Applications of Computational Simulation to Morphological Class Change: A Case Study in Romance*

TYLER LAU

University of California, Berkeley

1. Introduction

It is no coincidence that many closely related languages undergo parallel changes. While close contact is one factor in precipitating parallel changes, the seeds of change may also be found in an ancestor language. Although the factors may be speculated upon, it is difficult to ascertain the extent to which each plays a role. To isolate causes, one must carry out an experiment manipulating factors. However, as language change occurs in natural settings over generations and is generally observed after the fact, it is not feasible to run an experiment in order to answer this question.

Computational simulation offers a tool to control parameters, providing insight into the roles of different factors in language change. Latin is well known for having a nominal case system that simplified greatly in all the daughter Romance languages. I explore the question of what the seeds of change in Latin were by employing computational simulation of morphological change on a corpus of Latin and test how closely history can be approximated given minimal data. Two parameters are toggled: 1) whether sound changes affecting nominal suffixes have taken place, and 2) whether nouns are introduced to the model only by the *type frequency* of a specific case/number combination or also by the *token frequency* of the lemma.

The simulation reveals that the merger of Declensions IV with II and of the neuter with the masculine as well as the merger of the genitive and dative cases in

* Many thanks to Maria Polinsky, Andrew Garrett, Terry Regier, Kevin Ryan, James Kirby, Mairi McLaughlin, and Yang Xu for comments and guidance through various stages of the project, to the audience at the UC Berkeley historical linguistics discussion group *dhworom, at the 2nd Edinburgh Symposium on Historical Phonology, at the 90th annual LSA meeting, and at the 28th UCLA Indo-European Conference for helpful comments. I would also like to thank Ezra Van Everbroeck for providing the code for the simulation in Polinsky and Van Everbroeck (2003), Jake Seaton for helping with the early stages of the connectionist simulation, and to Marek Majer for advising on Romance sound changes.

David M. Goldstein, Stephanie W. Jamison, and Brent Vine (eds.). 2017.
Proceedings of the 28th Annual UCLA Indo-European Conference. Bremen: Hempen. ###–###.

Romance could have occurred regardless of sound change, whereas the merger of the nominative, accusative, and ablative cases was crucially brought about by sound change. The simulation also reveals that only type frequency and not token frequency is relevant to morphological class change, corroborating previous simulation work on morphological change.

Section 2 provides background on the Latin nominal system and its reflexes in the daughter languages. Section 3 discusses computational simulation and its applications to questions of language change and explains the model used in this study. Section 4 details the results of the simulation, while Section 5 discusses the implications. Finally, Section 6 draws conclusions.

2. Background

2.1. The Latin nominal system

The Latin nominal system consists of five declensions, three genders, two numbers, and six cases. It is probable that the masculine, feminine, and neuter in both the singular and plural were inherited from late Proto-Indo-European (Weiss 2009:195). Meanwhile the case system simplified from eight cases to six: nominative, accusative, genitive, dative, ablative, and vocative. Both the Proto-Indo-European instrumental and locative cases merged with the ablative, with the exception of some toponyms and relic forms (Weiss 2009:213–4). The Latin paradigm for masculines (primarily in II, III, and IV) and feminines (primarily in I, III, and V) are provided in Table 1.

Table 1. The Latin nominal system

		I	II	IIIa	IIIb	IV	V
	Root	*silv(a)-*	*ann(o)-*	*color-*	*ign(i)-*	*lac(u)-*	*fid(e)-*
	Gloss	‘forest’	‘year’	‘color’	‘fire’	‘lake’	‘faith’
Sg	Nom	*silv-a*	*ann-us*	*color*	*ign-is*	*lac-us*	*fid-ēs*
	Gen	*silv-ae*	*ann-ī*	*colōr-is*	*ign-is*	*lac-ūs*	*fid-eī*
	Acc	*silv-am*	*ann-um*	*colōr-em*	*ign-em*	*lac-um*	*fid-em*
	Dat	*silv-ae*	*ann-ō*	*colōr-ī*	*ign-ī*	*lac-uī*	*fid-eī*
	Abl	*silv-ā*	*ann-ō*	*colōr-e*	*ign-e/ ign-ī*	*lac-ū*	*fid-ē*
	Voc	*silv-a*	*ann-e*	*color*	*ign-is*	*lac-us*	*fid-ēs*

Table 1 continued

		I	II	IIIa	IIIb	IV	V
	Root	*silv(a)-*	*ann(o)-*	*color-*	*ign(i)-*	*lac(u)-*	*fid(e)-*
	Gloss	'forest'	'year'	'color'	'fire'	'lake'	'faith'
Pl	Nom	*silv-ae*	*ann-ī*	*colōr-ēs*	*ign-ēs*	*lac-ūs*	*fid-ēs*
	Gen	*silv-ārum*	*ann-ōrum*	*colōr-um*	*ign-ium*	*lac-ūm*	*fid-ērum*
	Acc	*silv-ās*	*ann-ōs*	*color-ēs*	*ign-īs/ ign-ēs*	*lac-ūs*	*fid-ēs*
	Dat	*silv-īs*	*ann-īs*	*colōr-ibus*	*ign-ibus*	*lac-ubus*	*fid-ēbus*
	Abl	*silv-īs*	*ann-īs*	*colōr-ibus*	*ign-ibus*	*lac-ubus*	*fid-ēbus*
	Voc	*silv-ae*	*ann-ī*	*colōr-ēs*	*ign-ēs*	*lac-ūs*	*fid-ēs*

Neuter nouns exist only in Declensions II–IV and differ slightly in that the nominative and accusative singular are identical, usually both ending in *-um* or nothing. The relative frequencies of the declensions and genders are presented in Table 2. The overall gender and declension frequencies in Latin were calculated by Polinsky and Van Everbroeck (2003). Table 2 also breaks down the gender frequency within declension, calculated with the study's corpus (see §3.2).

Table 2. Frequency of gender per declension class and overall frequency

	I	II	III	IV	V	Overall
F	97.73%	0.69%	37.96%	4.55%	87.5%	40.8%
M	2.27%	37.5%	38.9%	90.9%	12.5%	38.1%
N	0%	61.8%	23.15%	4.55%	0%	21.1%
Overall	21.6%	23.7%	52.6%	1.4%	0.7%	

Declensions I and V consist of primarily feminine nouns, Declensions II and IV of primarily masculine and neuter nouns, and Declension III contains a mix of all three genders. The neuter gender is a smaller class than the other two, and Declensions IV and V are very small, but contain high-frequency nouns.

2.2. Changes from Latin to Romance

All Romance languages share a number of innovations in the nominal system. By the time of late Latin, Declension IV had merged with Declension II, while Declension V nouns had primarily joined Declension I, though some joined Declension III (Maiden 2011:163). Most Romance languages merged the neuter class fully with the masculine class, both of which were well-represented in both Declensions II and IV, a process that possibly began in the preclassical period

(Wilkinson 1985). Many Romance languages also reanalyzed a minority of neuter plurals (such as *folia* 'leaves') as feminine singulars, causing migration of some neuters to the feminine class. Herman (1967:65) and others have argued that the reanalysis was driven by the interpretation of neuter plurals as collectives.

Remnants of the Latin neuter plural are preserved in some Romance varieties. The neuter class in Latin has reflexes in Daco-Romance as an ambigeneric class; rather than taking unique agreement as in Latin, singular nouns take masculine agreement while neuter plural nouns take feminine agreement (Alkire and Rosen 2010:282). However, only a subset of these nouns maintain the neuter plural *-a* ending of Latin. Of the feminine endings in Romanian, *-uri* is one that derives from the common *-ora* ending of Declension III in Latin (Alkire and Rosen 2010: 282). Italian and Old Provençal also have ambigeneric nouns descended from Latin neuters whose plural ends in *-a* (Petrucci 1999:92–7). The partial preservation of the neuter in Romanian has been attributed to contact with South Slavic languages, which also retain the Indo-European neuter class (Hořejši 1964:405, Graur 1967:27).

Three case systems are attested in Romance languages, with only the first two surviving into the modern descendants (Ledgeway 2012:333).

1) No case (most Romance languages)
2) Nominative/Accusative versus Genitive/Dative (Daco-Romance)
3) Nominative versus Oblique (older Gallo-Romance & Rhaeto-Romance)

The order of case mergers is rather murky. One view, shown in Figure 1, from Banniard (1992:518) proposes that the ablative merged early with the genitive/dative to form an oblique case in opposition to the nominative and accusative.

