Borja Herce*

# Quantifying the importance of morphomic structure, semantic values, and frequency of use in Romance stem alternations

https://doi.org/10.1515/lingvan-2022-0028
Received March 5, 2022; accepted July 28, 2022; published online October 19, 2022

**Abstract:** Stem alternations in Romance have recently been argued to be regulated largely by autonomously morphological (aka morphomic) organizational principles. Here, I assess the relative contribution of morphomic structures vis à vis alternative principles, namely semantic structure, and token frequency. Results confirm the exceptional importance of autonomously morphological domains on Romance verb stem alternations; however, inherent inflectional values and token frequency also play a decisive role in the overall stem-morphological similarity of different paradigm cells.

**Keywords:** autonomous morphology; frequency; morphomes; paradigm; semantics

## 1 Introduction

Morphological paradigms constitute complex grammatical objects whose organizational principles are controversial. Some theoretical frameworks like Distributed Morphology (Halle and Marantz 1994) do not allocate any ontological status to paradigms as such, and consider them epiphenomenal, since what really matters is how smaller morphological units (aka. 'morphemes': Bloomfield 1926; Bolinger 1948; Embick 2015) are licensed to express particular features and values. In other frameworks like the Word and Paradigm approach (Blevins 2016; Matthews 1965), paradigms are central to morphological architecture, as are word-to-word similarities and oppositions that allow speakers to produce all inflected forms, usually on the basis of an incomplete input (consider the Paradigm Cell-Filling Problem of Ackerman et al. 2009).

Research on the paradigm as an empirically accessible object has become more popular over the last decade, as tools and metrics from Set Theory (Stump and Finkel 2013) and Information Theory (Ackerman and Malouf 2013) have been adopted to explore different measures of complexity, and to assess quantitatively and objectively the predictability of some paradigm cells from others. In the domain of Romance stem alternations, a quantitative (consider research on stem-spaces by Boyé and Cabredo-Hofherr 2006; Montermini and Bonami 2013, etc.) and a more philological and qualitative tradition (Esher 2015; Herce 2020a; Maiden 1992, 2018, etc.) have explored the synchronic and diachronic properties of paradigmatic structures in more detail than in any other family. A view has emerged in these circles that Romance stem alternation patterns in particular, and maybe even paradigmatic structures more generally, are essentially (autonomously) morphological structures. Carstairs-McCarthy (2010: 210), for example, believes that the importance of features in morphological evolution "has been overrated". Similarly, Maiden (2016: 49) argues that, based on Romance stem alternations and change, morphomic patterns are not dispreferred. Blevins (forthcoming) goes as far as to say that "the contrast between 'natural' and 'unnatural' classes appears to reflect a priori assumptions about descriptive 'economy' and 'naturalness' which have never been shown to be relevant to language structure, acquisition or use".

Although experimental (Herce et al. forthcoming; Saldana et al. 2022) and typological literature (Cysouw 2009) would seem to argue quite forcefully in favour of the relevance of naturalness in language structure and

***Corresponding author: Borja Herce**, University of Zurich, Zurich, Switzerland, E-mail: borjaherce@gmail.com

acquisition in general, it might still be the case that in concrete families and systems (for example Romance stem alternations in verbal inflection) other principles take the upper hand. Despite the abundance of broad (and hard to test) claims in this respect, the precise weight of morphomic organisational principles in the paradigmatic structure of Romance or other families and languages is not known because it has not been subject to a dedicated empirical investigation.

In this paper, I attempt to do precisely this: quantify in a statistically responsible way the relative importance of morphomic domains, semantic structure and token frequency, in predicting the stem alternation patterns found in verbal paradigms across the family. Section 2 provides the necessary background on Romance stem alternation and morphological paradigmatic predictability structures in the family. Section 3 presents the data used for the present investigation and shows how an explicit statistical model might shed light on the relative weight of different explanatory principles. Section 4 discusses the results and their implications and limitations, and Section 5 summarises the paper and its conclusions, and presents ideas for future research.

