact relation of Proto-Romance with Latin is disputed, it is universally accepted that they are closely tied. Because of how well-known Latin was (and is) and that we know much less about the proto-languages of other families, Romance languages have often been used to vindicate methods in Historical Phonology [Posner 1996: 97].

Romance philology is old and said to have been started in 1304 by Dante, who divided the languages into groups based on their words for “yes” [Posner 1996: 2]. It has indeed been critical in the discussion of both theories discussed here: early Neogrammarians such as Gröber and Meyer-Lübke focused much

---

4 Nevertheless, this has not stopped linguists from trying, and there have been some attempts to get scientific support for certain theories, such as the proposed connection of diachronic lenition processes to drunk slurred speech, leading to studies which observed the speech of intoxicated subjects [see Kaplan 2009].

5 It is thought that “Proto-Romance” could be considered either Late Latin, a late form of colloquial (“Vulgar”) Latin, a closely related language to Latin within the Italic family, or collections of colloquial lower class dialects that tended to influence each other [For a fair summary of some of these possibilities, see Posner 1996: 97-103]
of their discussion on developments in Romance languages (especially French), as did Ascoli [Posner 1996: 4].

Of the Romance languages, French is the second most spoken (after Spanish), and possibly the first most studied. Because Latin was a well-known language for much of early modern academia and that many if not most of them also spoke French, French is one of the best studied languages in Historical Linguistics. It also provides an interesting case study, due to the roles of a fairly well-studied superstrate (Frankish) and a fairly well-studied substrate (Gaulish), the oscillation of historical relations between “Standard French” and its related dialects, the large corpus of Old French text and how extensive the sound changes that occurred in French were, being the most extensive of all major Romance languages. For these reasons, French has been frequently recruited as a testing ground for both the “how” and the “why” controversies of sound change presented in this paper.

4a. Diachronic Sound Change and Neogrammarian Regularity in French.

While many important early (typically German) Neogrammarians were also experts in French, their contemporaries in France at the time were skeptical, often having backgrounds in literature and being very focused on “communicative creativity”, the “vagaries of fashion” and “individual idiosyncracies” [Posner 1997: 221]. At the same time, the fact that French is very well attested throughout much of the Middle Ages, with extensive writing from the 16th century onward by the French about their own language and its history [see Pope 1934] including sound changes ongoing at the time makes French very fertile ground for testing theories about the initiation, propagation and resolution of sound change. It has been noted by some that the known histories of some languages and language families seem more conducive to Neogrammarianism than others, and French and Romance are often primary examples for those that were friendly to Neogrammarianism.
4b. Dialect and register relations in French

The historical relations between dialects and registers can have a profound effect on conditioning the spread of sound shifts, and they sometimes modify their ultimate results [Egerod 1982; Mazzola 2012; Andersen 1973]. Sociolinguistically conditioned stylistic restorations are also an important part of this phenomenon, as Mazaudon [1994] adroitly notes. The results of all this often do not look Neogrammian at first glance—but what “diffuses” is not actually the sound change, but rather the restoration. This forms a critical part of Janda & Joseph’s [2011] “Big Bang Theory of Sound Change”—sound changes start out as strictly Neogrammian and solely phonetically or phonologically driven, but their spread (across both geographic and linguistic environments) and even their effects may ultimately be warped by non-phonological factors.

The fact that they end up looking non-Neogrammian is what makes the effects of such factors (often sociolinguistic) so important for the debate between regularist Neogrammianism and lexical diffusionism. The linguistic history of French is replete with well-known instances of such effects. The vast knowledge about the sociolinguistic context in the case of historical French makes French historical phonology a great testing ground for what we can expect the effects of sociolinguistic interference to look like. Knowing that — and perhaps even being able to simulate it! — would be a great step forward to resolving the regularist vs. diffusionist controversy.

In the late Roman and Migration Age, the propagation of language contact effects (see following section) varied between both geographic dialects and class registers. Parisian speech had not yet gained national dominance, and the speech of the Frankish-origin rulers may have resembled that of the “more Frankish” northwestern regions of Picardy, Champagne and Lorrain [Mazzola 2012; Pope 1934]. In most regions, lenition (consonant softening) preceded unstressed vowel loss (apocope), with one of the
results being that the final palatal consonants in words like *manica* ended up voiced, as *-ge /ʒə* (modern /ʒa/), but in these “more Frankish” regions, it was the reverse, causing the final palatal to be devoiced, as *-che /ʃə* (modern /ʃa/). Due to association with the ruling class, however, these dialects gained prestige, and French speakers outside their native regions tried to imitate them to sound more prestigious ultimately adopting, but inconsistently so, the pattern of final palatal endings, resulting in doublets like *grange* and *granche* [Mazzola in Arteaga 2012:162-3].

Later periods of French also offer insight into the workings of interdialect influences in phonology, not least because we have six centuries of literature of French people discussing their language which touches upon the subject. From this, we know that it was fashionable to use provincial or dialectal words in 1500s [Posner 1996: 87], and that the situation became the opposite in the modern era, where the use of dialects (*patois*) became intensely stigmatized.

The Middle French period was another period of significant difference between upper class and lower class speech, but this time the upper class seems to have used Latin features, rather than Frankish, to assert prestige [Pope 1934: 38-41]. The upper class was increasingly educated but still wrote in Latin, and conducted much trade with Italy, and mockeries of upper class speech focused on how Latinate and un-Gallic they sounded. Certain aspects of court pronunciation seemed to make their speech sound more like Latin, or more like Italian [Pope 1934: 31], such as the sociolinguistically driven shift of /we/>/e/, which being sociolinguistic indeed led to non-Neogrammrian vocabulary splits like *François* and *Français*.

The rebuttal using sociolinguistically driven stylistic restoration for supposed cases of lexical diffusion, to the best of my knowledge, dates back to 1982 with Egerod’s response to Wang, and was only first formulated into a unifying theory over a decade later [Mazaudon 1994]. And yet, one of the most of striking things I found when doing the literature research for this thesis was how the fact that
socially driven restorations can occur and cause irregularity was not in any way novel in French historical linguistics. Indeed, Pope, writing in 1934, discusses numerous such effects that would easily demonstrate the relevant effects of restoration. Sound change has no memory, but orthography, grammarians and in some cases [Andersen 1973] elderly people do. A perfect example of these effects is the trajectory of the Middle French alteration of r in weak positions (intervocalic or unstressed word-final) [Pope 1934 : 156-159]. Initially, just as the theory (published 70 years later!) of Janda and Joseph [2004] would predict, the shift originated as a “buzzing” of the alveolar trill /r/ into a fricative, /z/. Given that during the same period French was experiencing another round of intervocalic lenition affecting /s/ and /l/ [Pope 1934: 82-83], it seems natural to view the “buzzing” of /r/ into an easier /z/ sound as part of that phenomenon. Slightly later, word-final /z/ was also increasingly effaced when not followed by a vowel, explaining the effacement of /r>z/ in that position. Also quite predictably, this “buzzing” r was deleted before every other alveolar consonant then present in the language: /l/, /s/ and /z/. Thus, originally, the shift was completely phonetically likely and regular. Then, non-phonological factors interfered. Grammarians wrote “persistently and vehemently” against the shift, causing the educated classes to try to avoid it—thus triggering an (attempted) stylistic restoration. The effects included the hypercorrective reversal of the shift where it had not occurred: Nemous became Nemours permanently. In other cases, the attempted restoration missed targets, such as chaise which became widely accepted and replaced chaire (“chair”). After four centuries of upper class resistance, in one context the shift was finally accepted and canonized by 1700 (after which it was no longer productive): /r/ was consistently deleted word finally only after the vowel /e/ and only in multisyllabic words (not typologically likely at all), whereas elsewhere the remaining effects of the shift were sporadic. Looking at the data here, a diffusionist would likely see “frequency effects” as most French verbs end in -er, but Pope clearly lays out the historical and linguistic reasons why the shift survived in some phonological contexts but not others, and they don’t include frequency effects. What is astonishing is that even
French opponents of lexical diffusion theory, like Mazaudon, seem to have omitted reference to such examples in French historical phonology as arguments, and much less to study them to develop linguistic knowledge on the properties of sociolinguistically-motivated stylistic restoration.

4c. Substratum and superstratum in French

The French detected early on that their language stood apart from other Romance tongues, and were quick to postulate non-Latin providence, be it “Germanic”, “Celtic” or even “Greek”. German philosophers similarly concurred, and argued amongst themselves whether French was more “Celtic” or “Germanic” [Posner 1996:3]. Neither identity has been without controversy, and the various sides have oft mapped onto political divides: France’s clergy and nobility identified with the Franks, while the common people, leftists and populists identified with the Gauls [Krepps 2010: 2-6]. Meanwhile, after the Franco-Prussian war, prominent right-wing Parisian Romance linguists begrudgingly admitted the possibility of “barbarous peasant” Celtisms in French, but thought Ascoli’s theory was not worth discussing, preferring to dismiss it and instead emphasize France’s supposed superior (over Germany) Roman pedigree of civilization [Hoyt 2006:94-98].