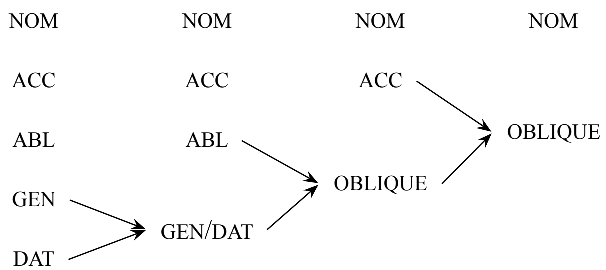


Fig. 1. Stages of loss of Latin case (Banniard 1992:518)

Clackson and Horrocks (2007:253, 277) note that there is already some evidence of confusion between the accusative and ablative around the 2nd century CE and that there is widespread evidence of the genitive/dative merger in the 5th and 6th centuries CE.

Two sound changes that occurred by the 5th century CE (Herman 1967:52) led to syncretism of case endings and may therefore be part of the reason for the collapse of the case system. The first was the overhaul of the Latin vowel system, which included the loss of length distinction. In Western Romance, short high vowels lowered such that /i/ and /u/ merged with both long and short /e(:)/ and /o(:)/ in case endings.[1] The second was the loss of word-final nasal *m*, the spelling of which likely only represented nasalization, if anything (Clackson and Horrocks 2007:97). These sound changes effectively eliminated the difference between the accusative and ablative singular endings, facilitating their merger.

The Latin cases are generally believed to have collapsed to the accusative, as can be seen in the lack of final -*s* in the singular and its presence in the plural in Western Romance (Herman 1967:55, Ledgeway 2011:461), although Penny (1980) argues that there is also descent from the nominative. In languages such as Italian and Romanian, final -*s* disappeared and thus it has been argued that the forms could have descended from the nominative, particularly as the Eastern plural endings -*e* and -*i* can be derived via sound change from the Latin nominative feminine -*ae* and masculine -*ī* (D'hulst 2005:1304). However, Maiden (1996:178) argues that the lack of palatalization in these plural forms, along with other evidence, suggests that these were descended from the accusative. D'hulst (2005: 1314) proposes a pathway in which word-final *s* became *i* and subsequent monophthongization occurred (-*as*, -*os* > *-*ai*, *-*oi* > -*e*, -*i*) to support an accusative origin. Regardless, there do appear to be vestiges of the nominative in various Romance languages, particularly in animate nouns (Ledgeway 2012:333–5). The merger of the nominative and accusative was likely very late, as it was hardly confused even in 6th and 7th century Merovingian texts (Herman 1967:54–5). How exactly the cases merged is one of the primary objects of study in this paper.

Latin is used for the current study for two reasons. First, it is a documented ancestor language to the Romance languages, which all underwent convergent evolution to lose both the case system and also the neuter gender (except for

1 In stressed positions, however, short /e/ and /o/ became the open-mid /ɛ/ and /ɔ/. In Romanian, the same pattern occurs for the front vowels, but the back vowel /u/ merges with /u:/ instead of with /o/.

Romanian). Second, the large corpora of Latin provide frequency information, a crucial factor in language change. Rather than chalking the parallel changes up to contact or chance, I propose that the seeds of these changes may be found in Latin itself, a hypothesis that can be tested because of the fortunate availability of data from the ancestor language.

3. Modeling language change

Computational simulation is a method that has been increasingly applied to language change. Connectionist simulation is one method that has been extensively employed in the learning of phonological and morphological patterns, (Rumelhart et al. 1987, Daugherty and Seidenberg 1994, Plunkett and Juola 1999). Connectionist simulations, which will be described in detail in §3.1, train a model to link features of an input to those of an output in order to predict outcome forms. Implicit in connectionist models is an analogical, rather than a rule-based, model of morphological learning. Other models have been proposed; a list may be found in Albright and Hayes 2003:122, and the paper itself details the authors' own implementation of stochastic rules to morphological learning and their argument for its superiority over analogical models.

Connectionist models have also been applied to modeling morphological *change*. One landmark study was that of Hare and Elman (1995), who were able to reproduce English weak and strong verb class shifts with high consistency by teaching a computational model verb endings. Polinsky and Van Everbroeck (2003) followed up on Hare and Elman by applying their model to Latin nouns and factoring in contact with Gaulish to show that the loss of the neuter class in Old French could be attributed to contact. This study adds to the growing body of simulation of morphological change.

3.1. Connectionist modeling

Connectionist modeling is a computational technique that models phenomena as emergent processes arising from connections between simple units (Elman et al. 1998:xii). *Neural networks* are one form of connectionist modeling, so named because they simulate neurons firing in a brain. Neural nets attempt to learn the outputs corresponding to given inputs by forming associations, essentially tackling a categorization problem. The simplest kind of neural net consists of three layers: an *input layer* where the input is introduced, the *output layer* where the output is produced, and one *hidden layer*, which provides a stochastic element to the associations between input and output. Each layer may be imagined as a sequence of nodes with given values. Each node in a layer is connected to every node in an adjacent layer. In training, the neural net is initialized with a randomized *connection weight* for each connection and a randomized *activation value* for each hidden node. Figure 2 schematizes initialization. The lowest set of two nodes is the input layer, the middle two are the hidden layer, and the top two are the output layer.

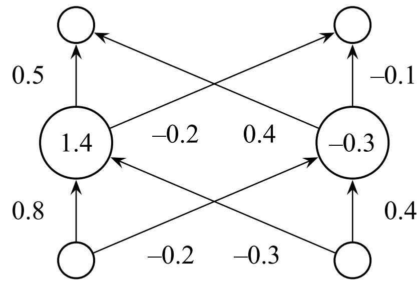


Fig. 2. Neural net initialization

In order to concretely understand how the neural net works, we take the word *focus* 'hearth', a Declension II masculine noun in the nominative singular form. We will imagine, as a toy example for the purpose of demonstration, that the input is merely the nodes representing *masculine,* coded into the binary sequence 1, 0 (see §3.2 for explanation of coding). The input nodes attempt to stimulate the hidden nodes via the following process: the products of each input node and its connection weight with respect to the hidden node are summed. In the equation below, connection weights are indicated by hyphens between the two relevant nodes.

1	×	0.8	+	0	×	–0.3	=	**0.8**		1	×	–0.2	+	0	×	0.4	=	**0.2**		
I1		I1–H2		I2		I2–H1		H1		I1		I1–H2		I2		I2–H2		H2		

The resulting value is then compared to the hidden node's activation value. If it is lower, the hidden node remains unactivated (0.8 < 1.4), receiving a value of 0, but if it is higher, the hidden node is activated (0.2 > –0.3), receiving a value of 1. This process is shown in Figure 3. The summation process is then repeated between the hidden and output layers (Figure 4):

0	×	0.5	+	1	×	0.4	=	**0.4**	0	×	–0.2	+	1	×	–0.1	=	**–0.1**	
H1		H1–O1		H2		H2–O1		O1	H1		H1–O2		H2		H2–O2		O2	

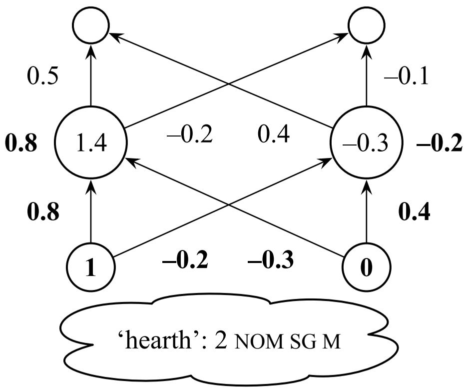


Fig. 3. From input to hidden layer

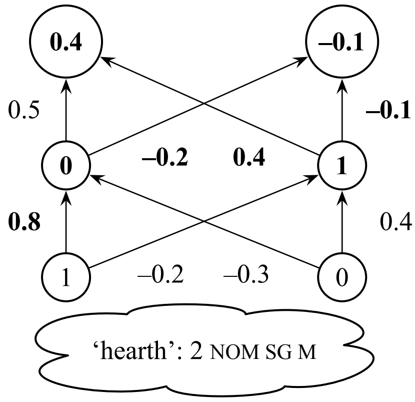


Fig. 4. From hidden layer to output

The computed values are compared to the actual (= expected) values (1, –1), provided by the phonological features of the suffix -*us* (we limit ourselves to the [+high] and [–front] features of /u/ for simplicity), as in Figure 5, and used to recalibrate the connection weights and hidden node activation values in order to update the model’s accuracy. The resulting net is then used for the following token in the corpus.