## 2 Morphomes and stem alternations in Romance verbal inflection

Romance verbal inflection expresses the contextual-inflectional values of person (1, 2, 3) and number (SG, PL) of the subject, various TAM categories (between 4 and 9, depending on the language), as well as a small number of nonfinite forms.

The Portuguese paradigm of 'give' in Table 1 illustrates the inflectional categories that Romance languages maximally inherited from Latin, to which we would need to add the 2SG imperative *dá*, 2PL imperative *dai*, infinitive *dar*, gerund *dando*, and participle *dado*.

Because the semantic description of imperative and nonfinite forms into different TAM values is not straightforward (or impossible) and because person and number values do not apply here as in finite forms, imperatives and nonfinite forms will be excluded from analysis in the rest of this paper. Due to the choice to focus on forms inherited from Latin, the same applies to the future (*darei, darás, dará* …) and conditional forms (*daria, darias, daria* …), which emerged from periphrastic constructions only in Western Romance.

Although the paradigm in Table 1 does not show stem alternations (that is, it has a stem *d-* everywhere), many other Portuguese and Romance verbs do. The most prominent and widespread stem alternation patterns have been named (N, L, PYTA) and discussed quite extensively over the last decades (see Maiden 2018 for an extensive summary). In the Romance paradigm, these alternants have the distribution illustrated in Table 2.

Each of these alternation patterns goes back to morphology in the ancestral language that would have been inherited by Romance varieties. PYTA is the oldest of all, and was already present in Classical Latin, where many verbs showed alternations between imperfective and perfective stem (for example *fak-* vs.

**Table 1:** TAM and person-number categories and forms in Portuguese *dar* 'give'.[a]

| | PRS.IND | PRS.SBJV | IMP.IND | PRT.IND | PLUP.IND | PLUP.SBJV | FUT.SBJV |
|---|---|---|---|---|---|---|---|
| 1SG | *dou* | *dê* | *dava* | *dei* | *dera* | *desse* | *der* |
| 2SG | *dás* | *dês* | *davas* | *deste* | *deras* | *desses* | *deres* |
| 3SG | *dá* | *dê* | *dava* | *deu* | *dera* | *desse* | *der* |
| 1PL | *damos* | *demos* | *dávamos* | *demos* | *déramos* | *déssemos* | *dermos* |
| 2PL | *dais* | *deis* | *dáveis* | *destes* | *déreis* | *désseis* | *derdes* |
| 3PL | *dão* | *deem* | *davam* | *deram* | *deram* | *dessem* | *derem* |

[a]Individual Romance varieties may have additional syncretisms or may have lost some of these TAMs. In addition, future and conditional tenses are widespread across Western-Romance but were not inherited from Classical Latin, as they grammaticalized from verbal periphrases involving the infinitive.

**Table 2:** Distribution of N, L, and PYTA stem alternants in the Romance paradigm.

| | PRS.IND | PRS.SBJV | IMP.IND | PRT.IND | PLUP.IND | PLP.SBJV | FUT.SBJV |
|---|---|---|---|---|---|---|---|
| 1SG | N/L | N/L | | PYTA | PYTA | PYTA | PYTA |
| 2SG | N | N/L | | PYTA | PYTA | PYTA | PYTA |
| 3SG | N | N/L | | PYTA | PYTA | PYTA | PYTA |
| 1PL | | L | | PYTA | PYTA | PYTA | PYTA |
| 2PL | | L | | PYTA | PYTA | PYTA | PYTA |
| 3PL | N | N/L | | PYTA | PYTA | PYTA | PYTA |

*fe:k-* 'do', *po:n- posw-* 'put', *fer- tul-* 'carry', etc.). Despite the fact that these tenses are no longer all perfective, these alternations were often inherited by the daughter languages (for example Portuguese *faz-* vs. *fiz-*, Spanish *hac-* vs. *hic-* 'do', Italian *fac-* vs. *fec-*, etc.) and occasionally analogically innovated.