Today, French is a widely cited case of both substrate and superstrate effects, and indeed was central to the very foundation of the substrate theory. Indeed, French can be called an ideal scenario for the theory, and even generally skeptical authors like Sala [2013] and Mees [2003] concede there are likely some Celtic substrate effects in French, and definitely Germanic superstrate effects, though they view them as more restricted than other authors do.

---

6 For the sake of fairness to Mazaudon, one should note that her paper was primarily made in response to the claims of diffusionists, and the phonological history of French typically is not brought up for examples by diffusionists.
The Frankish superstrate comes from the speech of the invading Franks, who founded the French kingdom and gave it their name. The Franks retained their language until the 9th or 10th century [Holmes 1938: 29] when they finally adopted the Romance tongue of their surroundings. Given that Old French is one of the best attested medieval languages, that Frankish was fairly well attested with a living descendant (Dutch), and that both members of well-studied language families makes French a near-ideal case of proposed historical superstrate effects.

The substrate, meanwhile, is mostly Gaulish, the indigenous Celtic language of the region, which was called Celtica, Gallia or Gaul before it was France. “Celtica” referred only to France’s northern half\(^7\), whereas Gallia referred to the whole. “Gaul” meanwhile, was the Frankish name for the land, “Gaulish” its inhabitants\(^8\). Substantial population continuity between ancient Gaul and modern France is not widely controversial\(^9\). Today, the identification of the common Frenchman with the Gaul can be seen in popular media such as the children’s comic, *Asterix*. While language replacement by Latin in Italy and Iberia happened during the height of the Empire, in the 200s primary sources state that Gaulish is spoken in both Gallia and northern Italy whose people both considered themselves kin, while a 228 degree allows Gaulish in court [Stifter 2012: 110]. It seems it wasn’t until the late 5th century that Auvergne’s nobles are written to have set aside their “barbarous Celtic tongue” [Apollinaris 474], while the commoners probably retained Gaulish longer. The last primary sources in the late 6th century seem to link Gaulish to rural paganism [Gregory of Tours 575]. It is unclear whether these sources actually refer to persistent spoken Gaulish [Stifter 2012 : 110-112], but as Kerkhof notes there is enough

\(^7\) I.e. the half that gave us the French language; the southern half, known as *Gallia Narbonensis*, gave us instead the Occitan language of the troubadours. The name “Celtica” was not forgotten : in Old French, “celtique” was used to refer to the French people.

\(^8\) Etymologically unrelated to “Gallia” but instead from *walha*, the same root as “Welsh” in English

\(^9\) As per Gaul’s high pre-Roman population density and the small male-heavy Frankish coterie of Germanic invaders. In contrast, Dacia (Romania) may have seen an ancient ethnic cleansing by the Roman state and/or later Medieval Vlach-Romanian immigration. Worse still, the issue is intertwined with emotive ethnic nationalisms.
evidence, including evidence of late Gaulish phonological developments (some resembling those in
Gallo-Roman) and Gaulish words in the *Lex Salica* to posit a survival at least in northern, mountainous
and/or remote areas into the migration age of Medieval Europe and significant period of
bilingualism[2016], with it being difficult to confirm or disconfirm any exact date of the
extinction of Gaulish as it may have survived “unseen” in remote valleys for quite some time, even
“centuries” [2000: 12].

In the lexical realm, there was a massive influx of Frankish vocabulary [1996: 250; Pope
1934: 14], including some things as basic as colors (*blanc* “white”, for example). The confirmed Gaulish
vocabulary retained in French is less, but includes significant core vocabulary. There is a problem of
attestation, and large parts of the French lexicon remain etymologically “unknown”; some of this is now
being attributed to Gaulish due to new discoveries of Gaulish inscriptions and/or new Proto-Celtic
reconstructions, especially “rural” words like *mouton* (sheep), from Gaulish *multonis* (cf. Welsh *molt*).
There are also known calques from Gaulish, such as *aveugle* (“blind”) and *oui* (“yes”)[10]. Recent
simulation data also suggests a considerable number of shifts noun gender arose from Gaulish influence,
in which the gender of a Latin noun was changed to match its gender in Gaulish [2013].

Frankish also transformed French morphology, adding various new morphemes of which some
are still productive (-*ais* < Frank -*isc*, -*enc* < Frank -*ing, -art, mes*, etc) [1934: 14-15], while the
known effects of Gaulish here are less, including the intensifier prefix *re*- [2004:294-5]. Either
Frankish [1934: 16] or Gaulish [2004: 26] could be source of the French 2nd person plural

---

10 *Aveugle*: from Lat. *ab oculis* calquing on Ga. *exs ops*, one of the best attested cases of an ancient calque [1995:158; 2007:279-89]. *Oui* : from Lat. *hoc (est) ille* “this is the one”, the rare and very Celtic affirmative structure underlying the words for “yes” in all surviving Celtic languages, Fr. *oui* and Occitan *oc* [2007:106; 1997:15]
conjugation -ons, while it is Gaulish likely underlies feminine plural -es (<Celtic -as\textsuperscript{11}, not Latin -ae) [Savignac 2004: 26; Adams 2007: 287].

French syntax was also heavily influenced by Frankish, ranging from adjective placement to the temporary adoption of the verb-second syntax system typical of Germanic languages, and that French became a pro-drop language [Posner 1996: 53, 248-249]. Traits attributed to Gaulish include clefting, periphrastics to indicate and the use of à (“to”) to denote possession [Savignac 2004:26; Matasović 2007:106-107; Filppula 2008:77-82].

In the phonological realm, the Frankish superstratum gave Old French a “strong expiratory” and heavy syllabic stress system that was to play a crucial role in its development. This caused, as in German, the devoicing of final consonants and the closing of unstressed vowels [Pope 1934: 15; Posner 1996: 246-248]. Gaulish, meanwhile, seems to be associated with tendencies to blur segmental boundaries, still characteristic of modern French phenomena like liaison, enchainment and elision [Guiter 1995]. Another major effect seems to be the migration-era two waves of “Celtic lenition” experienced by French [Pope 1934: 6, 136-140; Posner 1996:237], a “strong diagnostic Celtic trait” [Ball 1995:7-13,22]. A convincing (though not conclusive?) model exists how the phenomenon originated internally in Celtic languages as a drag process caused by their lack of phonemic spirants [Charles-Edwards 2013:79-80], becoming regular in the late classical era [Stifter 2008: 107] and only being phonemicized by later degemination and unstressed vowel loss [Martinet 1952:192-5]. The effects from Frankish and Gaulish may have cumulated, engendering in Old French a language with an extremely heavy stress accent (first from Gaulish but greatly reinforced by Frankish) and weak segment boundaries, which together

\textsuperscript{11} If -as derived endings in French come from Gaulish, it makes sense that similar -as feminine plural endings in Spanish and other Western Romance languages have the same Celtic origin. Unlike Latin, Celtic languages had nominative plurals ending in -s, and the Western Romane languages, spoken in formerly Celtic areas, now have it while other Romance languages lack the feature. However, although I've never encountered it, a reasonable internalist counterpoint would argue for morphological analogy from Latin plural accusative endings in -s.
generated interaction effects of lengthening and the diphthongizing stressed vowels in open syllables, and then slurring and deleting unstressed vowels [Cerquiglini 2007:26; Meillet, 61-2].

Other Germanic effects include the reintroduction of /h/ as a phoneme [Pope 1934: 15], while miscellaneous Celtic effects are more widespread. Of the Gaulish substrate effects in French initially posited by Ascoli, some are now widely supported, including the merger of /p/ and /k/ when before /s/ or /t/ [Posner 1997: 242-3], and then the spirantization of /k/ to /x/ in the same context [Pope 1934: 6; Mees 2003: 13], the gradual fronting of /u/ to /y/ [Pope 1934: 6; Pescarini 2016: 190; Posner 1997: 250], while another of Ascoli’s claims, the second palatalization (/k g/ > /c ɟ/ > /tʃ dʒ/), is no longer typically included. Also discussed as either Celtic or Germanic in origin regressive vowel nasalization, but there is insufficient evidence [Pope 1934: 6; Posner 1996: 24-29]. All of these except nasalization, as well as “Spezia-Rimini/ Celto-Romance” intervocalic lenition, occurred only in languages that replaced Celtic and also occurred in surviving Celtic languages [Pope 1934: 6; Posner 1996: 237; Martinet 1952: 214; Savignac 2004: 26; Pescarini 2016: 190; Mees 2003: 16]. For the first two shifts, as well as lenition, consistent with Thomason’s [2010] criterion for language contact explanations, have been attested [Pope 1934: 6] to have occurred first [Herman 2000: 46-47; Martinet 1952: 214-5] in Gaulish inscriptions and then, without known prior kindred tendencies, appeared in regional Latin inscriptions [Adams 2007: 286-7; Lambert 1995: 46-7].

