DECOUPLING SPEED OF CHANGE AND LONG-TERM PREFERENCE IN LANGUAGE EVOLUTION: INSIGHTS FROM ROMANCE VERB STEM ALTERNATIONS

Chundra A. Cathcart[*1,2], Borja Herce[1,2], and Balthasar Bickel[1,2]

[*]Corresponding Author: chundra.cathcart@uzh.ch
[1]Department of Comparative Language Science, University of Zurich, Switzerland
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Switzerland

Romance verb stem alternations (e.g., Spanish ***teng****o* 'I have' vs. ***tien****es* 'you have') constitute seemingly unnecessary but highly inheritable morphological traits. Using novel phylogenetic methods, we assess the impact of frequency and alternation patterns on properties of their evolution, specifically on the speed of change and the long term preference for pattern types within lemmata. We find credible differences in long-term trends between alternation patterns, and confirm the notion that frequency drives the maintenance of irregular patterns. However, our model reveals no or only weak effects of either predictor on the speed of change. Our findings call for modeling the multiple dimensions of language change jointly but with distinct parameters for speed (or rates) of change and long-term (i.e., stationary) preferences.

1. Introduction

Stem alternations in Romance verb paradigms are well studied (Esher, 2016; Maiden, 2018; Herce, 2019, etc.). Their importance lies in that they constitute a prime example of unmotivated yet systematic patterns (Aronoff, 1994; Bickel, 1994). Unlike morphological alternations that express differences in meaning (e.g. English *drive* PRESENT vs. *drove* PAST) or are phonologically conditioned (e.g., *cat*[s] vs. *dog*[z]), Romance stem alternations do not mark a specific semantic value and are phonologically unpredictable. Despite this lack of motivation, these patterns exhibit remarkable diachronic stability, frequently spreading to new verbs. Here we seek to quantify the evolutionary dynamics of these patterns in an explicit phylogenetic framework, testing effects of frequency and alternation type.

The method we develop allows us to assess the effect of these two predictors on different components of language change, namely speed and stationary probability, interpretable as the long-term preference for a feature. Past research has either focused on speed of change (Pagel, Atkinson, & Meade, 2007; Greenhill et al., 2017) or on biases in change and preferred configurations (Maslova, 2000; Cysouw, 2011; Bickel, 2015; Jäger & Wahle, 2021). Little work to date has considered these properties of change jointly. We provide a principled means of

teasing apart the dynamics of these two facets of language evolution, providing a more flexible understanding of the role of different factors in change.

2. Background

The philological literature identifies three main types of stem alternations in Romance verbs, labeled N, L, and P(YTA); these can co-occur within verbal paradigms. The emergence of the N and L patterns (see Table 1) occurred as a result of sound changes after Classical Latin but (largely) before the break-up of Romance into different languages. Thus, the alternation between *e* and *ie* in the stem of Spanish *perder* is due to different trajectories of change affecting stressed (leading to *ie*) and unstressed (leading to *e*) syllables. The alternation between *g* and *c* (phonetically [g] and [θ], respectively) in the stem of Spanish *decir* results from the palatalization of Latin *c* ([k]) before the vowels *e* and *i*.

Table 1. Partial paradigms of Spanish *perder* 'lose' and *decir* 'say'

	N			L		
	Pres. ind.	Pres. sbjv.	Ipf. ind.	Pres. ind.	Pres. sbjv.	Ipf. ind.
1SG	pierd-o	pierd-a	perd-ía	dig-o	dig-a	dec-ía
2SG	pierd-es	pierd-as	perd-ías	dic-es	dig-as	dec-ías
3SG	pierd-e	pierd-a	perd-ía	dice	dig-a	dec-ía
1PL	perd-emos	perd-amos	perd-íamos	dec-imos	dig-amos	dec-íamos
2PL	perd-éis	perd-áis	perd-íais	dec-ís	dig-áis	dec-íais
3PL	pierd-en	pierd-an	perd-ían	dic-en	dig-an	dec-ían

While the alternations seen in these lexemes are expected developments due to regular sound change (e.g. *digo*, *di*[θ]*es* < *dīc*[k]*ō*, *dīc*[k]*is*), other lexemes have lost their alternations (e.g. *cue*[θ]*o*, *cue*[θ]*es* ← *coquō*, *coquis* 'cook' vs. expected †*cuego*, *cue*[θ]*es*), or have acquired the alternations in analogy with other verbs (e.g. *caigo*, *caes* ← *cadō*, *cadis* 'fall' vs. expected †*cao*, *caes*).