L alternations emerged later as a result of sound changes involving consonant palatalizations (of coronals before /j/, and of velars before /i/ and /e/). Thus, Latin *dī[k]ō* 'say.1SG.PRS' *dī[k]is* '2SG.PRS.IND' for example, become Portuguese *digo dizes*, Italian *di[k]o di[tʃ]i*, Romanian *zi[k] zi[tʃ]i*, etc. At the same time, many L-shaped (that is, 1SG.PRS.IND+PRS.SBJV) alternations in modern Romance varieties must be analogical (for example, *cadō* 'fall.1SG.PRS.IND' *cadis* '2SG.PRS.IND' would not be expected to alternate but does in many varieties like Spanish *caigo caes*), which suggests that the domain and/or the alternation pattern must have been acquired as a (semi)productive grammatical unit (that is, as a 'morphome') by language users at some point (but see Nevins et al. 2015).

N alternations emerged somewhat later still, as a result of sound changes that created divergences between stressed and unstressed vowels. Those cells where stems were stressed (SG+3PL present) preserved a greater number of phonological distinctions, and often also underwent diphthongizations in a way that unstressed vowels did not (for example Latin /ˈk**o**mputo:/ 'calculate.1SG.PRS' > Spanish /ˈk**we**nto/, while Latin /k**o**mpuˈta:mus/ 'calculate.1PL.PRS' > Sp. /k**o**nˈtamos/). A number of other alternations, for example suppletive ones like Italian *vado* 'go.1SG.PRS' versus *andiamo* 'go.1PL.PRS' must have emerged in analogy to the ones generated by sound change.

It is analogical morphological changes like these, which respect the inherited domains of stem allomorphy, that have fuelled the notion of the 'morphome' and the claims that paradigms can have autonomously morphological structures and categories that do not correspond to semantic or syntactic natural classes. However, and although domains like SG+3PL.PRS, or 1SG.PRS.IND+PRS.SBJV are certainly not well-defined values as traditionally conceived, there is still a measure of semantic similarity among the cells involved. Stem alternants are not haphazardly distributed across semantic values, as for example, both N and L involve present tense cells exclusively, even if not all of them.

Upon further scrutiny, Romance stem alternations also abide by general typological tendencies such as the seeming greater relevance (in need of statistical quantitative confirmation here) of inherent inflection (that is, TAM) relative to contextual inflection (that is, person and number) (see Booij 1996; Bybee 1985: 57). Thus, even if cases of suppletion based on person agreement do exist (see for example Corbett 2007: 20–23) stem allomorphy is found to be cross-linguistically much more sensitive to inherent inflectional categories, which are also more relevant semantically to lexical meaning, and also tend to be expressed closer to the stem than contextual inflection.

Romance verb stem alternations also match other general trends, such as the horizontal homophony hierarchy of Cysouw (2009: 300), which observes that, in line with the semantic transparency of plural number in different persons (associative 1PL: 1 + 2, 1 + 3, 2PL: 2 + 3 vs. cumulative: 3PL: 3 + 3), number distinctions/morphology are cross-linguistically less prominent in 3 than in 2, and less common in 2 than in 1. A look at Romance morphomes (Table 3) reveals that they also respect this semantically motivated hierarchy by which morphological neutralizations (in stems or affixes) are most common in 3.

**Table 3:** Morphological systems with number distinction in 1 and 2 but no distinction in 3.

| | PRS.IND Romance | | PRS.SJV Romance | | Ecuadorian Quechua | | Chickasaw | |
|---|---|---|---|---|---|---|---|---|
| | SG | PL | SG | PL | SG | PL | SG | PL |
| 1 | N/L | | N/L | L | -ni | -nchik | sa- | po- |
| 2 | N | | N/L | L | -ngi | -gichik | chi- | hachchi- |
| 3 | N | N | N/L | N/L | -n | -n | ∅- | ∅- |

What we are missing, thus, in order to assess just how important different factors and structural principles are in regulating stem alternations in the family, is a statistical analysis based upon extensive quantitative data. Fortunately, thanks to decades of research into Romance synchrony and diachrony, and also thanks to the extensive documentation of the ancestral language Latin, this data exists and is readily available. Section 3 will elaborate on the data that this research relies on, and on how we can best operationalize semantic and morphomic structure, as well as token frequency.