5. Computational Models of Diachronic Change: Reconstruction and Simulation

If sound change occurs algorithmically, as the Neogrammarians proposed, then naturally one should be able to represent them with algorithms. The use of computational models to the chronological development of language dates back at least to 1964 with the work of Morris Swadesh [Dunn in Bowern
The modeling of historical word forms, although not properly their historical phonologies (i.e. forms were not represented *phonologically* as they are in this paper) came soon after. In 1969 the first [Piwowarczyk 2016] computational simulation of sound changes was made to historically derive Russian words from Proto-Indo-European [Smith 1969]; in 1976 the second such model was made, deriving Old French from Latin [Burton-Hunter 1976]. After a bit of a lull, more of these “historical derivation” computational simulation models were made, including Spanish from Latin [Hartman 2003] and Polish from Proto-Slavic [Kondrak 2002]. These models, however, modeled words in simple script, without any\(^\text{12}\) representation of the features of the phonetic units of which they consisted, [Burton-Hunter 1976; Piwowarczyk 2016: 1]; in contrast, implementing a computational simulation that takes into account proper phonological representation is a central aim of this paper. Some degree of similar phonological (and also morphological) modeling is, as far as I know, currently being undertaken by D. R. Piwowarczyk of Jagellonian University Krakow for Indo-European as a whole [Piwowarczyk 2016].

The beginning of proper *computational phonology* is said to have come in 1981, and it came in a model of phonological rules, a deeply important concept for Historical Phonology, with a finite-state transducer [Jurafsky & Martin 2009: 395; see Kaplan and Kay 1981]. Computational “phylogenetic models” have become increasingly important in Historical Linguistics as a whole, and since the turn of the century, there has been an “explosion” in their use, starting from Indo-European languages but also seeing important breakthroughs using various Australian and Austronesian languages [Dunn in Bowern and Evans : 7], with works such as that by Nakhleh, Ringe and Warnow [2005] being particularly influential in developing models of *evolutionary* diachronic change.

---

\(^\text{12}\) Despite his own notable work in automated representation of phonological features for diachronic purposes, Kondrak’s Polish simulation admittedly “has no notion of phonological features” [Kondrak 2002: 141]
There is rapid ongoing expansion of computational methods and there is much that could be discussed. One particularly interesting and relevant project was Kondrak’s ALINE system, which, similarly to this project, encodes phonological features, doing so the purpose of aligning phonological segments that are likely to come from the same source in related languages, for purposes of auxiliary usage in diachronic reconstruction. Also quite notable is Bouchard-Côté et al’s [2013] automated Austronesian reconstruction program, which uses a Monte Carlo inference algorithm to attain an accuracy of 85%, allowing one phone of deviation from the manual reconstructions currently held by linguists of Austronesian, described as “a first step toward a comprehensive computational model of sound change” [Atkinson 2013].

6. Description of this diachronic language simulation package

This package is built, with some deviation, around the phonological theory of distinguishing factors, described at length in section 6c. Accordingly, the system uses phonetic units (section 6a), which are organized in a class hierarchy that both represents the theoretical division of types of phones and serves pragmatic purposes as well (section 6b). These Phone class objects are represented and defined in such a way that they are considered equal if and only if they have the same values for their phonological feature parameters (see section 6c). Lexemes are modeled as having phonological representations made up of these Phone units (section 6d), although the possibility is left open for incorporation of non-phonological information (for more discussion, see section 8a). These representations are used by the Alteration class (section 6e), which is an object representing a diachronic change with a method to implement the change as a function, with the class parameters being the shift’s targets, destinations and the prior and posterior context. For various purposes of convenience, also included are parsers that can convert text either in IPA or Latin to appropriate phonological structures (section 6f). Lastly, there is a Lexicon class (section 6g) which was originally intended to represent all relevant lexemes in the language, but which is currently actually used mostly
as a convenient container for a much smaller set of words used for demonstrative purposes in tester classes (with the intention to use it for its original purpose still intact).

6a. Representational of phonetic/phonological units

As is the majority position, though not universally accepted, I treat the phone as the basic phonological unit, rather than the lexeme (i.e. the word stripped of any morphological additions). I acknowledge that a minority of Historical Phonologists do instead treat the lexeme as the basic unit, and that this minority is particularly concentrated among the lexical diffusionist camp for obvious reasons. Given this, for the sake of neutrality, I have also programmed my phonological representations in such a way that the word is still treated as a discrete unit. It is worth noting that some Phonologists would disagree with this as well, arguing that the only relevant unit is the phone (or phoneme) and that the lexeme is in some cases an artificial boundary between segments; for once again obvious reasons, these linguists tend to be of the regularist/Neogrammarian persuasion. Coding for both the phone and the lexeme, with the lexeme constructed as a unit that is built in part from the sequencing of smaller phone units, seems like the best compromise between these two extreme positions, as a way to maintain the neutrality of the model in the context of the regularist vs. diffusionist controversy over sound change.

Another controversy, in both Phonology proper and Historical Phonology, is the relationship between phones and phonemes, although it is not covered at length in this paper. In short, while for a time the standard position across the field of Phonology had been that there is a distinction between phones, which, being “phonetic” are the units of sound as they are actually produced by speakers of a language, and phonemes, which, being “phonemic” are the supposed underlying phonological units that are realized as phones. Although the distinction may be hard to grasp at first, its relevance is apparent as the same phoneme can be pronounced as different phones by individual speakers. The difference becomes even more apparent in situations where allophones are present, where the same phoneme is
said to be pronounced with different phones—for example, the “t” is realized phonetically as an aspirated alveolar stop in “tip” (phonemes: /tɪp/; phones: [tʰɪp]), a (non-aspirated) alveolar stop in “stay” (phonemes: /steɪ/ a glottal stop in “mitten” (phonemes: /mɪttn/, phones: [mɪʔtən]), a preglottalized alveolar stop in “mit” (phonemes: /mɪt/; phones: [mɪt]) and an alveolar tap in “pitting” (phonemes: /pɪtɪŋ/; phones: [pʰɪɾɪŋ]). Some linguists more recently have objected to the view that such underlying representations actually exist subconsciously in the minds of language users. I also omit the difference between phonemes and phones from the model—not because I necessarily endorse this dissident view, but because in the diachronic perspective it would seem simultaneously trivial, overly particular, and hubristic to try to assign discern differences between surface phones and underlying phonemes at a stage in a language that existed many centuries ago. Therefore, for the purposes of the Latin to French model, allophonic relations between phones and phonemes are only relevant for the input source language (Latin) and the output language (French), and essentially irrelevant for all the stages in between, in which the phone units within a lexeme are vaguely assumed to be essentially phonemes, not in the strict sense of the word, but for the purposes of a diachronic perspective.

6b. A class hierarchy of types of phones.

For purposes of coding in Java, an object-oriented programming language, objects that represent phone units may be of one class in a nested class hierarchy. The overarching class represents not the phone, but an abstract class Phonic (abstract in both the computer science and English senses), which exists for pragmatic programming purposes rather than theoretical ones, as it can be generalizable to both of its subclasses, PseudoPhone and Phone. Phone is another abstract class (again, both senses)
which represents a real phonetic unit, while *PseudoPhone* is used to represent various non-phonetic concepts that typically represent boundaries. They are considered “phonic” only for pragmatic programming purposes, where sound shifts embodied in the *Alteration* class may treat them as “phonic” context variables (see section 6e). At the moment, the interface is prepared to use PseudoPhones to represent word onsets, word codas, and syllable boundaries, although this potential is only implemented for the former two. The abstract *Phone* class is itself extendable to *Consonant* and *Vowel* classes, which unsurprisingly represent their namesakes, while the *Vowel* class itself has a *Phthong* subclass which is used to represent diphthongs and triphthongs, and could theoretically be used to represent a “phthong” with up to five component elements.

6c. Representation of phonological features of phone units

One of the hallmarks of this model is the representation of phones not as simple character strings (as was done in many earlier models) but rather, as units defined and differentiated by their *phonological features*. When discussing one language, the *phonological features* generally refer to the features that distinguish each phoneme from the others (or, less commonly, distinguish their phonetic realizations from other phonetic realizations). If the distinctiveness between the corresponding feature sets of two or more phonemes (or phones) is blurred over time, they ultimately may tend to merge. Within the discussion of multiple languages, however, *phonological features* are used to distinguish all the possible articulations that are produced in any one or more languages. It is this sense that I adopt when creating an interface for representing phone units as entities whose identity is defined by the values of parametrized *phonological features*.

Some phonological features are represented in a *boolean* fashion, taking on a value of true or false. These include:
Voicing – a phone (almost always a consonant) is voiced if and only if it is pronounced with discernible vibration of the vocal cords or not. In English, this makes the distinction between $p$ and $b$, $f$ and $v$, and so on.

Rounding – a phone is rounded if and only if it is pronounced with rounded lips. Rounded vowels in languages with Latin script typically include $u$ and $o$ variants. An example of a rounded consonant would be the approximant English /w/ phone, which is the most common rounded consonant. However, occasional trends of labiovelarization and labiopalatalization result in many other sorts of rounded (read: rounded and labialized) consonants. Latin $qu$ /kʷ/ was a rounded velar stop.

Aspiration – a phone (almost always a consonant) is aspirated if and only if it is pronounced with a short burst of breath. In English, initial stops are aspirated- hence why the $p$ in pit has a slightly different phonetic realization than the $p$ in spit.

Nasality – a consonant or vowel is nasal if and only if it is pronounced with a lowered velum so that air also escapes from the nose. English $m$, $n$ and $ng$ are nasal consonant phonemes. French has phonemic nasal vowels.