Unlike the other two patterns, the P pattern is inherited from Latin (see Table 2). In Latin, *fēc-* vs *fac-* marked a semantic distinction that is no longer present in modern Romance languages, leading to alternations that are arbitrary from the perspective of meaning. Like the other patterns, however, the P pattern has also been lost in some cases or expanded by analogy in others.

Irregular alternations like these have increasingly been studied in quantitative frameworks and cross-linguistically, exploring the roles of different predictors (e.g., frequency, word length, uniqueness points) in their maintenance (Marzi, Ferro, & Pirrelli, 2019; Sims-Williams, 2021). However, data-driven studies to date have not made use of a phylogenetic framework, which has the potential to shed light on evolutionary pressures that foster and militate against irregularity. In response, we develop a novel phylogenetic model to obtain a detailed understanding of the dynamics of change of these irregular patterns.

Table 2. Partial paradigms of Spanish *hacer*, and Latin *faciō* 'do' illustrating P

	Spanish			Latin		
	Ipf. ind.	Pret. ind.	Ipf. sbjv.	Ipf. ind.	Perf. ind.	Plup. sbjv.
1SG	hac-ía	hic-e	hic-iese	fac-iēbam	fēc-ī	fēc-issem
2SG	hac-ías	hic-iste	hic-ieses	fac-iēbās	fēc-istī	fēc-issēs
3SG	hac-ía	hi[θ]-o	hic-iese	fac-iēbat	fēc-it	fēc-isset
1PL	hac-íamos	hic-imos	hic-iesemos	fac-iēbāmus	fēc-imus	fēc-issēmus
2PL	hac-íais	hic-isteis	hic-ieseis	fac-iēbātis	fēc-istis	fēc-issētis
3PL	hac-ían	hic-ieron	hic-iesen	fac-iēbant	fēc-ērunt	fēc-issent

3. Data

In the Oxford Online Database of Romance Verb Morphology (Maiden, Smith, Cruschina, Hinzelin, & Goldbach, 2010; Beniamine, Maiden, & Round, 2020) each lemma is represented with its paradigm and is coded for its Latin etymon. For each lemma in each variety, we manually assessed whether it contained a reflex of the stem alternation patterns (N, L, and P) presented in Section 2. We gloss over language-specific idiosyncrasies in the inherited distribution of alternants in the paradigm for the purposes of this study (e.g., the L pattern affects the 3PL present indicative of Italian *dire* 'say' but is absent from the 1–2PL present subjunctive, in contrast to its Spanish cognate in Table 1), coding patterns as PRESENT/ABSENT/MISSING DATA. Among Latin-descended lemmata, 231 cognate sets are found in the database. We target items with wide coverage, excluding cognate sets that are limited to dialects of a single language or attested in fewer than five varieties. Additionally, we exclude uninformative lemma-pattern pairs that are completely present or absent across all languages with attested data. In total, we analyze 171 lemma-pattern pairs involving 66 lemmas from 67 varieties.

4. Method[1]

We investigate the historical dynamics of stem alternation using a phylogenetic comparative model. Methods of this sort require a timed phylogenetic representation of the languages in our data set in the form of a Bayesian tree sample, which we infer using RevBayes (Höhna et al., 2016) on the basis of both automatically generated lexical cognacy data (Jäger, 2018) and sound class data indicating which speech sounds are present in each variety (Heggarty et al., 2019). We impose uncontroversial clade constraints on the tree topology along with lower and upper bounds for each clade's date calibration drawn from the literature on Romance languages (Hall, 1974). We use a a Birth-Death tree prior (Yang & Rannala, 1997) and a General Time-Reversible model of character evolution (Tavaré, 1986), along with a relaxed clock with log-normally distributed branch-level rate multipliers and gamma-distributed variation across 4 rate classes. We run 1,000,000

[1]Code available at `https://github.com/chundrac/JCoLE2022-morphomes`

iterations of Markov chain Monte Carlo over 4 chains, thinning the sample to 100 trees after discarding the first half as burn-in. We scale branch lengths so that one unit represents a millennium of change.