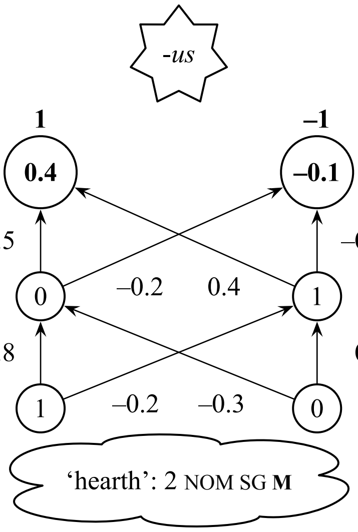


Fig. 5. Comparison to actual/expected output

Once the model is trained, a test set of forms is provided to the model. The now-established connection weights and activation values are used to calculate the output values. Figure 6 shows an example test run.

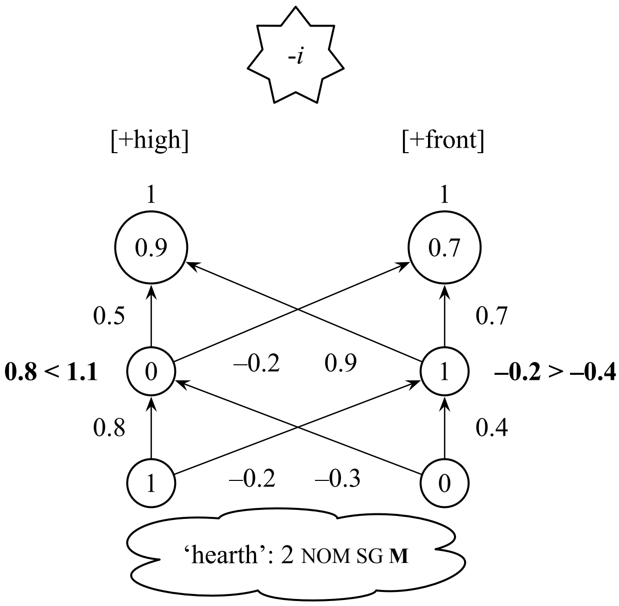


Fig. 6. Test run

The closest point to (0.9, 0.7) that represents a valid feature is chosen. Here the choices are (1, 1), (1, –1), (–1, –1), and (–1, –1). (1, 1) is chosen, translating to a *+high, +front* vowel /i/.

As in this toy example, the neural net may yield incorrect results. The modeling of morphological change takes advantage of these potential errors to simulate diachronic change. Imagine that during the training session, the inputs and expected outputs are provided to a learner by a teacher. The learner forms hypotheses by calibrating the connection weights and activation values such that certain values of certain nodes become associated with certain outputs. In the test session, the learner applies the hypotheses. The learner inevitably miscategorizes some inputs and the "erroneous" outputs become the expected outputs for the following generation, in which the learner becomes the new teacher. This process of generational learning is schematized in Figure 7.

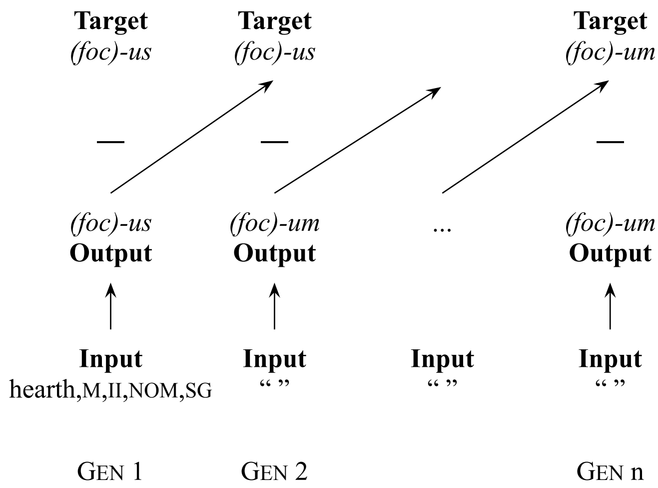


Fig. 7. Simulating generational learning

The miscategorization that occurs is in part due to noise but may also occur because of analogy with other forms. For example, if the nodes representing *feminine* become associated with those representing *a*, the input for a feminine word *manus* 'hand' may be predicted to output a suffix beginning with *-a* rather than the masculine-looking *-us* ending. These associations via phonological similarity incur analogical processes that underlie morphological class mergers and shifts, creating a model that can simulate language change.

3.2. Corpus

The corpus used for the current study was compiled using Delatte et al.'s (1981) frequency dictionary of Latin. The dictionary consists of counts of Latin nouns gathered from twenty-six written sources from between the 1st century BCE and the 2nd century CE. The counts are divided by genre (prose: 161,652; poetry: 68,600). The 500 most frequent nouns were culled from the dictionary, excluding words that appeared ten or fewer times in either genre. Words that disproportionately appear in one genre, such as *sceptrum* 'scepter' (prose: 2; poetry 59) or *colonia* 'colony' (prose: 55; poetry 0), may be specialized and thus unlikely to be heard by a child at high frequency.

Also gathered from the frequency dictionary were the counts for each case/number combination. The relative ratios were calculated with the lowest occurring combination, the dative plural, set to 1 and treated as *type frequency multipliers*, to be used to determine the number of times each form of each token is added to the corpus.[2] The vocative case was excluded because its low count would greatly increase the multipliers, leading to an intractably large corpus.[3] The counts and frequency multipliers may be found in Table 3.

Table 3. Counts for case/number combinations and type multipliers

	Total counts		Ratios	
	Singular	Plural	Singular	Plural
Nominative	41,617	12,738	8.23	2.52
Accusative	42,709	30,327	8.45	6.00
Genitive	21,639	10,467	4.28	2.07
Dative	8,222	5,056	1.63	1.00
Ablative	39,345	14,440	7.78	2.86
Vocative	2,243	611	N/A	N/A

2 Using the same type multiplier regardless of lemma is for model simplification. Different classes of nouns are expected to appear in different cases at different rates. For example, animate nouns should be more common in the nominative case than in the accusative whereas the opposite is expected for inanimate nouns. A more complex model with type multipliers specified by noun should be more accurate in predicting retention of some cases for certain nouns (as in the retention of the nominative rather than the accusative form for some animate nouns in French).

3 This unfortunately abstracts away from reality, as there are Romanian dialects that retain the vocative in masculine nouns.

The case/number combination of each form is coded into an input and expected output. The input contains a *root identifier* demarcating the specific lemma and its *declension, gender, case,* and *number*. All this information is converted into a sequence of 18 bits. As an example, the feminine Declension I lemma *silva* is the 412th word in the corpus and thus its root identifier is 412 converted into binary. The entire binary sequence of the genitive plural form *silvārum* is shown below.

[1 1 0 0 1 1 1 0 0]	[0 0 0]	[0 0]	[0 1 0]	[1]
Root ID	Declension	Gender	Case	Number
'forest'	I	F	Genitive	Plural

The expected output is the phonological form of the case suffix. The case suffix of *silvārum, -ārum*, is represented abstractly in the first row of Figure 8, with the accusative singular suffix *-am* in the second row for comparison. The case suffix is first converted to a template of seven potential phonemes (long vowels are treated as two separate phonemes), with the first three (in a *VVC* skeleton) representing the first syllable and the second four (in a *CVVC* skeleton) representing the second syllable. All case suffixes in Latin begin with a vowel, so the first syllable lacks an onset slot.

a a – r u – m

a – m – – – –

Fig. 8. Case suffix template

This template lines up the relevant parts of each case suffix; otherwise if the phonemes were read linearly, the same nodes associated with the second *a* in *-ārum*, which marks Declension II, would be associated with the *m* in *-am*, which marks accusative case. Each phoneme is decomposed into six nodes, each representing a phonological feature: [±sonorant], [±continuant], [±high], [±front], [±low], [±back]. These features were chosen as the minimal set distinguishing the nine phonemes that appear in Latin suffixes: /b s m r i e a o u/. 1 represents feature presence, –1 feature absence, and 0 feature irrelevance.[4] The lack of a pho-