Rhoticity – a rather vague concept which refers to if a consonant or vowel is perceived by speakers to be “R-like”. Most languages world wide have some “rhotic” consonant; only a few (including English) have rhotic vowels. Typically rhotic consonants are alveolar in place and/or taps or trills in articulation manner, but the typical French $r$ is an uvular fricative (a sound that is typically not rhotic in many languages). Meanwhile, the rhotic consonant of English is an approximant. English has an alveolar tap allophone for intervocalic /t/ or /d/ that is not rhotic despite typically being treated as rhotic in other languages (Spanish, Albanian, etc.).
- **Laterality** – if the phone is “L-like”, meaning air escapes off the sides of the tongue. Most languages have at least one lateral phoneme. Many have two (ex. modern Albanian) while some have three, such as Old French, which had an alveolar /l/, a velarized dental /h/, and a palatal /ʎ/, the mouillé.

Other phonological feature parameters can take a limited number of ordered categorical values represented as integers. These include **length**, which may be used to represent the length of one, two or three phonetic units. In practice, it is used typically for vowels, while **geminate** consonants are just represented as sequences of identical consonants. The other phonological feature parameter that is represented this way is stress, which may take values representing statuses of **unstressed**, **stressed**, or **secondarily stressed** (“atonic”, “tonic”, and “countertonic” respectively, in the terminology used for some older French linguistics like that done by Mildred K. Pope).

Last are the three parameters that are represented quantitatively as if placed on a spectrum, which are defined in our paradigm as **articulation place**, **articulation manner**, and **coarticulation**. **Articulation place** refers to the place in the mouth (or throat) where friction is produced to produce the phone, and can be **glottal/laryngeal** (a value of 10), **pharyngeal** (20), **uvular** (30), **velar** (40), **palatal** (50), **postalveolar** (60), **alveolar** (70), **dental** (80), **labiodental** (90), or **bilabial** (100). In English, almost all consonants are either bilabial (b,p,m), labiodental (v,f), alveolar (d,t,n,z,s,r,l), postalveolar (sh,ch,j,zh) or velar (g,k,ng), although there is also one palatal phoneme (y) plus one palatal allophone (h before a y sound, as in huge), as well as one glottal phoneme (h) and one glottal allophone (the t-s in mitten). **Coarticulation** refers to a second element of articulation with a different place than the main one, and uses the same integer paradigm; most common among these are **palatal**, **velar** and **glottal** coarticulations. The articulation manner, meanwhile, refers to how the sound is produced, including possible categorizations such as **stop/plosive** (10), **sibilant affricate** (13), **spirant affricate** (15), **sibilant fricative** (23), **spirant fricative** (25), **trill** (27), **tap** (28), or **approximant** (30) for consonants, and various
values from 40 to 70 denoting vowel height for vowels. It is intentional that the difference between vowels and consonant articulation manners is defined on a spectrum, as the high vowels (40-50) are more likely to be consonantalized, typically into approximants, while approximants and other more vocalic consonant manners, which are assigned higher integer values, are the most likely ones to be vocalized, typically into high vowels. To help familiarize unfamiliar readers with the relations between different articulation manners and places, below is a list of major consonants that have historically existed in English and French, including only those that are distinguished solely by their place and manner (voiced consonants are followed by their voiceless counterparts) for the sake of simplicity (all the others, such as laterals and nasals, are omitted), in their International Phonetic Alphabet symbols.

Blue denotes those that occur or have occurred in French only, while red denotes the same for English.

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labio-dental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>b</td>
<td>p</td>
<td>d</td>
<td>t</td>
<td>d</td>
<td>t</td>
<td>j</td>
<td>c</td>
<td>g</td>
<td>k</td>
</tr>
<tr>
<td>sib. affr</td>
<td>dz</td>
<td>ts</td>
<td>dʒ</td>
<td>tʃ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp. affr</td>
<td></td>
<td></td>
<td>z</td>
<td>s</td>
<td>ʒ</td>
<td>ʃ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sib. fric</td>
<td>β</td>
<td>φ</td>
<td>v</td>
<td>f</td>
<td>δ</td>
<td>θ</td>
<td>j</td>
<td>ç</td>
<td>γ</td>
<td>x</td>
</tr>
<tr>
<td>spi. fric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td>ç</td>
<td>γ</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tap</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>approx.</td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6d. Representation of the lexeme

Lexemes are represented as entities that are defined by the values of parameters in the same way that phones are. In the case of lexemes, the most salient parameter is the phonological representation, which consists of a sequence of phones that, as an ordered collective unit, serves as the symbol for whatever semantic referent the lexeme in question has.

6e. Alteration – object representation of diachronic sound change
An essential, perhaps the most essential, part of the package constructed for this paper is that diachronic sound changes themselves are programmed as unit entities that, just like phones and lexemes, are given identity by the values that certain feature parameters are set to. Since one of the ultimate goals of this project is to create a system that can identify, construct and assess the likelihoods of putative historical diachronic sound changes, it is essential that they be treated as such an object that could be constructed by an automated algorithm.

Here, the diachronic sound change is represented as an Alteration class object, with four parameters. The Alteration class is built according to the paradigm that a sound change has a set of target phones, each of which is mapped to a certain destination phone, being mutated in a specifiable context, occurring at a certain time (or stage) in a language’s history. While time is not coded into the Alteration class, instead being represented in where an Alteration object is declared in the simulation class itself, all the other parameters listed above are.

The set of target phones for a given diachronic change is coded as the targets class parameter, represented by a List<List<Phone>> object (a list of lists of phones). The set of destination phones is similarly coded as the destinations class parameter, a List<List<Phone>> of the same size as the targets parameter, with corresponding indices for pairs of target and destination.

The representation of context is more complicated. Most importantly, it is programmed the way it is for the sake of efficiency as well as adherence to the phonological theory of distinguishing features. Accordingly, context requirements are not represented as lists of possible phones for each contextual relative place, but rather as sets of restrictions on the features of phone that decide whether it is legal to be in a certain place, contextually relative to the target. Place is specified as the number of phones before or after the target. If the target is a segment of multiple phones, then this means the number of phonic-places after the last phone, or before the first phone. Also of note is that while the target and
destination can each only consist of Phone class objects, the context is specified for Phonic objects, allowing the starts and ends of words to be used as relevant context in addition to proper Phones.

The restrictions are expressed as a positive or negative value for any phonological distinguishing feature. Thus, “+nasal” would imply that a phone at the specified position must be nasal -- either a nasal vowel or a nasal consonant, since the phone type is not specified. Negative and positive constraints are not, in fact, symmetrical in behavior. If a negative constraint is violated, the wider context is always immediately considered illegal, and the diachronic change will not occur. On the other hand, a positive restraint represents a condition that could be true to determine the phone at the given contextual position to be legal. The only situation where it is the singularly deciding specification occurs when the positive restraint is the only one on that context position that pertains to its parameter (place, manner, coarticulation place, length, stress, etc.). So, in practice, if constraints on a certain position are “+alveolar,+nasal”, since these positive constraints refer to different parameters (place and nasality respectively) both must be true for the position (and thus the wider context as a whole) to be legal. On the other hand, if the constraints on that position are “+alveolar, +glottal”, one of the two must be true (but they don't both need to simultaneously be true) to determine that the phone is legal for the given position, since both pertain to the same parameter (in this case, articulation place). To represent the overall context constraints for shifts, the set of each positions constraints is placed into an overarching set of the sets of context constraints indexed for each relative position.

6f. Parsers for extracting phonological structure from String objects

The restrictions are in fact represented with their own class, Restriction, and when we speak of the set of restrictions on a certain phone position in the context, this is actually represented with its own class, CandRestrictPhone, which represents the set of restrictions on a certain phone. There is also a method, parseCandRestricts(), that exists in various classes, made for convenience which parse properly coded String information object into list of CandRestrictPhone objects. When calling the method, different restrictions are separated with commas, while the delimiter between different CandRestrictPhone objects is the semicolon.
For the sake of convenience, there are two classes of text parsers that allow the user to input String objects to construct corresponding objects of the phonological representation structures described above. The first, `PhoneStructureTranslator`, “translates” International Phonetic Alphabet (IPA) text into its corresponding phonological form. It can do so to return either a single `Phone` object (using the `getPhone()` method), a specified `Vowel`, `Consonant` or `Phthong` type `Phone` object (using `getConsonant()`, `getVowel()` or `getPhthong()`). It can also parse more than one phone at a time. The method `parseSegment()` can parse a succession of Phones, as long as they are separated by commas ("","), returning a List<Phone>. The class can also parse a list of segments, separated by semicolons (";"), using the method `parseSegList()`, which returns a List<List<Phone>> object. In practice, this latter method is used primarily for constructing target and destination input parameters for the `Alteration` class, which currently requires that both of these two parameters be entered in as List<List<Phone>> objects.