# 3 Data coding and variables

The *Oxford Database of Romance Verb Morphology* (Maiden et al. 2010) constitutes the key resource on which the present investigation relies. It contains (mostly) complete paradigms in phonological form of 73 Romance varieties. Of these, 57 (see Figure 1) were documented well enough for a relatively complete picture of stem alternation to be obtained. The chosen threshold was having at least 15 lexemes with complete paradigms. The paradigmatic distribution of the stem alternations that occurred in all these verbs[1] were manually coded. The result was a database of the paradigmatic distribution of 2,151 stem alternation patterns, 212 of them unique in their paradigmatic extension. The number of inspected lexemes and the number of paradigmatically different stem alternation patterns found per variety is displayed in Figure 1.

The way alternation patterns were encoded relied on identifying segments which, within the stem, are not shared throughout all word forms in the paradigm. Only morphological alternations were considered as far as possible, thus ignoring automatic phonological operations such as word-final devoicings, trivial vowel reductions, etc.[2] Data collection proceeded by coding presence (1) versus absence (0) of those alternating segments across all paradigm cells. An example is given in Table 4. Missing data, usually in tenses that have become extinct in individual varieties, but also occasionally in undocumented or missing (that is, defective) forms, were coded as *NA*.

With this information we can assess quantitatively how often are stems the same or different between all possible pairs of cells in the paradigm of verbs with stem alternations. The most stem-different cells in the paradigm were found to be 3SG.PRET.IND and 3SG.PRS.SBJV, which were found to be distinct in 82.1% of verbs with stem alternations. On the opposite side, many cells were found to always share their stem (for example all PLUP.IND and PLUP.SBJV cells). The complete dissimilarities are provided in the form of a distance matrix in the supplementary materials.

**1** Reflexes of the verb *sum* 'be' were excluded because of the indeterminacy of segmentation in this extremely irregular verb.

**2** Examples include: the fact that /a/ changes to /ɐ/ in Portuguese in unstressed environments (for example /ˈfaz-iʃ/ and /fɐˈz-emuʃ/), the fact that /g/ changes to /k/ word-finally in Catalan (for example /ˈbeɣ-a/ and /ˈbek-∅/), etc. Which operations are automatic (that is, synchronically active, required, and predictable as part of a language's phonological and phonotactic system) is not foolproof, of course, as the borders between phonology and morphology are blurry. Individual coding decisions can be consulted in the supplementary information to this paper.