4 Thus, consonants are coded as 0 for the high, front, low, and back features. While vowels are sonorant and continuant, these features are coded as 0 for this simulation in order to maximally distinguish vowels from consonants. Coding them as 1 for these features leads the simulation to predict suffixes such as *-rs* or *-rm* as *r* is the only consonant coded as 1 for both these features. The use of these features is in essence a way to code the four consonants as maximally

neme is represented by a vector of seven zeroes. A template consists of seven phonemes and each phoneme is defined by six features, so the output layer has 42 nodes. The number of hidden nodes is set to 30, the average of 18 (number of input nodes) and 42 (number of output nodes).[5]

To determine the number of times each form is included in the training corpus, the *base 10 log* of the frequency count of the noun is taken and multiplied by the *type frequency multiplier* of the case/number combination (shown in Table 3) and rounded down, following Polinsky and Van Everbroeck (2003:369). The log is preferred to the raw frequency under the assumption that as words become more frequent, their relative frequencies with respect to words with similar frequencies decreases exponentially—thus the difference between appearing one and ten times is much greater than between appearing 1,001 and 1,010 times. A word with token frequency of 1,000 will be included in the corpus eighteen times in the accusative plural: log(1000) × 6.00 (accusative plural multiplier) = 18. The total number of tokens in the corpus was 47,722. The training session was set to three epochs (an epoch refers to one full round of introducing the corpus), also following Polinsky and Van Everbroeck (2003:374).[6] After training is done, each lemma in each case/number combination is introduced once for the test set; thus the size of the test set is 5,000 words (500 lemmas × 10 cases). The results of the test session are used in the following generation and the process is repeated for ten generations.

3.3. Parameters and hypotheses

For this study, I toggle two parameters: 1) token frequency and 2) sound change. Many studies have shown that token frequency is irrelevant to morphological class changes and that using it actually *worsens* computational models (Bybee 1995:434, Bybee 2001:13, Pierrehumbert 2001:198, Albright and Hayes 2003: 133). Thus, large word classes attract words from other classes, but small word classes do not, even if the words inside them have a high frequency. If the token frequency parameter is off, only type frequency multipliers are used in creating the training corpus.

---

distant from one another (i.e., they lie on a plane at the points (1, 1), (1, –1), (–1, –1), and (–1, 1)) and the five vowels in a pentagonal shape roughly corresponding to their locations in the vowel space.

5 See Heaton 2008:158–9 for other methods for choosing the number of hidden nodes.

6 For comparison, Hare and Elman (1995:75) use ten epochs.

If the sound change parameter is on, the following sound changes are applied to the end results of each simulation:

1) Final *m* loss
2) Lowering of high vowels
3) Shortening of long vowels

With these changes in place, the accusative and ablative singulars should merge for all five declensions to a form consisting only of the vowel signifying the declension. Note that Romanian merges /u/ not with /o/, but with /uː/ to /u/, which may prevent the accusative (*-um*) and ablative (*-ō*) in Declension II from merging in a simulation involving Romanian sound changes. I leave the toggling of language-specific sound changes to future models. The four combinations of parameters are each run for fifty trials. Each trial may be thought of as a potential descendant of Latin.

4. Results

Two types of measures were taken from the final outputs of each simulation. One measure was the percent of output suffixes that matched the expected suffix in the descendant Romance languages (i.e., *-a* for feminine singular nouns, *-o* for masculine singular, etc.). The second measure was the percent similarity between given declension/gender/case/number combinations. For this measure, the most common ending for a given combination was taken as the representative of that combination. For example, if 15 Declension II masculine nominative singular nouns yield *-us* but 175 yield *-um*, *-um* is coded as the representative of that combination. In the toy example in Table 4, the percentage similarity between the nominative and the accusative is calculated as 75%.

Table 4. Example of percent similarity calculation

	Declension I			
Case	Singular		Plural	
Nominative	F: ***-a***	M: ***-a***	F: ***-as***	M: *-i*
Accusative	F: ***-a***	M: ***-a***	F: ***-as***	M: *-as*

The histograms in §4.1 and §4.2 plot either the percentage matching an expected ending or the percentage similarity between noun classes on the x-axis against the percentage of trials in which the given result occurs on the y-axis.

4.1. Case mergers

Charts 1 and 2 explore the genitive/dative merger. Toggling sound change (compare left and right graphs) yields little difference, but turning token frequency off (compare Chart 2 to 1) causes the percentage of complete genitive/dative match to jump from 42% to 56%.

Chart 1. Token frequency on, sound change off versus sound change on

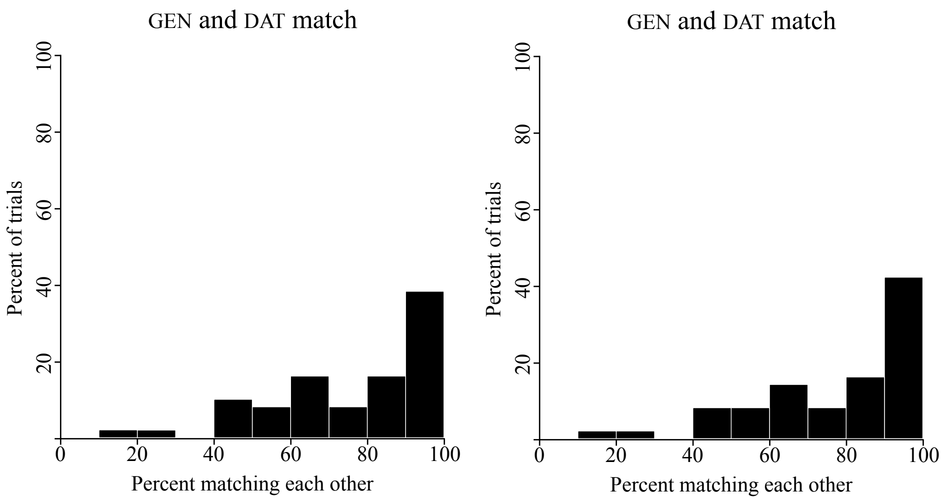


Chart 2. Token frequency off, sound change off versus sound change on

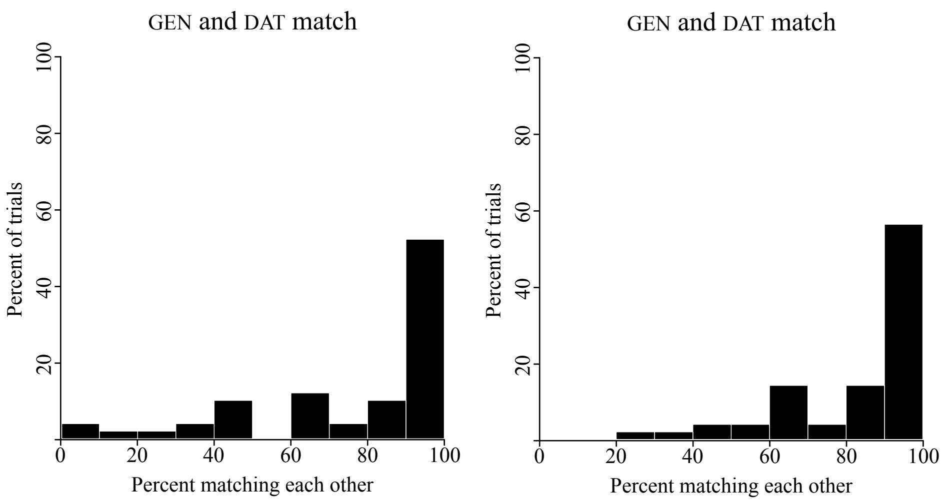


Chart 3 demonstrates that, as predicted, including sound changes greatly increases the percentage of accusative and ablative forms matching one another. The model is improved even further when token frequency is turned off (Chart 4), bringing the percentage of trials with over 90% match from 26% to 56%.