There is also a parser for Latin words, the `LatinParser` class, given that with a few exceptions the phonologies of Latin words in Classical Latin are completely predictable given their orthographies. In practice, this class is mainly used for constructing the Latin words to fill the lexicon of words that are to be given as inputs to the simulation algorithm, and, in some isolated cases, constructing target and destination inputs for the `Alteration` class. Given this, `getSegment()` and `getSegList()` are implemented for `LatinParser` just as they are for `PhoneStructureTranslator`, but `getPhone()` method is not yet supported. Delimiters are also different for parsing Latin—no comma is needed to separate phones when calling `getSegment()`, while a comma rather than a semicolon is used to separate segments when calling `getSegList()` for `LatinParser`. At the moment, because of character encoding issues and differences between how vowels with macrons are represented in different fonts, as well as more critical macron encoding inconsistencies between Windows and Macintosh, it is recommended that instead of using macrons, users wishing to enter long vowels do so by entering a colon (":"), after the long vowel.

6g. The **Lexicon object**
The program also has a *Lexicon* class, which was originally created to represent the set of all words in a language at a point in time. This set of words is internally represented as a HashMap object, with the keys being the English translations of the words’ meanings, paired to the words’ phonologies represented as lexeme *WordPhon* objects (i.e. lexemes for which only phonological information is stored). The class also contains a method which implements a sound change (given as input as an *Alteration* object) on all words in the lexicon.

At the moment, in practice the *Lexicon* class is primarily used as an auxiliary class for simulation tester classes, which allow one or more words, all encapsulated in a *Lexicon* object, to have their development via sound shifts simultaneously represented (and usually printed). Typically, only 1-10 words are used, a small fraction of the actual “lexicon” of the language chosen for demonstrative purposes. However, it is intended that the same class will ultimately be used to represent a language’s whole “lexicon” in the future, for purposes of statistical analysis (see section 8c) and for calculating patterns of distribution for specific phones.

7. An automated simulation of the development of Latin into French

The central piece of this project has been the construction of a simulation of the phonological development of Latin into Modern French, written in Java, using the phonological-lexical apparatus described above. The simulation models different stages of the development of French, starting with the differentiation between the Classical Latin, the “Vulgar” Latin that developed among the common people, the development of the Gallian regional variant of Vulgar Latin, then the development of Gallo-Romance in the Dark Ages, followed by Early Old French, Classical Old French, Later Old French, Middle French, Early Modern French, and Modern French. It is now available on GitHub, with the URL granted upon request. Also included is the *MerciJacques* demonstration class, which automatically shows the
development of specific words in French over time (depicted below), as well as the \textit{LTFTester} class, which allows the user to enter whatever word or words they want simulated.

\begin{center}
\begin{tabular}{lllllllll}
English meanings: & bed & tongue & four & ice & cow & dog & table & sea \\
Latin written forms: & lectum & lingua & quattuor & glacia & bovem & canem & tabula & mare \\
Latin phonemic forms: & \textit{<le\textquotesingle k\textquotesingle to\textquotesingle m>} & \textit{<cl\textquotesingle pg\textquotesingle s>} & \textit{<k\textquotesingle\=a\textquotesingle \textquotesingle t\textquotesingle pu\textquotesingle r>} & \textit{<gl\textquotesingle k\textquotesingle\~a>} & \textit{<b\textquotesingle\~w\textquotesingle m>} & \textit{<k\textquotesingle g\textquotesingle m>} & \textit{<\textquotesingle s\textquotesingle\~b\textquotesingle l\textquotesingle a>} & \textit{<m\textquotesingle\~re>} \\
Vulgar Latin & \textit{<le\textquotesingle k\textquotesingle to\textquotesingle m>} & \textit{<cl\textquotesingle pg\textquotesingle s>} & \textit{<k\textquotesingle\=a\textquotesingle \textquotesingle t\textquotesingle pu\textquotesingle r>} & \textit{<gl\textquotesingle k\textquotesingle\~a>} & \textit{<b\textquotesingle\~w\textquotesingle m>} & \textit{<k\textquotesingle g\textquotesingle m>} & \textit{<\textquotesingle s\textquotesingle\~b\textquotesingle l\textquotesingle a>} & \textit{<m\textquotesingle\~re>} \\
Later Vulgar Latin, with Gallian accentation: & \textit{<le\textquotesingle\~x\textquotesingle to\textquotesingle m>} & \textit{<cl\textquotesingle\~x\textquotesingle s>} & \textit{<k\textquotesingle\~a\textquotesingle \textquotesingle t\textquotesingle pu\textquotesingle r>} & \textit{<gl\textquotesingle\=c\textquotesingle\~j\textquotesingle o>} & \textit{<b\textquotesingle\~i\textquotesingle c\textquotesingle>} & \textit{<k\textquotesingle\~i\textquotesingle c\textquotesingle>} & \textit{<\=s\textquotesingle\~b\textquotesingle l\textquotesingle a>} & \textit{<m\textquotesingle\~i\textquotesingle c\textquotesingle>} \\
Gallo-Romance, circa 600: & \textit{<le\textquotesingle j\textquotesingle o>} & \textit{<le\textquotesingle\~n\textquotesingle o>} & \textit{<k\textquotesingle\~a\textquotesingle \textquotesingle t\textquotesingle\~j\textquotesingle o\textquotesingle>} & \textit{<b\textquotesingle\~w\textquotesingle\~e\textquotesingle>} & \textit{<k\textquotesingle\~e\textquotesingle\~g\textquotesingle>} & \textit{<t\textquotesingle e\textquotesingle l\textquotesingle\~b\textquotesingle a\textquotesingle>} & \textit{<\textquotesingle m\textquotesingle\~i\textquotesingle c\textquotesingle>} \\
Early Old French, circa 800: & \textit{<l\textquotesingle\~i\textquotesingle t\textquotesingle>} & \textit{<l\textquotesingle\~n\textquotesingle g\textquotesingle s\textquotesingle>} & \textit{<k\textquotesingle\~a\textquotesingle \textquotesingle t\textquotesingle\~r\textquotesingle\~o\textquotesingle>} & \textit{<gl\textquotesingle\~a\textquotesingle\~j\textquotesingle e\textquotesingle>} & \textit{<b\textquotesingle\~u\textquotesingle\~e\textquotesingle>} & \textit{<t\textquotesingle\~f\textquotesingle\~y\textquotesingle\~g\textquotesingle>} & \textit{<\=t\textquotesingle a\textquotesingle\~b\textquotesingle l\textquotesingle a\textquotesingle>} & \textit{<\textquotesingle m\textquotesingle\~i\textquotesingle r\textquotesingle>} \\
Classical Old French, circa 1100: & \textit{<l\textquotesingle\~i\textquotesingle t\textquotesingle>} & \textit{<l\textquotesingle\~n\textquotesingle g\textquotesingle s\textquotesingle>} & \textit{<k\textquotesingle\~a\textquotesingle \textquotesingle t\textquotesingle\~r\textquotesingle\~o\textquotesingle>} & \textit{<gl\textquotesingle\~a\textquotesingle\~j\textquotesingle e\textquotesingle>} & \textit{<b\textquotesingle\~u\textquotesingle\~e\textquotesingle>} & \textit{<t\textquotesingle\~f\textquotesingle\~y\textquotesingle\~g\textquotesingle>} & \textit{<\=t\textquotesingle a\textquotesingle\~b\textquotesingle l\textquotesingle a\textquotesingle>} & \textit{<\textquotesingle m\textquotesingle\~i\textquotesingle r\textquotesingle>} \\
Late Old French, circa 1300: & \textit{<l\textquotesingle\~i\textquotesingle t\textquotesingle>} & \textit{<l\textquotesingle\~n\textquotesingle g\textquotesingle s\textquotesingle>} & \textit{<k\textquotesingle\~a\textquotesingle \textquotesingle t\textquotesingle\~r\textquotesingle\~o\textquotesingle>} & \textit{<gl\textquotesingle\~a\textquotesingle\~s\textquotesingle>} & \textit{<b\textquotesingle\~e\textquotesingle\~f\textquotesingle>} & \textit{<t\textquotesingle\~f\textquotesingle\~y\textquotesingle\~g\textquotesingle>} & \textit{<\=t\textquotesingle a\textquotesingle\~b\textquotesingle l\textquotesingle a\textquotesingle>} & \textit{<\textquotesingle m\textquotesingle\~i\textquotesingle r\textquotesingle>} \\
\end{tabular}
\end{center}

Using the \textit{LTFTester} class, one can input a Latin word, or a potentially large set of Latin words to fill the lexicon, and it should implement the simulation, printing out the phonological forms the words are projected to have had during each period of the linguistic history of French. While it is to be admitted the division of a language’s history into discrete periods can be quite subjective, the periods used for this project are roughly based on the division employed by Pope [1934], although some changes assigned by her to the largely unattested “Early Old French” period are instead assigned to the largely unattested “Gallo-Romance”, and vice versa mainly for programming purposes. There were some revisions for the dates of certain shifts based on more modern research, for example relying on Buckley’s more recent [2003] survey of Gallo-Roman and Old French a-fronting. There is also some dependence on later authors like Posner [1997] and Arteaga [2012].