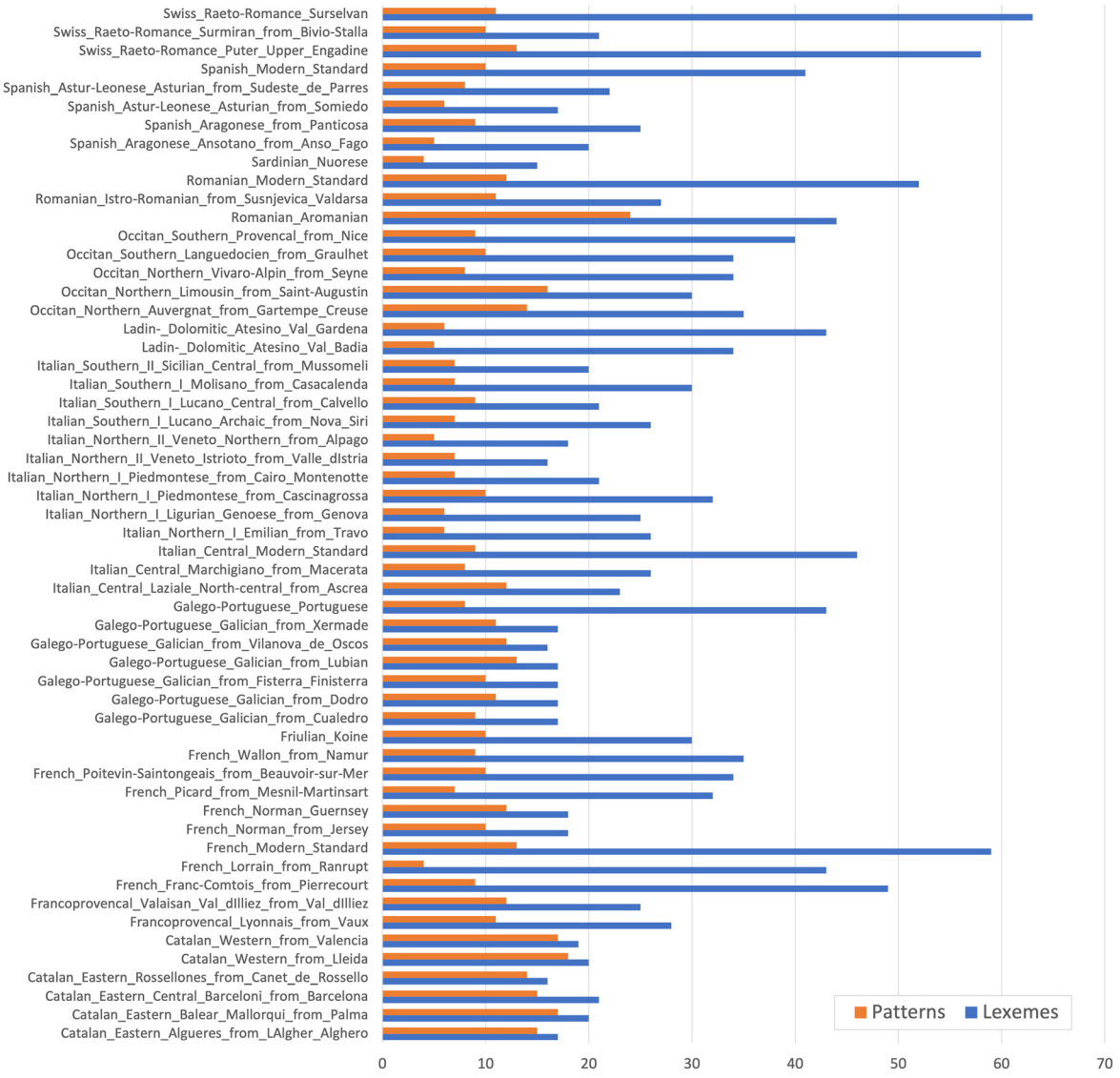


**Figure 1:** Number of lexemes and number of unique alternation patterns per variety.

**Table 4:** Example of the coding of the paradigmatic distribution of alternations.

| | 1SG.PRS.IND ˈposu | 2SG.PRS.IND ˈpɔdiʃ | 3SG.PRS.IND ˈpɔdi | 1PL.PRS.IND puˈdemuʃ | 2PL.PRS.IND puˈdejʃ | 3PL.PRS.IND ˈpɔdẽĩ |
|---|---|---|---|---|---|---|
| /s/[a] | 1 | 0 | 0 | 0 | 0 | 0 |
| /ɔ/ | 0 | 1 | 1 | 0 | 0 | 1 |
| /o/ | 1 | 0 | 0 | 0 | 0 | 0 |

[a]Because being coded as 1 or 0 is irrelevant (that is, it is only having the same or a different number that counts), having a line 10000 for /s/ and also a line 01111 for /d/ would be redundant in that the same alternation pattern would be counted twice.