Under our model, each binary lemma-pattern pair evolves independently over the phylogeny of Romance according to a continuous-time Markov process. This process is parameterized by a gain and loss rate, or alternatively, the speed of change (irrespective of direction) and the stationary probability of feature presence, interpretable as the long-term preference for a given feature over a phylogeny (irrespective of the speed of change). The gain and loss rate of a lemma-pattern pair with index $d \in \{1, ..., D\}$ are $\pi_d s_d \rho$ and $(1 - \pi_d) s_d \rho$. Here, $s_d \rho$ represents the speed of change for feature d, s_d being a multiplier of the global speed ρ. We place a Uniform$(0, 10)$ prior over ρ, preventing changes from happening more frequently than ten times per millennium. The parameter π_d is the stationary probability of presence for the lemma-pattern pair in question.

While evolutionary parameters in phylogenetic comparative methods are usually estimated without predictors, we model them in a hierarchical distributional regression framework (Bürkner, 2018), allowing both the speed of change and stationary probability for each lemma-pattern pair to vary as a function of multiple predictors. We assume speed multipliers $\boldsymbol{s}$ and stationary probabilities $\boldsymbol{\pi}$ to be normally distributed, with a logit link to keep values within $(0, 1)$:

$$\text{logit } s_d \sim \text{Normal}(\alpha^s + \beta^{s,\text{LEMMA}}_{\text{LEMMAID}_d} + \beta^{s,\text{PATTERN}}_{\text{PATTERNID}_d}, \sigma^s) \tag{1}$$

$$\text{logit } \pi_d \sim \text{Normal}(\alpha^\pi + \beta^{\pi,\text{LEMMA}}_{\text{LEMMAID}_d} + \beta^{\pi,\text{PATTERN}}_{\text{PATTERNID}_d}, \sigma^\pi) \tag{2}$$

In each sampling statement, α denotes the intercept, β^{LEMMA} represents the contribution of each lemma type, and β^{PATTERN} represents the contribution of each alternation type. The standard deviations of these distributions represent the variance in speed and stationary probability that are not explained by the predictors included. We model the contribution of lemma type as a monotonic function of each lemma's frequency in Latin texts (Tombeur, 1998); this involves a combination of a parameter representing the effect of moving from the lowest to the highest frequency and a simplex parameter representing the effect of moving along the cline of frequency (Bürkner & Charpentier, 2020). We treat pattern type as two dummy-coded factors, comparing the levels L and P to N, respectively. We place Normal$(0, 1)$ priors over all model parameters in statements (1–2) with the exception of simplex parameters and standard deviations σ, which receive Dirichlet$(1, ..., 1)$ and HalfNormal$(0, 1)$ priors, respectively. Posterior distributions for parameters are inferred using the R package CmdStanR (Gabry & Češnovar, 2021). In line with the most conservative criteria for hypothesis evaluation, we infer decisive evidence for the effect of a predictor if the 95% credible interval (CI) of the corresponding parameter excludes zero, and strong evidence in cases where the 95% CI overlaps with zero but the 85% CI does not.

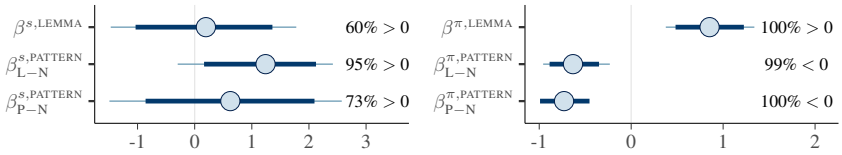

Figure 1. Medians and posterior 95% and 85% (shaded) credible intervals for model parameters of interest on the logit scale, along with percentage of samples above or below 0. PATTERN refers to the difference between N, L and P alternations; LEMMA refers to Latin lemma frequency.