For Latin phonology and its relation to the orthography, the program mainly relies on the work of Allen [1978] and Oniga [2007]. For later developments and Vulgar Latin, Palmer [1988] and Herman [2000] are consulted, and for Gaul-specific regional Latin developments, reference is given to Adams [2007], Pope [1934] and Grandgent & Moll [1999]. From Gallo-Romance and Early Old French onward,
the program is mostly reliant on the chronology given by Pope[1934]'s impressive work, with additional more modern input from Arteaga [2012] and specific input from Buckley [2003], Operstein [2010] and Mazzola [2012]. For various specific shifts, source attribution is given in the comments above the shift in the LatinToFrench java file.

7a. Current accuracy of the simulation system: successes and remaining issues

At the current moment, the performance simulation specific to the derivation of inherited French words from Latin could be called a pilot program and its accuracy is inconsistent. The current aim is to do a statistical analysis on the correctness of the outputs of the simulation for a large and representative set of inherited French words as derived from their respective Latin forms. At the moment, such an analysis has not been completed. However, various lower level and more casual testing of the program has found that for various sets of words, the program typically has roughly a 40-60% rate of correctness. The reasons for this are varied, and described at length below. As also noted below, there is some variation in different domains: for example, for ordinal numbers, the rate of errors is only 11.5%, rather than around 50%.

There are some cases undoubtedly some cases, yet to be discovered, where the program fails to implement historical changes that are in fact present in the literature on French historical phonology and were merely missed in the design of the simulation. These are expected to be much more numerous and significant for the Middle French, Early Modern French, and Modern French periods although preliminary examination of the simulation’s performance found a few in Early Old French, of which some were subsequently fixed. In some cases, caveats had previously been missed. For example, consider the treatment of /ɛ/ in closed palatals. When stressed and in open syllables, it had already become the diphthong /iɛ/ in Gallo-Romance, but elsewhere when /ɛ/ combined with the /j/ that ejected backwards from palatals in Early Old French, in most syllables, the two merged to /ei/. However,
it was discovered that this in fact only applied to countertonic (i.e. secondarily stressed) syllables—the few cases of fully stressed /ɛj/ clusters in closed syllables in fact passed to /ei/ and then ultimately /iɛi/ and then /i:/ (i.e. as happened in dix, from Latin decem).

Some apparent errors may arise from issues in the ordering of shifts. For example, the system currently derives /lɛʁɛ/ *lière from lege:re, not the lexical phonology of the correct French word, lire /liʁ/. This appears to have occurred because the simulation, has, in the Gallo-Romance stage, the word being modified by a sound shift that acted before /r/ to change palatalized intervocalic g, /ɟ/, to d, causing the word to pass to /lieðra/ in Early Old French. This was the same shift that saw Latin plangere become (correctly) plaindre. Ultimately, erroneously, the descendant of lege:re is thus presented by the simulation as showing the same pattern as pierre (pronounced historically as /pieðra/ in Early Old French, from Latin petra). The dentalization of /ɟ/ before r indeed affected words where former front vowels had been lost through apocope, as was the case with lege:re. However, it is notable in lege:re’s case that it was originally the second vowel, not the first, that was stressed, and should have originally not been lost to apocope – the stress may have shifted later, allowing for apocope of the second vowel only after the dentalization shift was no longer productive, and at which point the /ɟ/ should have passed instead to /j/, which ultimately would regularly render lire (via /liejr/).

In some cases, in the later stages of the programming, shifts were added in to the simulation file LatinToFrench to fix apparent errors. One example is the case found with closed tonic /ɛj/ clusters as in decem > dix. In general, there is more confidence for issues of precision of the contexts of shifts and specific targets vis-à-vis stress and position for the earlier stages, but a significant number may remain with respect to the Middle French, Early Modern French, and Modern French stages which at the moment still await more thorough investigation. These eras may prove more difficult because although there were fewer shifts in these areas, many had very complicated conditioning factors and nonphonological sources of instability (such as the onset of prescriptionist schooling for speech among
French speaking children during the relevant period). In particular, there is still uncertainty in the program’s chronology about developments concerning the distribution of mid-close /e/, /o/ and /ø/ as opposed mid-open /ɛ/, /ɔ/ and /œ/ respectively during the Middle French and Modern French stages. There may also remain errors with the ordering of specific shifts in earlier stages, not least because the ordering of shifts in these less attested eras such as Gallo-Roman and Early Old French is still in fact disputed among historical phonologists [see discussion by Buckley 2003 on a-fronting, for example].

In some cases, it seems that apparent inconsistencies in the performance of the simulation are not its own fault, but in fact reflect inconsistencies between social classes in Latin in antiquity: for example, what was *qui:nque* /kʰi:nkʷe/ among the upper classes and in the Latin literary register was in fact *ci:nque* /ki:nkʷe/ in the speech register that gave rise to French, due to a sporadic dissimilation effect [Pope 1934: 318 section 823iii]. As such, it should be entered into the program as “*ci:nque*”, not “*qui:nque*”—for which, the system outputs the correct output of /sɛk/.

As the simulation architecture was built in part as a test of Neogrammarian beliefs about sound change, it was built using Neogrammarian assumptions. Therefore, if there are in fact cases of lexical diffusion causing lexical splits and irregularity, the machine is expected to fail for the words they affected. At present, no confirmed cases of this have been found where the observed effects of irregularity have not yet been confirmed to not be any of the following: (a) errors in the code or omission of shifts that are supported in the historical phonological literature, (b) apparent regular shifts that do not seem to have known exceptions but have not (yet) been found to be supported in the literature, or (c) non-phonological interference.

In cases where non-phonological interference is known to have acted upon words, the simulation is not predicted, according to Neogrammarianism, to be correct—in fact its output is predicted to be incorrect. There are numerous cases of this, and the Middle French era is particularly affected because
of the combination of the fad of using dialectal loans and the heavy influence of prescriptionism, which caused cases of hypercorrection and inconsistent restoration to spelling pronunciations. Some shifts, like the shift of /ɛ/ to /a/ before /r/ came to be regarded as vulgar, as did r-buzzing, but the hypercorrective attempts to reverse the shift produced inconsistencies -- such as both e-lowering before r and intervocalic buzzing being attenuated except for in certain words like par and chaise, while r was hypercorrectively inserted in some places where it had not previously existed (ex Nemours). The simulation was not built to handle these effects, precisely because they are non-Neogrammian, and that is exactly what is expected. The simulation also asserts that Latin septem should render French sit\textsuperscript{14} /si/. Although modern French has sept /sɛt/, this is not in fact a fault of the program as septem would indeed regularly render sit /si/. Clusters of pt rarely ever occur in native French words, because p in that position was regularly merged with k, then opened to /x/ and finally /j/ in such positions (as indeed happened to the /k/ in sex /sɛks/ > /sɛks/ > /sis/ six). This raises the suspicion that sept, despite being a cardinal number, could be influenced by either the Latin literary register or other Romance languages. Italian, which was spoken by much of France’s elite at times, and where /pt/ clusters regularly became geminate /tt/, is a likely culprit. One could also postulate that such influences on sept may have been encouraged by the fact that were it not for their effects, septem would have ended up sounding very similar to six, as well as the contrapositive affirmation si (‘yes’ said in rebuke of a negative statement).

In addition to the non-Neogrammian effects of sociolinguistic interference, syntactical effects have been notable especially in articles; this impossible to predict for a simulation that uses only phonological factors, and for this reason, the simulation has not yet been tested on simulating the

\textsuperscript{14} I am asserting that the program “thinks” it would be spelt sit because this is closest to its projected lexical phonology in the Classical Old French and Late Old French stages, which are what the French written forms tend to reflect, with a few exceptions.
development of articles. Analogy has also had a historically large role that can't be predicted by the simulation as it is currently. For example, this includes a number of words where final /l/ was vocalized to /u/ just as it was before consonants, due to analogy because of the influence of flexional -s forms [Pope 1934: 156 section 393]. Thus the system, incorrectly but “Neogrammari anly”, outputs /bɛl/ bel rather than /bo/ beau for the descendant of Latin bellum -- just like, regularly, caballum renders /ʃəvəl/ cheval, a word that was not affected by analogy in this way.

In a test on the domain of ordinal numbers, the system did quite well, while it is also notable where the errors did occur in French there are 26 numbers that are not recent loans or neologisms or combinations (such as vingt-et-un, dix-sept, etc.). Of these 26, the breakdown on the performance of the program is as follows:

1. The largest group consists of words that are expected to adhere to Neogrammari an rules and indeed do so in the simulation, being outputted correctly. These comprise 19 of the total 26 (73.1%), and include : (1) un /œ̃/, (4) quatre /kaʁwa/, (5) cinq /sɛk/, (6) six /sis/, (8) huit /ɥi/, (10) dix /ɡis/, (11) onze /ɔz/, (12) douze /ɡuz/, (13) treize /tʁɛz/, (14) quatorze /kaʁwaʁ/, (15) quinze /kɛz/, (16) seize /sez/, (20) vingt /vɛt/, (30) trente /tʁɑ̃t/, (40) quarante /kɑʁɑ̃t/, (50) cinquante /sɛkɑ̃t/, (60) soixante /swasɑ̃t/, (90) nonante /nɔnɑ̃t/ and (100) cent /sɔt/.