To assess how important morphomic structure, semantic/syntactic structure and frequency are, we need to assess how good they are as predictors of these relative cell-cell stem (dis)similarities. To do this, we need to operationalize these in a practical way. Some of these operationalizations are trivial. With regards to contextual inflectional values, the Romance paradigm cells in Table 2 can be classified into first (1), second (2),

and third (3) person, and singular (SG) and plural (PL) number. We can hence incorporate these values into a statistical model to see if/how well they predict the stem-similarity of any two cells: are for example first person cells stem-morphologically more similar to other first person cells than to second or third person cells? Morphomic structure can also be operationalized with relative ease, with the classification of paradigm cells, as per Table 2, into N/L/∅, N/∅/∅, ∅/L/∅, ∅/∅/P and ∅/∅/∅ cells. Pairs of cells can thus be ranked for their relative morphomic dissimilarity: 0 (if they belong to the same morphomic category, like 1SG.PRS.IND and 2SG.PRS.SBJV, both N/L/∅ cells) to 3 (if they differ on every morphomic affiliation, like 1SG.PRS.IND and 2SG.PRET.IND, N/L/∅ and ∅/∅/P cells respectively).

Other factors are more problematic and subject to different potential operationalizations, which would lead to somewhat different results. With regards to cell frequency, and in the absence of detailed corpora of most of the documented varieties, I had to resort to the frequency of the cells in the attested Latin corpus (as registered in Delatte et al. 1981, see Table 5). Although the frequency of the reflex cells will certainly differ in the Romance daughter languages (most markedly across inherent TAM inflectional categories), the frequency in Latin will be used here as a proxy for the frequency of cells across the family.

Latin and Romance daughter languages' cell frequencies are highly correlated (for example 0.907 Correlation Coefficient with the Spanish frequencies in CORPES XXI [subcorpus from Spain]), which is why the use of Latin frequencies is appropriate here. This correlation between Latin and Romance cell frequencies is expected from i) the fact that the range of uses of different inflectional values constitute inheritable grammatical properties, and ii) from the fact that a degree of universality must exist regarding what people tend to talk about the most. The combined frequency of a pair of cells (for example 3,270 tokens of IPV.IND.1PL+506,517 tokens of PRS.IND.2SG) will be used to predict how stem-morphologically (dis) similar these cells are. As per Bybee (for example 2006) and others, the morphological autonomy of a cell (that is, the extent to which it can have idiosyncratic traits) depends, among other things, on its token frequency. Thus, for a given pair of cells (A vs. B), the chance of having a different stem in cells A and B is expected to be higher the higher the token frequency of each individual cell. For this reason, operationalizing this predictor as combined frequency (rather than for example as the difference in frequency) was deemed the most sensible option.

More challenging still is the operationalization of inherent inflectional structure (that is, TAM) into sensible predictor variables. Unlike contextual inflection, which is structured into relatively uncontroversial and orthogonal features and values, TAM categories are, in Romance and many other languages, much messier. Different analyses abound and the number of semantic dimensions along with different TAMs are structured (tense and aspect in particular) is more than just three (see Coseriu 1976 for a detailed summary). The functional specialization of the Romance TAM categories is least controversial with respect to mood, with classification into indicative (IND) and subjunctive (SBJV) values as indicated by the labels in Tables 1 and 2. With regard to tense, the division into present (PRS) versus non-present tenses is the least controversial (note that future and conditional tenses are not analysed). Aspect is the most problematic category, with many (maybe most) forms being aspectually neutral. Because of this, a specific classification into different aspectual values will not be provided.

**Table 5:** Cell frequencies of the Latin verbal paradigm (according to Delatte et al. 1981).

| | PRS.IND | PRS.SBJV | IMP.IND | PRT.IND | PLUP.IND | PLUP.SBJV | FUT.SBJV[a] |
|---|---|---|---|---|---|---|---|
| 1SG | 739,362 | 335,837 | 10,150 | 203,068 | 4,882 | 3,747 | 9,744 + 7,280 |
| 2SG | 506,517 | 204,562 | 5,685 | 39,150 | 3,032 | 2,277 | 29,119 |
| 3SG | 2,899,946 | 607,274 | 207,802 | 912,185 | 83,290 | 50,071 | 170,915 |
| 1PL | 184,043 | 86,034 | 3,270 | 71,202 | 1,565 | 1,018 | 7,484 |
| 2PL | 103,439 | 24,985 | 1,426 | 14,207 | 470 | 432 | 7,855 |
| 3PL | 725,044 | 187,805 | 91,708 | 164,664 | 27,462 | 13,943 | 41,372 |

[a]As the FUT.SBJV tense is generally considered to result from the merger of two different Latin tenses (future perfect and perfect subjunctive, which were only morphologically distinct in the 1SG), their combined frequency has been considered.