Chart 3. Token frequency on, sound change off versus sound change on

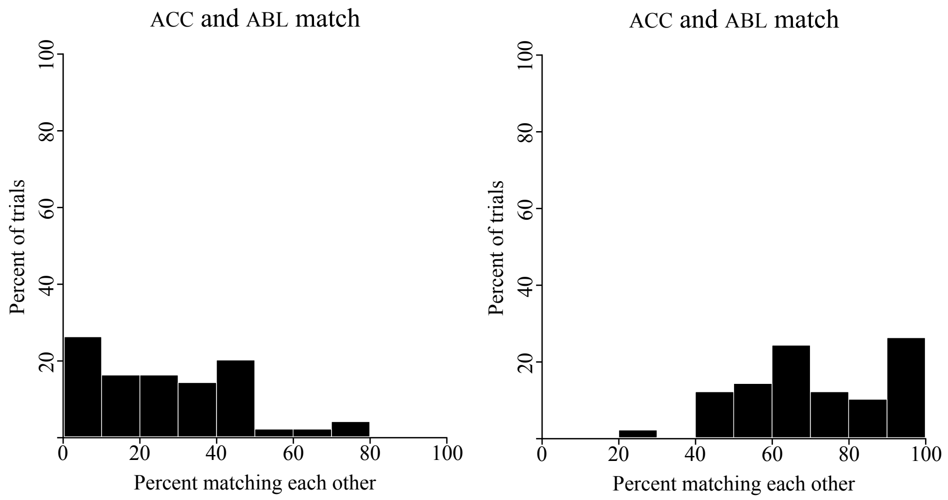


Chart 4. Token frequency off, sound change off versus sound change on

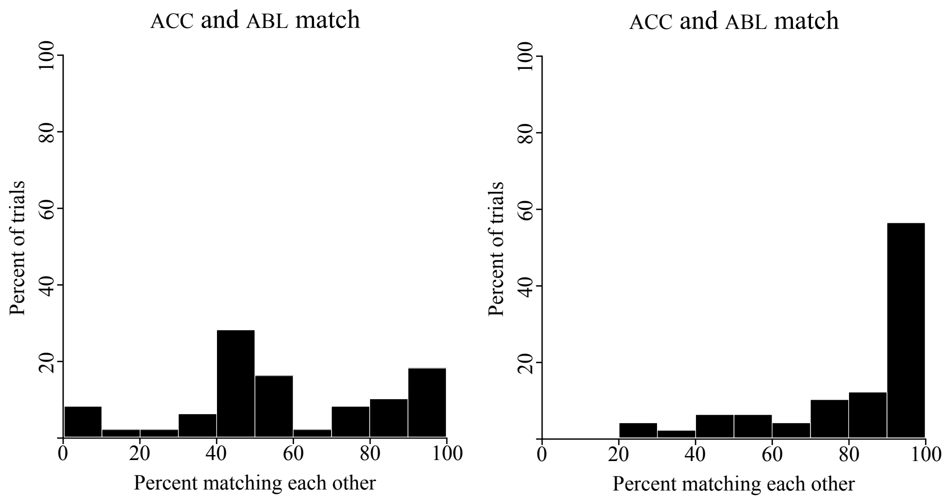


Chart 5 shows that the nominative also appears to pattern with the accusative and ablative in a fair percentage of cases (assuming sound change). Once again, disregarding token frequency improves the model greatly.

Chart 5. Sound change on, token frequency on versus token frequency off

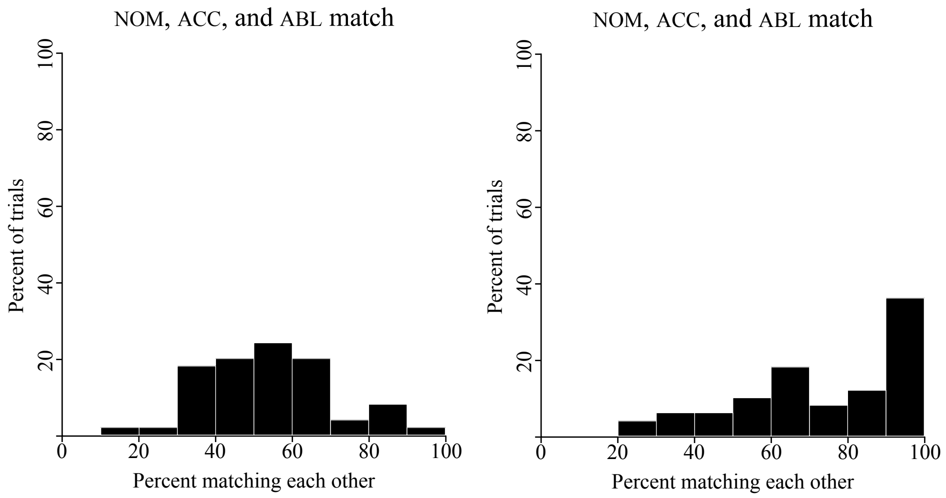
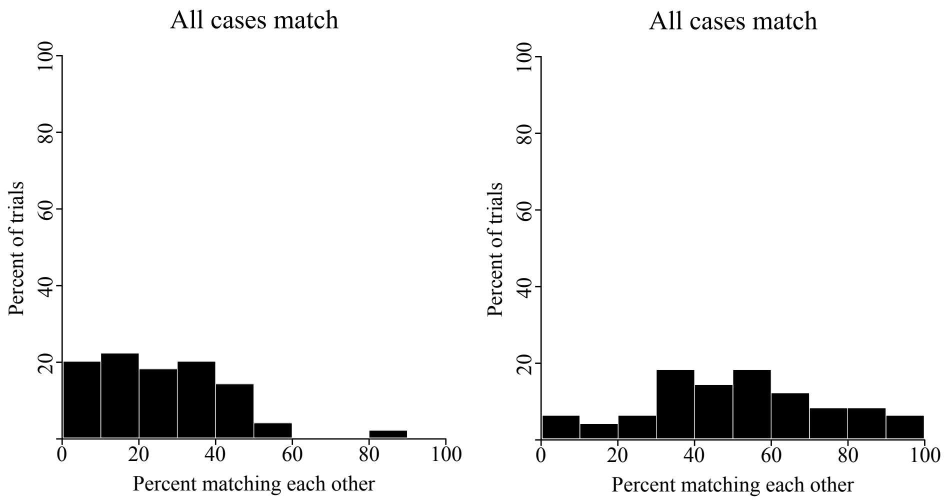


Chart 6 compares the percentage match of all the cases. Once again, removing token frequency great improves the model.

Chart 6. Sound change on, token frequency on versus token frequency off



However, the mere partial match suggests that while phonology and frequency alone can predict some of the case mergers, other factors must be taken into account in order to derive the complete loss of case in of most Romance languages.

4.2. Gender and declension mergers

Since it has been established that ignoring token frequency improves the model, the remaining results will be presented with this parameter off.

Chart 7. Token frequency off, sound change off versus sound change on

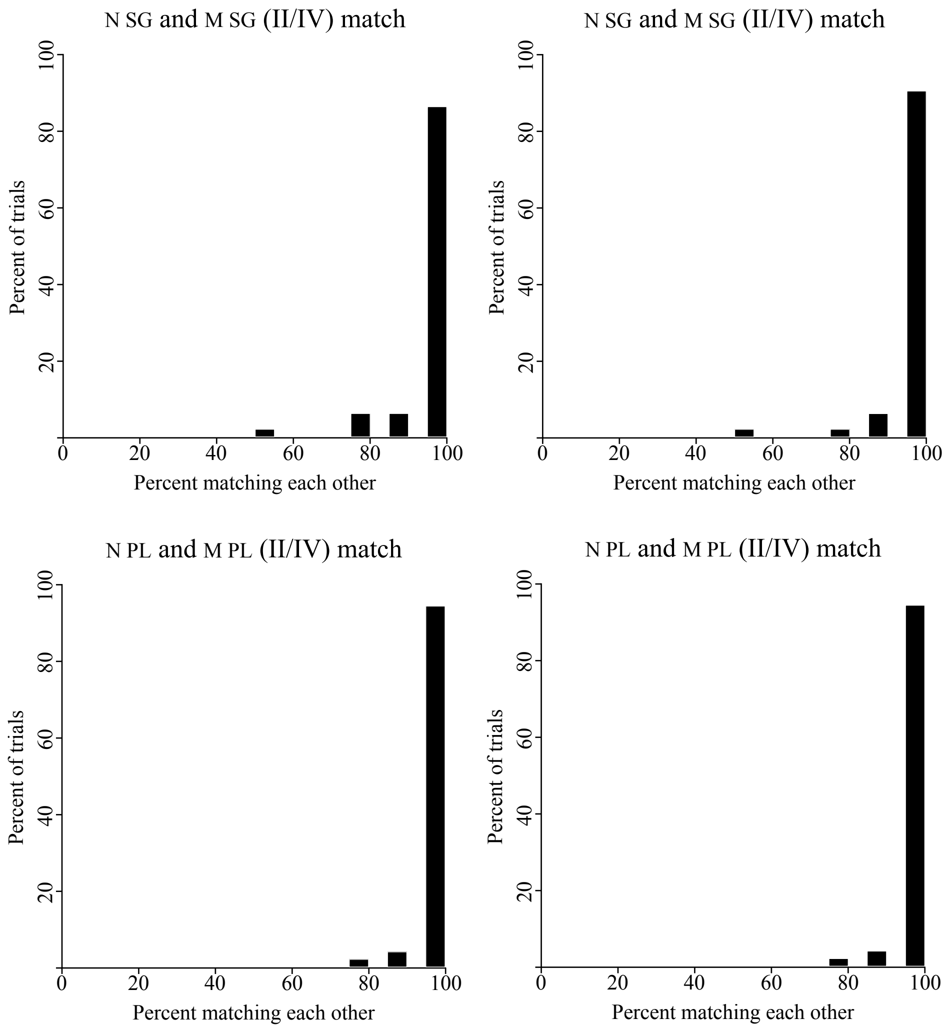
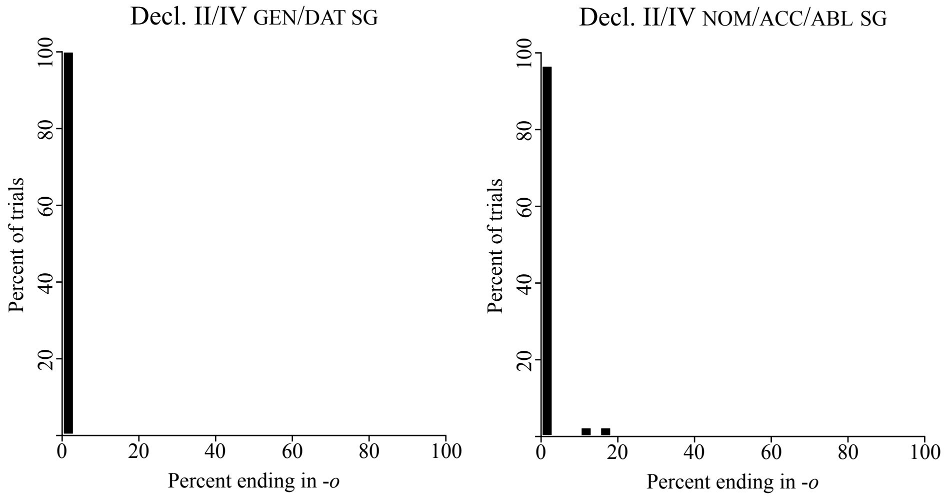