5. Results

Posterior distributions for model parameters are given in Figure 1. In general, the predictors in our model do not appear to have a decisive effect on features' speed of change (s; left panel), though there is strong but not decisive evidence that the L pattern shows a higher speed of change than the N pattern. In contrast, we see decisive evidence for effects of all predictors on long-term pattern preference, i.e. the stationary probabilities π (right panel). The 95% credible interval for the parameter $\beta^{\pi,\text{LEMMA}}$ is positive, indicating that more frequent lemmata are more likely to display a stem alternation pattern and less frequent lemmata are likelier to regularize. Additionally, we see decisive evidence that the long-term preference for N is greater than that of L and P, though post-hoc inspection of model parameters showed no evidence for a L vs. P contrast.

6. Discussion

Our results show that the N pattern outranks L and P in long-term preference (stationary probability) but not in stability (speed of change). A possible explanation might be that that the N pattern affects the 3SG present indicative, which tends to be highly frequent, whereas the L and P patterns do not.

More generally, we confirm the role of lexical frequency in the maintenance of irregular patterns. Arbitrary stem alternation patterns fly in the face of communicative and acquisitional efficiency in that they introduce more than one form per meaning. At the same time, irregularity enhances discriminability between forms in a paradigm, especially in more frequent lexemes (Nübling, 2011; Blevins, Milin, & Ramscar, 2017). Our results support the idea that more frequent lexemes are more likely to preserve and acquire irregular morphology over time.

The frequency effect is limited to stationary probabilties (long-term preferences), however, and does not affect how fast languages move to the preferred state (cf. $\beta^{\pi,\text{LEMMA}}$ vs $\beta^{s,\text{LEMMA}}$ in Fig. 1). This finding is consistent with the notion that cognitive and communicative pressures (such as frequency effects) — or indeed "language universals" more generally — bear primarily on stationary probabilities, and not on stability (Maslova, 2000; Cysouw, 2011; Bickel, 2015).

Our approach decouples these two aspects of evolution and facilitates explicit

assessment and comparison of them in a single model. This flexibility is not generally found in recent work in evolutionary linguistics which tends to focus on variation in speed. We suspect that this emphasis stems from two sources: first, most prominent biological models of rate variation focus solely on speed variation (Huelsenbeck & Suchard, 2007; Heath, Holder, & Huelsenbeck, 2011), with few exceptions (e.g., Lartillot & Poujol, 2011). Second, a large body of work in language evolution revolves around change in basic vocabulary, where the replacement of one word by another cannot necessarily be construed as incurring a communicative cost in the same way that the development of unpredictable allomorphy might (though see Martin, 2007).

7. Conclusion

We investigated the development of irregular patterns in Romance verbal morphology using a novel method that unites phylogenetic modeling with hierarchical distributional regression modeling. Our results in some ways confirm received wisdom regarding the role of frequency in the maintenance of irregularity, but also shed light on poorly understood issues in language change, showing that frequency largely does not explain variation in speed of change but only in long-term preferences (stationary probabilities). One possible explanation for this is that speed of change responds to social pressure for differentiation such as schismogenesis, which is most strongly associated with vocabulary (Greenhill et al., 2017). Stem alternation and irregularity might be less accessible as markers of social differentiation and are therefore relatively immune to differences in speed of change. By contrast, they have direct implications for efficiency in processing, learning and communication, with effects on long-term preferences.

The model presented here can be expanded in a number of ways. It is straightforward to build in branch-level variation for speed and stationary probability, which can help to identify events of drastic change coinciding with language contact, schismogenesis, or other changes in the linguistic system. Additionally, our model used Latin lemma frequency as a proxy for an etymon's frequency throughout Romance history, a simplification that does not fully capture the dynamics of vocabulary change. In theory, it is possible to treat relative frequency as a continuous trait that varies over the tree (see Ringen, Martin, & Jaeggi, 2021 for flexible, complex models of the co-evolution of continuous and discrete cultural traits). A next step will involve building additional predictors and interaction terms into the model to investigate among other things whether phonologically and semantically similar lemmata undergo similar patterns of change, and whether or not other variables, such as conjugation class, play an interpretable role.