2. 4 of the 26 (15.4%) are cases where the system fails to output the modern results for words whose respective developments were warped by non-phonological factors. Mille /mij/ (1000) is a
historic loan from Latin [Pope 1934: 318 s823iv] that replaced native regular /mil/ (which the
machine outputs). As noted above, sept (7) and derived septante (70) arose due to a
combination of Latin (or Italian?) influence and dissimilation pressures. Lastly, the -ante ending
of (80) huitante (< Latin octoginta ) is a case of analogy from the other multiples of ten
(quarante, cinquante, etc).

(3) The remainder are words where the simulation was incorrect, 3 out of 26 (11.5%). These include
(2) deux / dø/, (3) trois /tʁwa/ and (9) neuf /nœf/, which the simulation errantly predicted as
rendering /døs/, /tʁɛ/ and /nœf/, coming fairly close to the real forms each time. In the case of
trois, it seems the error lies in the treatment of final -s, which the simulation considers to have
the effect of making the monosyllable closed (blocking the diphthongization of the long e in
Early Old French). In reality, in many words in Late Latin, tre:s included, the word final -s was
omitted and thus didn’t close the syllable, although the exact contextual patterning of this
phenomenon remains unclear. In the case of deux and neuf, the errors lie in the Middle French
and Modern periods, for which the patterning of shifts is less thoroughly ascertained in the
program at the current moment: in particular, the errors here lie in the failure to open /ø/ in
neuf and to delete word final -s in deux.
Type in the Latin words you would like to test:
Mark long vowels with a following colon--ex) a:
Separate different words with a semicolon (;)
Now, please enter the Latin words you’d like to test:
rubrum; galbinum; viridem; purpura; badium; nigrum
Now, please enter their English meanings.
red; yellow; green; purple; bay; black
Vulgar Latin
<ruˈbeʊm> <puˈrpur̩a> <ɡaˈlbiɡœm> <waˈriɡœm> <baˈdʒlœm> <ɡiˈɡr̩œm>
Later Vulgar Latin, with Gallian accentation:
<ruːˈboːo> <puːˈrpoʊɾa> <ɡaːˈlbeɡoe> <baːˈʒreʃa> <ɡeːˈgroʊ>
Gallo-Romance, circa 600:
<ruːˈboʊo> <puːˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjroʊ>
Early Old French, circa 800:
<ruːˈʒaʊ> <puːˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjjaʊ>
Classic Old French, circa 1100:
<ruːˈʒaʊ> <puːˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjjaʊ>
Late Old French, circa 1300:
<ruːˈʒaʊ> <puːˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjjaʊ>
Middle French, circa 1550:
<ruːˈʒaʊ> <puːˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjjaʊ>
Early Modern French, circa 1780:
<ruːˈʒaʊ> <puːˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjjaʊ>
Modern French, circa 1930:
<suˈʒaʊ> <puˈrpoʊɾa> <ɡaːˈlpoʊ> <veːˈɾoe> <boːˈzo> <ɡeːˈjjaʊ>

The domain of colors, pictured above, although initially looking like another domain of relative success for the program, actually provides a cautionary tale against overoptimistic interpretations of the system’s results. Above, we see the development of the six major inherited color terms in French (others were borrowed from Frankish, such as bleu and blanc). At first glance, it appears that the only color that is derived wrong by the system is bai /be/, which was somehow misderived as */bo/, perhaps because the deletion of intervocalic yod occurred either errantly or too early in the Gallo-Romance stage. This would imply a rate of success above 80%. However, in fact, the rate of success here is an unremarkable 50% : although the outcome for noir and vert was correct, the process wasn’t. Vert should have had a final t at least until the Classical Old French stage, and noir should not have had the final schwa that it retained up until the most modern stage. In both of these cases, the error likely lies in the ordering of shifts in the Early Old French stage: the absorption of j into a preceding historic ei (>oi > /wa/) likely happened too late in noir, causing the jr to emit a final schwa, while in vert, the final consonant should have been devoiced.
Lastly, there are a few results of the simulation that have been discovered that may call into question some traditional etymologies that have in the past been questioned. One case is the etymology of *glaive* /glev/ ("sword") as coming from Latin *gladius* (acc: *gladium*). The system does not derive /glev/ from *gladium*, but interestingly, when *gladebum* (a Latinization of the case ending of Gaulish accusative *gladebon*, which has been proposed as an alternative etymology) is entered, the result is /glev/. It is too early to draw conclusions from this or similar cases. However, with further investigation, a larger class of scattered words which are traditionally considered to be inherited from Latin but may instead be from other sources (other Romance languages, dialects, Gaulish, or Frankish, to name a few) may emerge.

8. Where to go from here -- plans for expansion and improvement

Aside from fixing up some of the problems of alteration chronology seen in the previous section, there are many other directions this project may ultimately go. Although this project has already been quite an extensive, for reasons of personal interest as well as belief in the potential of the program to be useful outside of this thesis project, it is currently my plan to expand and extend the project. The current code has been done in a way to make it maximally compatible with as many languages as possible, and to allow multiple future routes of expansion.

8a. Incorporation of non-phonological information into lexemes

With regard to lexemes, there is currently no implementation of any possible parameters that would represent the lexeme’s *referent*, and indeed in this interface there is currently no implementation of semantic matters at all. This is because this model is primarily concerned with Historical Phonology, rather than Historical Semantics. However, this remains an area where the model can be improved and expanded if such an interest arises. After all, it is known that the phonological development of words may be influenced by other words in the same semantic class, as well as by other semantic factors such as gender.
Other areas where the interface’s handling of lexemes may be extended include *morphological* and *syntactical* considerations. So far, most of the words that the system has been tested on have been noun lemmas and the infinitives of verbs, because these are thought to be the least affected by morphological analogy effects or syntactical and/or position effects. Word class (nouns, verbs, adjectives, prepositions, articles, etc) does have an impact on Historical Phonology, as prepositions, articles, particles and the like tend to end up as unstressed monosyllables in languages with historic stress accents due to syntactical reasons, while adjectives may be effected by the historic pressures of maintaining semblance between masculine, feminine, and plural forms. It is theoretically possible to make a class hierarchy of word class types just as was done for types of phones, though it remains outside the scope of this project.

For morphological reasons, meanwhile, syllable stress patterns occasionally differ between other word classes as well. In English, disyllabic words tend to have stress in the first syllable if they are nouns, whereas it tends to fall in the second if they are verbs. Furthermore, morphological classes (grammatical gender and/or declension for nouns, conjugation for verbs, etc.) can have consequence for Historical Phonology due to episodes of *analogy*, such as the emergence of new feminine forms ending in a schwa in French due to the existing pattern of feminine nouns and adjectives ending in a schwa.

However, these grammatical factors vary markedly between languages—for example, English has no gender, French has masculine and feminine, German has a three-way distinction between masculine, feminine and neuter, and Basque has animate and inanimate. At the moment, it is hard to think of how support for them would be built into an interface that is meant to support a wide array of different languages. On the other hand, inclusion of information of this sort may ultimately be necessary to make a wholesome simulation of a language’s diachronic development, as these sorts of effects are by no means an insignificant force. At present, it remains out of the scope of this project, though it will be
interesting to see how in the future such aspects may (or may not) be incorporated into computational simulation models.

8b. Testing etymologies for “unknown etymology words” in French

As mentioned earlier, a significant percent of French words at the moment remain without an accepted etymology. This model can be used to test possible etymologies, given a source and a time it entered the language. To do so, one would just do any necessary sound changes to adjust that would be appropriate for a loanword into French at that time, as is typical when words are loaned and assimilated (to varying degrees) into the native sound system, and then put it through all the periods up until the modern day, or up to wherever is desired.

To do so, one might need to craft fairly simple functions for assimilating loanwords. Of course, in the case of one of the most likely foreign source of words in the ancient period, Gaulish, we already know a fair amount about the periodization of sound shifts going on within Gaulish at relevant periods, and how Gaulish words were assimilated into the local Latin. With a fair amount of research, making a system to assimilate Gaulish words wouldn’t be that hard. It would be even easier to deal with Frankish or other Germanic sources, though to be fair this might be less useful for discovering new etymologies, as Germanic etymologies are likely already discovered. Of course, if desired, the system can already be used as it is to check possible Latin etymologies for modern French words – or to challenge suspicious-seeming Latin etymologies (i.e. *ambulare* for *aller*). As for other sources, there might need to be a bit more work, but given that this system was designed to be very versatile, one does not foresee it being a monumental task.

Essentially, in this way, if one wants to see if a certain word might have a certain etymology from a certain language at a certain time, one can simply input the source word into the system at the given time, and if the result is either identical to or significantly close to the word you are checking for,
this would be supporting, though not conclusive, evidence that the proposed etymology is likely. If such a system could be made robust and able to accommodate various different possible source languages, it would be a great system for helping check etymologies.

8c. Large scale statistical analysis on the French simulation’s results

Although there has been very extensive work over the past two centuries that has given us a fairly comprehensive picture of French historical phonology, there remain some controversies (for example: the nature of /a/-palatalization in Old French [Buckley 2003]), and there probably remain gaps in our knowledge. The simulation system constructed for this thesis could yield insight here. One could enter, for its input lexicon, the entire set of Romance vocabulary that is not known to be loanwords or affected by non-Neogrammian phenomena like analogy. If there are significant numbers of words for which the output does not match their modern forms, which there likely are, one of a fixed set of conclusions could be drawn:

1) The model, which is based on current knowledge of French historical phonology, is missing one or more shifts that may have occurred in the past, which still need to be discovered, and modified the words in question so that they don’t obey the current set of diachronic rules.