Chart 7 demonstrates that the neuter and masculine classes would merge regardless of sound change.[7] The merger of the neuter plural with the masculine plural in a majority of cases (and not with the feminine class despite the *-a* suffix) mirrors the history in most modern Romance languages.

Charts 8 and 9 demonstrate that the modern Romance two-way gender system of masculine *-o*(/*-u*/-∅) versus feminine *-a* is robust for the nominative, accusative, and ablative.[8] The implementation of sound change makes a large difference in the coalescence of the case suffixes to one vocalic element. Even with sound change, however, the genitive and dative still pattern differently from the nominative, accusative, and ablative and resist collapse to the expected endings, suggesting that sound change cannot in and of itself be responsible for the reassociation of the genders with a defining vowel.

Chart 8. Percentage of Declension II/IV nouns ending in *-o*

a. Sound change off



7 Declension III is left out of this analysis because of the inconsistency with which *-e* or -∅ is chosen as the representative ending.

8 Declension III may end in either *-e* or a consonant and contains nouns from both genders. It is also retained in both this model and in reality.

Chart 8 continued

b. Sound change on

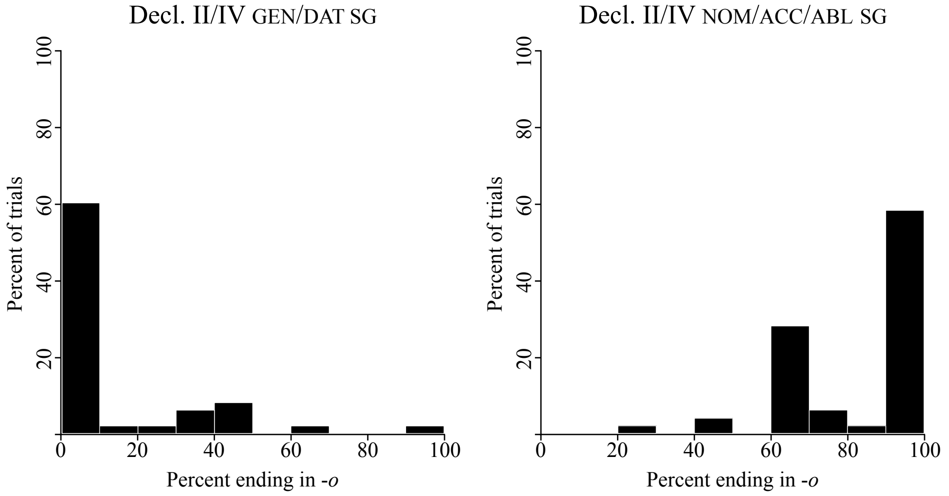


Chart 9. Percentage of Declension I nouns ending in -*a*

a. Sound change off

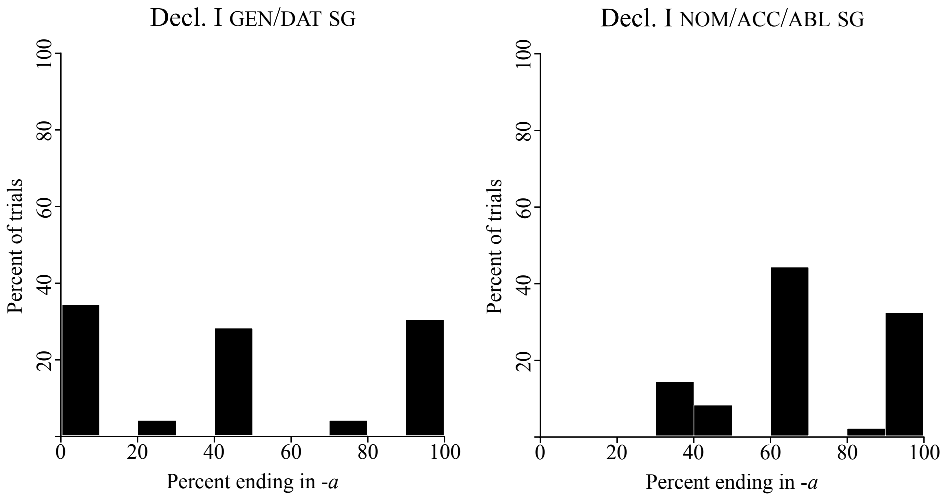
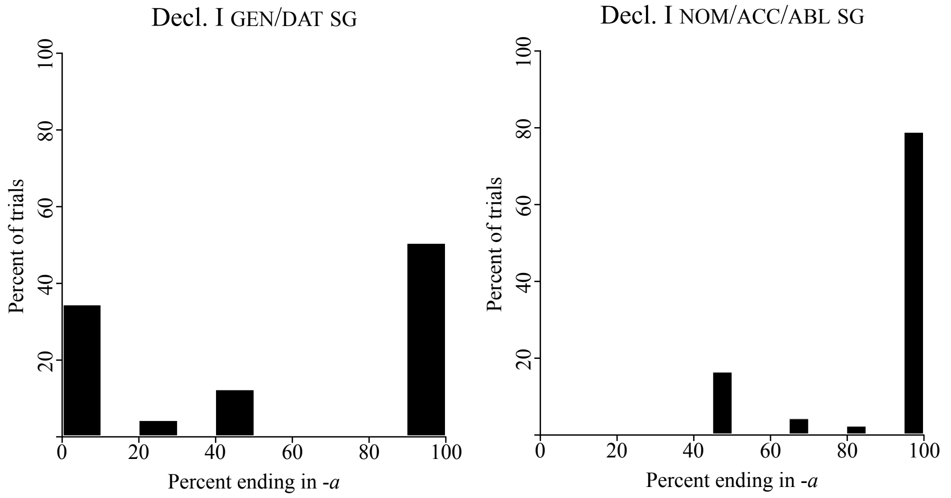


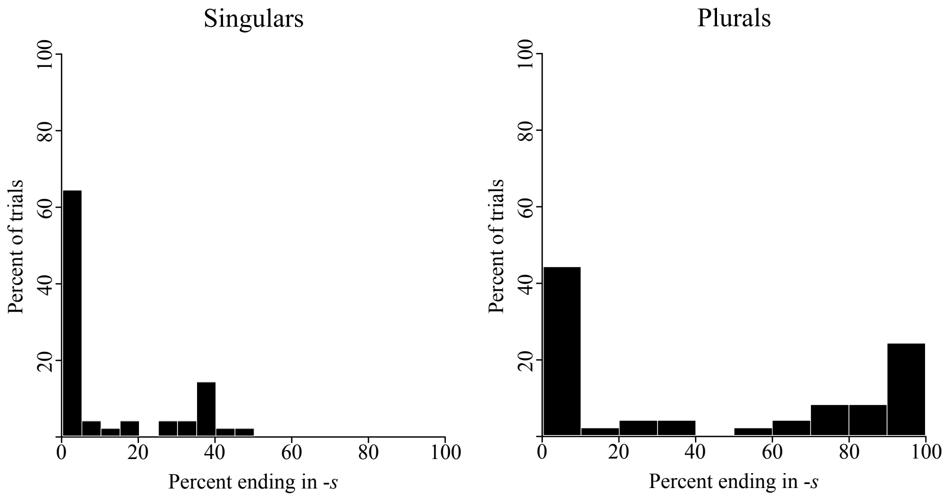
Chart 9 continued

b. Sound change on



Finally, the histograms in Chart 10 display the percentage of singular and plural nouns at the end of each trial ending in *-s*. *-s* never wins out as the ending for singulars, implying that the nominative case never takes over as the main form.