Acknowledgements

This work was supported by the NCCR Evolving Language, Swiss National Science Foundation Agreement Nr. 51NF40_180888.

References

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes.* Cambridge, MA: MIT Press.

Beniamine, S., Maiden, M., & Round, E. (2020). Opening the Romance Verbal Inflection Dataset 2.0: a CLDF lexicon. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3027–3035).

Bickel, B. (1994). In the vestibule of meaning: Transitivity inversion as a morphological phenomenon. *Studies in Language*, *19*, 73–127.

Bickel, B. (2015). Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis, 2nd edition* (pp. 901–923). Oxford: Oxford University Press.

Blevins, J. P., Milin, P., & Ramscar, M. (2017). The Zipfian paradigm cell filling problem. In *Perspectives on morphological organization* (pp. 139–158). Brill.

Bürkner, P.-C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420–451.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Cysouw, M. (2011). Understanding transition probabilities. *Linguistic Typology*, *15*, 415–431.

Esher, L. (2016). Morphomes and predictability in the history of Romance perfects. *Diachronica*, *32*(4), 494–529.

Gabry, J., & Češnovar, R. (2021). *Cmdstanr: R interface to 'CmdStan'*. (https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org)

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, *114*(42), E8822–E8829.

Hall, R. A. (1974). *Comparative Romance grammar.* New York/London/ Amsterdam: Elsevier.

Heath, T., Holder, M., & Huelsenbeck, J. (2011). A Dirichlet Process Prior for estimating lineage-specific substitution rates. *Molecular biology and evolution*, *29*, 939-55.

Heggarty, P., Shimelman, A., Abete, G., Anderson, C., Sadowsky, S., Paschen, L., Maguire, W., Jocz, L., Aninao, M. J., Wägerle, L., Dërmaku-Appelganz, D., Silva, A. P. d. C. e., Lawyer, L. C., Michalsky, J., Cabral, A. S. A. C., Walworth, M., Koile, E., Runge, J., & Bibiko, H.-J. (2019). *Sound Comparisons: Exploring Diversity in Phonetics across Language Families.*

Herce, B. (2019). Morphome interactions. *Morphology*, *29*(1).

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore,

B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, *65*(4), 726–736.

Huelsenbeck, J. P., & Suchard, M. A. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology*, *56*(6), 975–987.

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Reports*, *5*.

Jäger, G., & Wahle, J. (2021). Phylogenetic typology. *Front. Psychol.*, *12*.

Lartillot, N., & Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution*, *28*(1), 729–744.

Maiden, M. (2018). *The Romance verb: Morphomic structure and diachrony.* Oxford University Press.

Maiden, M., Smith, J. C., Cruschina, S., Hinzelin, M.-O., & Goldbach, M. (2010). *Oxford Online Database of Romance Verb Morphology.* University of Oxford. Online at http://romverbmorph.clp.ox.ac.uk/.

Martin, A. T. (2007). *The evolving lexicon.* Unpublished doctoral dissertation, University of California, Los Angeles.

Marzi, C., Ferro, M., & Pirrelli, V. (2019). A processing-oriented investigation of inflectional complexity. *Front. Comm.*, *4*, 1–23.

Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology*, *4*, 307-333.

Nübling, D. (2011). How do exceptions arise? On different paths to morphological irregularity. In H. Simon & H. Wiese (Eds.), *Expecting the unexpected: exceptions in grammar* (p. 139-162). Berlin: de Gruyter.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 717-720.

Ringen, E., Martin, J. S., & Jaeggi, A. (2021). Novel phylogenetic methods reveal that resource-use intensification drives the evolution of "complex" societies. *EcoEvoRxiv.*

Sims-Williams, H. (2021). Token frequency as a determinant of morphological change. *Journal of Linguistics*, 1–37.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*(2), 57–86.

Tombeur, P. (1998). *Thesaurus formarum totius Latinitatis a Plauto usque ad saeculum XXum.* Turnhout: Brepols.

Yang, Z., & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular biology and evolution*, *14*(7), 717–724.