2) The words in question did not actually develop directly from Latin in French. Most likely, the source is another Romance language. There are plenty of known cases of this-- *loup* (wolf) in French does not develop directly from Latin *lupus*, which would have regularly rendered *leu*, but rather was a borrowing from Occitan. Alternatively, the ultimate source may not be Latin at all. For example, *aller* (to go) was once attributed to Latin *ambulare*, but this would regularly render, and *did* render *ambler*; others suggest the origin is Gaulish and cognate to Cornish *ellev*. 
3) The words in question were affected by non-Neogrammarian phenomena like stylistic restoration, spelling pronunciations, or analogy, which are not presently incorporated into the model.

4) Lexical diffusionists might have been right all along that the Neogrammarian doctrine is wrong, and much of Historical Phonology needs to be rethought.

In addition to alerting us in this way to diachronic irregularity for certain words, this opens up the possibility of doing statistical analyses with the irregularity of outcomes as the dependent variable. One could test any number of phonological (“is a certain phone present at a given time?”), morphological, syntactical, or sociolinguistic (“is the word likely part of a higher or lower register?”) predictors, and if any of them are correlated strongly with outcomes under the current model that don’t match the modern (or attested historic) forms, this may lead to new discoveries for French historical phonology.

It should be noted that even currently, there are some diachronic changes that have been added to the model that are not part of the current consensus on the chronology of diachronic changes in French. These were added because the simulation had failed to yield the correct outcomes for various words. These are currently marked in comments in the code, and should be removed before any statistical analysis as described above is executed – but why it was necessary to add them in the first place should also be investigated.

8d. Analogous automated simulation for Latin loans in Albanian

It is easy to see how one could use this model as an example to embark on projects of the same vein, simulating the diachronic developments of other languages. A seemingly natural direction to go would be to construct a similar model for the development of Latin loans in Albanian from the Roman Imperial period until the present. On the one hand, Albanian is full of Latin loanwords that came
primarily during the Imperial period. These loans penetrate deep into the core vocabulary of the language, including the words peshk (fish, <Latin piscis), qen (dog, <Latin canis) and even function words like që (that, <Latin que). Indeed it is very hard to speak Albanian and avoid using Latin loans, and the name of the language itself, shqip, comes from the verb shqipoj, to enunciate, which is itself thought to from Latin excipere.

On the other hand, despite notable work, Albanian remains understudied. Albanian historical phonology does not have any equivalent to great work produced by Pope [1934], which advanced French historical phonology remarkably by giving a comprehensive chronology and analysis from various linguistic perspectives of the shifts that affected and produced French over the centuries. The closest thing was a survey started by Eqrem Çabej and completed by Vladimir Orel (2000), but it is done from a much less fine-tuned perspective than that of Pope’s 1934 treatise. In addition to neglect due to its peripheral position, unlike French, Albanian linguistic study is hampered by the fact that the earliest attested Albanian corpus of any size only dates back five centuries and is confined to the northern Gheg dialect.

The historical phonology of Albanian is important. Albanian is the only surviving language of its branch within the Indo-European language family, and for this reason it could play a significant role in the reconstruction of Proto-Indo-European, similar to the role played by Gothic for proto-Germanic, if only we knew more about its phonological history. Albanian has a particular role to play in at least two major debates in Indo-European linguistics. The first is the nature of the centum-satem isogloss. Albanian may present a case of a part-satem, part-centum language, because the modal “satem” shift of Proto-Indo-European /k~c/ fronting seems to have been blocked in Albanian in certain contexts; some

---

15 In the 1300s, Catholic priests appear to have made reference to Albanians writing in their own vernacular separately from the typical tendency to write in Latin, but these writings tantalizingly appear to have been lost
researchers instead believe this represents partial “desatemization” [Matasović 2012: 14]. Albanian also plays a major role in the Indo-European Laryngeal Theory, with some linguists claiming that Albanian uniquely preserves reflexes of proto-Indo-European laryngeals (though others disagree— for a summary of the debate, see [Matasović 2012: 8-9]). Of course, to make any real conclusions, we need more info on the history of these supposed reflexes. The question of the provenance of Albanian also remains disputed, as it is still disputed whether it descends from Illyrian, Thracian, Dacian, or some other unknown ancient Balkan language – further historical phonological work could help clear this up.

Although the forms that native Albanian words had millennia ago is sometimes disputed, we do know how Albanian’s ancient Latin loans were pronounced two millennia ago because they are Latin. Once assimilated into the language, these words were affected by the same diachronic changes as the rest of the lexicon, so a system built to simulate the development of the Latin words could also be used to test possible forms the native vocabulary may have had during the classical period\(^\text{16}\). Helping reconstruct the state of Albanian at that time would be a great step forward for the advancement of Albanian historical phonology.

8e. Application of Machine Learning algorithms

In addition to the increasing use of computational methods for Historical Phonology, especially automated simulation methods (discussed in its own section above), it is my belief that Machine Learning applied in various ways in the field could yield interesting insights. In the realm of wider phonology, the past few decades have seen very significant expansions of the role of learning algorithms, often in association with Optimality Theory [for a summary, see Jurafsky & Martin 2007: 409-410, 414-416]. The first that comes to mind is work towards the goal of having Machine Learning

\(^{16}\) That is, once the changes involved in assimilating the Latin loans into the Albanian sound system are accounted for. Thankfully, considerable work on how Albanian assimilates Latin loans has been done already [see Orel 2000]
algorithms that can, to some degree of accuracy, detect what shifts occurred and when they occurred, given large sets of input-outcome pairs. This would be made possible by using the Lexicon class to represent the entire lexicon of relevant words for comparison as pairs of Latin (or Gaulish) inputs and modern French outcomes. It will be interesting to see what such a Machine Learning algorithm can or cannot accomplish. In building such an algorithm, one could test first how well it works for smaller time scales – say, Latin to Gallo-Roman or Old French, or Middle French to Modern French – and then, once it seems to clearly work for those smaller time-scale cases, apply it to the whole learning challenge of detecting the Latin to French phonological developments.

The results of how a learning algorithm comes to decide what shifts are likely to have occurred when may also turn out to be interesting, especially if factors like the language’s phonological inventory at a given time, and the distribution of phones in words are considered. In the realm of grammatical gender, a notable work by Polinsky and Everbroeck [2013] interestingly shows that their connectionist learning algorithm had better performance for a similar simulation program of predicting gender shifts among nouns when information on the gender of the noun in Gaulish was incorporated into the model, providing inspiration for this idea. Of course, to consider these things, the lexicon class will need to realize its original intent as representing all the relevant words in the language at the given time, rather than a demonstrative subset. If it turns out that the learning algorithm ends up assigning higher likelihood to, for example, a shift involving the fronting of a vowel which formerly had a back articulation (as happened with both historical Latin /u/ and /a/ in French phonological history) during a period where there had been fewer front vowels than back vowels, this could be interesting as it would seem to match theories that assigned that relative dearth to be the cause [see Posner 1997: 252]. It would be more interesting if those sorts of results could be replicated for other languages.

Additionally, in the specific case of French, historical linguistic scholarship already has a pretty good picture of what major foreign influences (Gaulish, Frankish, later etc) affected French, in what
periods they affected French, and even how their effects manifested in French. We can therefore assign certain phenomena (for example, spirantization of k before s and t) to language contact phenomena (in this case the Gaulish substrate), while knowing that certain other shifts in French were likely driven by solely internal factors. This opens the possibility of using a Machine Learning system to try to predict specifically what shifts will occur internally to the language, excluding those that were promoted by outside influence, given factors like the language’s phonological make-up before the onset of an internally driven shift. If it is indeed found that internal shifts exhibit characteristics that are in any way predictable given various internal factors and different in patterning than externally promoted shifts, this might point to possibly useful insight for the still-ongoing debate about Ascoli’s hypothesis and the proper role of language contact in Historical Phonology.

9. Bibliography


Kerkhof, Peter Alexander. 19 January 2016. “The case for Late Gaulish: the old and the new evidence reevaluated”. Public lecture, SV-Astérix, notes uploaded by Pr Kerkhof at: [https://www.academia.edu/20508869/The_case_for_Late_Gaulish_the_old_and_the_new_evidence_reevaluated](https://www.academia.edu/20508869/The_case_for_Late_Gaulish_the_old_and_the_new_evidence_reevaluated)


ALINE System also available online : [https://webdocs.cs.ualberta.ca/~kondrak/](https://webdocs.cs.ualberta.ca/~kondrak/)


http://www.pescarini.it/pubblicazioni/2016_Beninc%E0_Parry_Pescarini_NorthDial.pdf


Sidonius Apollinaris. 474. *Letters, III.3.2*


Stifter, David. 2008. “Old Celtic Languages: IV. Gaulish”. Available online:  

Stifter, David. 2012. *Old Celtic Languages*. University of Kopenhagen