# 4 Statistical analysis and results

To assess the relative importance of inherent and contextual inflectional semantic values, morphomic categories, and frequency of use on the distribution of stem alternations in the paradigm, I fit a linear regression model (function lm() in R, R Core Team 2014; R Studio Team 2020) with the proportion of alternation patterns in which every pair of paradigm cells has a different stem as the predicted variable, and with inherent-inflectional similarity, contextual-inflectional similarity, and morphomic similarity of the cells as predictors, along with combined cell frequency:

$$im(Stem\_distance \sim Contextual\_infl + inherent\_infl + Frequency + Morphomes$$

Where:

'Stem_distance' is the proportion of alternations where a given pair of paradigm cells has a different stem (i.e. a '1' vs. '0' as per the coding in Table 4). For the complete list of 861 distances see the appendix.

'Contextual_infl' is a measure of contextual inflectional similarity of two cells. It ranges between 0 (no shared values, e.g. 1SG vs. 3PL) and 2 (both values shared, e.g. 1SG vs. 1SG).

'Inherent_infl' is a measure of the inherent inflectional similarity of two cells. It ranges between 0 (no shared values, e.g. PRS.SBJV vs. IPF.IND) and 3 (all shared values, e.g. IPF.IND vs. IPF.IND).

'Frequency' is the combined token frequency of the pair of cells in Latin as per Delatte et al. (1981).

'Morphomes' is a measure of the morphomic similarity of two cells, ranging between 0 and 3 as explained above.

The results are summarized in Table 6 and Figure 2. Table 6 reports the results for each predictor variable of the above model (Adjusted *R*-squared 0.9584). Figure 2, in turn, displays all datapoints; that is, the 861 possible cell pairs in the surveyed Romance paradigm (see Table 1) classified for every variable. They show a statistically highly significant (***) effect upon Romance verb stem alternations of i) morphomic structure, ii) frequency, and iii) inherent inflectional structure, but no significant effect of contextual inflectional structure (that is of person and number).

These results confirm the received wisdom that stem alternation patterns are much more sensitive to inherent than to contextual (that is, agreement) inflectional semantic structure. While sharing (more) TAM values makes cells more likely to also share a stem, sharing person and number values seems to have little effect overall.

The effect of frequency is also significant (larger than that of inherent inflection) and goes in the direction expected from the literature (consider Bybee's 2006 notion of 'autonomy'). A higher token frequency makes cells more autonomous and hence less likely to share a stem with other cells. Due to the well-established link between frequency and irregularity (Herce 2016; Pinker 1999; Wu et al. 2019; Zipf 1935), higher frequency cells (also lexemes) have a tendency to accumulate a greater degree of general idiosyncrasy than infrequent cells.

**Table 6:** Results of the linear regression model.[a]

| | Estimate | Std. error | *t* Value | Pr (>\|*t*\|) |
|---|---|---|---|---|
| (Intercept) | 7.64E-01 | 5.09E-03 | 150.171 | 2E-16*** |
| Contextual_infl | −7.72E-04 | 2.82E-03 | −0.274 | 0.784 |
| Inherent_infl | −2.04E-02 | 2.52E-03 | −8.109 | 1.76E-15*** |
| Frequency | 4.71E-08 | 2.93E-09 | 16.06 | 2E-16*** |
| Morphomes | −2.32E-01 | 2.22E-03 | −104.05 | 2E-16*** |

[a]The results do not differ significantly if each of the predictors is run in a separate model: Contextual_infl is deemed non-significant (*R* squared −0.004685), while the other variables are highly significant (Inherent_infl *R*-squared 0.3458, Frequency *R*-squared 0.1278, Morphomes *R*-squared 0.9432).