Chart 10. Percentage of plural nouns ending in *-s*



While a non-trivial percentage (24%) of trials show over 90% of plural nouns ending in *-s*, there is a curious bimodal distribution in the histogram. 44% of trials show less than 10% of plural forms ending in *-s*.

The breakdown of the representative plural endings for each trial is shown in Table 5. Most of the non-*s* endings are either *-e* (which would correspond to the nominative plurals) or a null ending.

Table 5. Most common plural suffix per trial

Suffix	% trials
-s	30
-∅	26
-e	24
-es	18
-o	2

While the singular case endings in Latin are well defined by the vowel associated with the declension, the plural case endings contain some variability in the vowels. If the neural net fails to associate *-s* with *plural*, it may be unable to make a decision due to the variability and converge on a null phoneme, which, being a string of zeroes, marks the midpoint between all the phonemes. Thus, the null morpheme may disproportionately be the output for plural if the model is indecisive about plural categorization. I leave the question of how to address this model limitation to future work.

5. Discussion

By feeding merely phonology and frequency information as inputs into a neural net, noun class shifts in Romance can be captured in the correct direction. The model supports a merger of the genitive and dative regardless of sound change. It also bolsters the view that sound change was responsible for the merger of the ablative and the accusative. These mergers support an early three-case system that would consist of the *nominative*,[9] *accusative*/*ablative*, and *genitive*/*dative,* which could have been the precursor to both the Romanian two-way *nominative*/*accusative* and *genitive*/*dative* system as well as the early Gallo- and Rhaeto-Romance *nominative* versus *oblique* systems. These two proposed pathways, both supporting Clackson's early accusative/ablative merger, are schematized in Fig-

9 With which the vocative likely merged (except in some Romanian dialects) as well considering the syncretism between it and the nominative.

ures 9 and 10. While most Romance languages collapsed the entire case system, Romanian preserves a case system that reflects an archaic intermediate form.

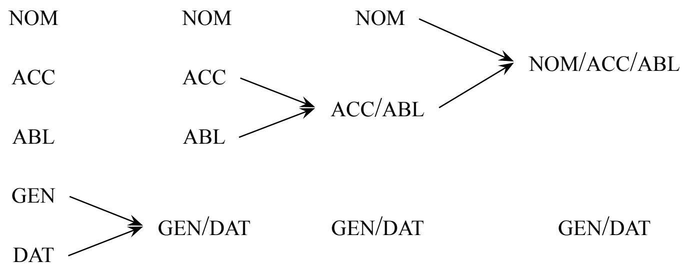


Fig. 9. Proposed case mergers for Daco-Romance

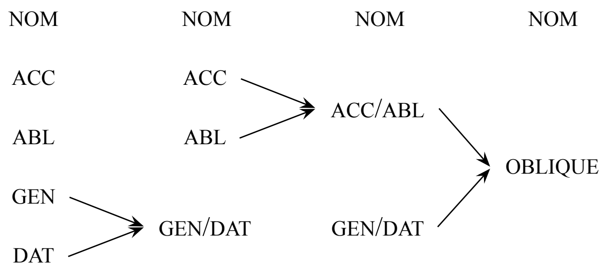


Fig. 10. Proposed case mergers for early Gallo-/Rhaeto-Romance

The complete merger of the nominative case with the accusative/ablative to the exclusion of the genitive/dative also comes about in approximately 40% of trials. Once sound changes have occurred, the nominative singular is differentiated from the accusative/ablative in Declensions II–V only by the *presence* of *-s* in the nominative, whereas in the plural it is differentiated from the accusative/ablative in Declensions I and II only by the *absence* of *-s*. If the model picks up on *-s* as a marker of the plural (which it seems to do often), then the output for the nominative singular should drop *-s* and for the plural should add an *-s*, successfully reproducing reanalysis. If the model stops producing *-s* as a marker of nominative case, it will naturally fall in with the accusative and ablative once sound changes have taken effect.

This same logic may also be applied to the merger of the genitive/dative. The genitive and dative plurals are the least frequent types and so will have little effect. However, the singulars match only in Declension I (*-ae*). In Declensions III–V they differ from each other only by *-s* following sound change. Once again, if *-s* becomes a plural marker, it will cease to be output in the singulars. Once the genitive/dative singulars of four declensions have merged, it is then natural for Declension II to merge its genitive/dative by analogy as well.

The simulation also captures the merger of Declensions II and IV and of the neuter and masculine classes. Both are consequences of the high degree of syncretism between the classes. The model, however, misses the merger of Declension V with I, instead merging it with Declension III. This consistent result suggests that the phonological similarity between Declensions III and V (namely, the stem ending *-e,* which merges with *-i* following sound change) trumps the identification of Declension I as the feminine class.

This study contributes to the growing literature on computational modeling of analogy and corroborates Albright and Hayes (2003) and others in demonstrating that the inclusion of token frequency worsens analogical models, suggesting that highly frequent classes and not individual tokens serve as attractors.

Future modifications of the model could consider having the neural net predict the phonological form of the entire word, a computationally intensive task. The current model assumes that learners are aware that only suffixes are important, but including the whole word would allow for modeling of analogy via association of phonologically common endings in roots. For example, the Romanian reanalysis of the common Declension III neuter ending *-ora* as a feminine plural suffix (> *-uri*) could potentially be modeled. One possible technique to aid the model would be to increase bias towards the end of a word, such that the suffix would still be considered the most relevant part of a word but later phonemes in the root would also be considered and the initial part of a word would be largely ignored. This technique would also remove the assumption that learners can segment a word into its root and suffix during acquisition.

A more realistic model might also include less frequent words in the corpus, such that learning is not biased to those with highest frequency. Furthermore, as the model was greatly simplified in order to tease apart the effects of phonology and frequency, another natural extension would be to add syntactic, semantic, or contact information to assess their roles in morphological change (see Maiden 2011:159–74; Salvi 2011:319–22; Ledgeway 2012:328–35). The current model provides a starting point for these more complicated models.

6. Conclusion

Connectionist modeling is an invaluable tool for simulating morphological change by treating learning of morphological classes as a categorization problem. Even by providing a neural net basic information from the well-known corpus of Latin, the model is able to predict the directionality of noun class mergers using analogy. The results suggest, specifically, that phonology and frequency are largely, but not exclusively, responsible for the morphological changes that occurred in Romance, and more generally, that morphological change may be modeled as a generational learning process with the use of neural nets. The current model also provides a baseline for future work in modeling morphological change in other language families. While Romance languages ubiquitously collapsed case distinctions, Slavic languages have notably retained much of the Indo-European case system. As this study has demonstrated, connectionist modeling of morphological change is an invaluable tool that will be useful in disentangling the differences between the precursors of change in different language families.

Appendix. Word list

aciēs, aditus, adulter, adventus, aedēs, āēr, aes, aestās, aestus, aetās, aevum, affectus, ager, agger, agmen, āla, altum, amīca, amīcus, amnis, amor, anima, animal, animus, annus, aqua, āra, arbiter, arbitrium, arbor, arcus, arēna, argentum, arma, armentum, ars, artifex, artus, arvum, arx, aspectus, astrum, auctor, aura, auris, aurum, auxilium, avis, avus, axis, bellua, bellum, bonum, bōs, brachium, caedēs, caelum, calor, campus, canis, cantus, capillus, caput, carcer, carmen, castra, catēna, cēna, certāmen, cervīx, cibus, cinis, cīvis, clādes, clāmor, classis, cliēns, coetus, cohors, collis, collum, color, coma, comes, concilium, coniūx, cōnsul, convīvium, cōpia, cor, cornū, corōna, corpus, crīmen, crīnis, cruor, cubīle, culpa, cultus, cupīdō, cūra, currus, cursus, custōdia, custōs, damnum, dea, decus, dēlicia, dēns, deus, dictum, diēs, digitus, discrīmen, dīvitiae, dolor, dolus, domina, dominus, domus, dōnum, dux, effigiēs, epulae, eques, equus, error, exemplum, exiguus, exitium, exitus, exsilium, exsul, fābula, faciēs, facinus, factum, fāma, famēs, fastīgium, fātum, favor, fax, fēmina, fera, ferrum, fētus, fidēs, fīlia, fīlius, fīnis, flamma, flētus, flōs, fluctus, flūmen, focus, foedus, fōns, foris, fōrma, formīdo, fortūna, forum, frāter, fraus, fremitus, frēnum, fretum, frīgus, frōns, frons, fructus, frūmentum, frūx, fuga, fulmen, fūnis, fūnus, furor, fūrtum, gaudium, gemitus, gemma, gēns, genū, genus, gladius, glōria, gradus, grātia, grex, habitus, hasta, herba, hērēs, hiems, homō, honor, hōra, hortus, hospes, hostis, humerus, hūmor, humus, ictus, ignis, imāgō, imber, imperium, impetus, incendium, īnfāns, ingenium, iniūria, īnsidia, īnsula, invidia, īra, iter, iūdex, iugum, iūs, iuvenis, iuventa, iuventūs, labor, lacrima, lacus, laetitia, lapis, latēbra, latus, laus, lectus, leō, lēx, liber, lībertās, libīdō, līmen, lingua, lītus, locus, luctus, lūcus, lūdus, lūmen, lūna, lūx, luxuria, magister, malum, manus, mare, marītus, māter, medium, membrum, mēns, mēnsa, mēnsis, mercēs, meritum, merum, metus, mīles, mīlitia, minae, minister, modus, moenia, mōlēs, mōns, mōnstrum, monumentum, mora, morbus, mors, mōs, mōtus, mulier, mundus, mūnus, mūrus, nātūra, nauta, nāvis, nemus, nepōs, nēquitia, nex, nix, nōmen, nota, nox, nūbēs, nūmen, numerus, nūntius, oculus, odium, odor, officium, ōmen, onus, oppidum, ops, opus, ōra, orbis, ōrdō, orīgō, ortus, os, ōs, osculum, ōtium, palma, palūs, parēns, pater, patria, patruus, paupertās, pavor, pāx, pectus, pecūnia, pecus, pellex, penās, perīculum, pēs, pestis, pietās, pignus, pōculum, poena, poēta, pondus, pōns, populus, porta, portus, praeceptum, praeda, praemium, praesidium, praetōr, pretium, prex, prīnceps, prīncipium, proelium, pudor, puella, puer, pugna, pulvis, puppis, querēla, quiēs, radius, rādīx, rāmus, ratiō, ratis, rector, rēgia, rēgīna, regiō, rēgnum, rēmus, rēs, rēx, rīpa, rōbur, rogus, rota, ruīna, rūmor, rūpēs, rūs, sacerdōs, sacrum, saeculum, sagitta, saltus, salūs, sānctus, sanguis, saxum, scelus, sēdēs, seges, sēmen, senectūs, senex, sēnsus, sententia, sepulcrum, sermō, serpēns, servitium, servus, sīdus, signum, silentium, silva, simulācrum, sinus, sitis, sōl, solum, somnus, sonitus, sonus, soror, sors, spatium, speciēs, specus, spēs, spīritus, spolium, spōns, status, stēlla, stirps, strepitus, studium, summa, supplicium, tabula, taurus, tēctum, tēlum, tempestās, templum, tempus, tenebrae, tergum, terra, terror, testis, theātrum, timor, torus, trabs, triumphus, tumultus, tumulus, turba, turris, tympanum, tyrannus, umbra, unda, urbs, uxor, vadum, vallis, vātēs, vēlum, vēna, venēnum, venia, venter, ventus, vēr, verber, verbum, versus, vertex, vērum, vestīgium, vestis, via, vicis, victor, vīlla, vinculum, vīnum, vir, virgō, virtūs, vīs, viscus, vīta, vītis, vitium, voluptās, vōtum, vōx, vulgus, vulnus, vultus.

References

Albright, Adam, and Bruce Hayes. 2003. Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study. *Cognition* 90.119–61.

Alkire, Ti, and Carol Rosen. 2010. *Romance Languages: A Historical Introduction*. Cambridge: Cambridge University Press.

Banniard, Michel. 1992. *Viva voce: Communication écrite et communication orale de IVe au IX siècle en Occident latin.* Paris: Institut des Études Augustiniennes.

Bybee, Joan. 1995. Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10(5).425–55.

———. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Clackson, James, and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language.* West Sussex: Blackwell Publishing.

Daugherty, Kim, and Mark Seidenberg. 1994. Beyond Rules and Exceptions: A Connectionist Modeling Approach to Inflectional Morphology. In Susan D. Lima, Roberta Corrigan, and Gregory K. Iverson (eds.), *The Reality of Linguistic Rules*, 193–218. Amsterdam: John Benjamins.

Delatte, Louis, Étienne Evrard, Suzanne Govaerts, and Joseph Denooz. 1981. *Dictionnaire fréquentiel et index inverse de la langue latine.* Liège: Université Laboratoire d'analyse statistique des langues anciennes.

D'hulst, Yves. 2005. Romance Plurals. *Lingua* 116.1303–29.
Elman, Jeffrey L., et al. 1998. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.
Graur, Alexandru. 1967. *The Romance Character of Romanian*. Bucharest: Publishing House of the Academy of the Socialist Republic of Romania.
Hare, Mary, and Jeffrey L. Elman. 1995. Learning and Morphological Change. *Cognition* 56(1).61–98.
Heaton, Jeff. 2008. *Introduction to NeuralNnetworks with Java*. St. Louis, MO: Heaton Research, Inc.
Herman, Józef. 1967 [2000]. *Vulgar Latin*. Transl. Roger Wright. University Park: Pennsylvania State University Press.
Hořejši, Vladimir. 1964. À propos des noms ambigènes ('neutres') roumains. *Philologica pragensia* 7.401–7.
Ledgeway, Adam. 2011. Syntactic and Morphosyntactic Typology and Change. In Maiden, Smith, and Ledgeway 2011, 382–471.
———. 2012. *From Latin to Romance: Morphosyntactic Typology and Change.* New York: Oxford University Press.
Maiden, Martin. 1996. On the Romance Inflectional Endings -i and -e. *Romance Philology* 50.147–82.
———. 2011. Morphological Persistence. In Maiden, Smith, and Ledgeway 2011, 155–214.
Maiden, Martin, John Charles Smith, and Adam Ledgeway (eds.). 2011. *The Cambridge History of the Romance Languages* I: *Structures*. Cambridge: Cambridge University Press.
Penny, Ralph. 1980. Do Romance Nouns Descend from the Latin Accusative? Preliminaries to a Reassessment of the Noun-Morphology of Romance. *Romance Philology* 33.4.501–9.
Petrucci, Peter. 1999. *Slavic Features in the History of Rumanian*. Munich: Lincom Europa.
Pierrehumbert, Janet. 2001. Stochastic Phonology. *Glot International* 5.6.195–207.
Plunkett, Kim, and Patrick Juola. 1999. A Connectionist Model of English Past Tense and Plural Morphology. *Cognitive Science* 23.463–90.
Polinsky, Maria, and Ezra Van Everbroeck. 2003. Development of Gender Classifications: Modeling the Historical Change from Latin to French. *Language* 79(2).356–90.
Rumelhart, David E., James L. McClelland, and PDP Research Group. 1987. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.
Salvi, Giampaolo. 2011. Morphosyntactic Persistence. In Maiden, Smith, and Ledgeway 2011, 318–81.

Weiss, Michael. 2009. *Outline of the Historical and Comparative Grammar of Latin.* Ann Arbor*:* Beech Stave Press.

Wilkinson, Hugh. 1985. The Latin Neuter Plurals in Romance. *Ronshu* 26.137–50.