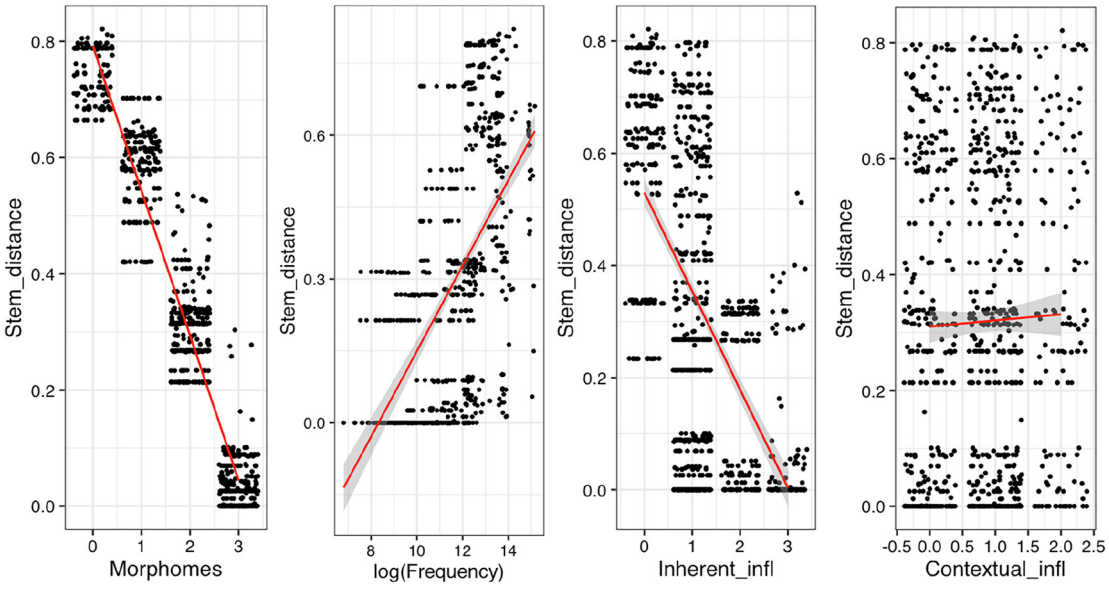


**Figure 2:** Correlation of the predicted and predictor variables.

Note in this respect that stem alternation is an irregular trait in Romance, with all N, L, and P occurring generally in under 5% of the verbal lexicon.

The most robust predictor of Romance stem alternation patterns according to this analysis is morphomic structure. The history of Romance and its accidents (that is, sound changes), and the (un)predictability relations these gave rise to (see Section 2) are, still in contemporary Romance varieties, the most significant predictor of stem alternation patterns across the family, with cells from within the same morphomic domain (as identified in Table 2) much more likely to share a stem.

These overall results speak, thus, clearly in favour of a scale *morphomes* > *frequency* > *inherent inflection* > *contextual inflection* according to the factors which most decisively drive stem alternation patterns in contemporary Romance. They support, thus, in a quantitative, rather than qualitative way, the extant opinion in the Autonomous Morphology literature that morphomic structures are the most important structural principle in Romance verb stem alternations. Results additionally reveal, however, that frequency of use, and inherent inflectional semantic structure are also highly significant predictors that should not be ignored. Any complete account of Romance stem alternation patterns, thus, requires reference not only to morphomes, but also to frequency and TAM categories.

Despite the relevance of these results, both to Romance philology and beyond as a methodologically straightforward way of quantifying the structural importance of different factors or grammatical components to explain a given phenomenon, several limitations should also be mentioned. The first and least critical one is that the operationalization leading to the statistical analysis and results in Table 6 is one among several similarly plausible/sensible options. Other such alternatives have been explored (for instance, including person and number, tense and mood, and N, L, and P as separate predictors, and using frequency difference rather than combined frequency) and results were not found to differ in the relevant aspects emphasized here. Thus, in a regression model *Stem_distance* ~ *pers* + *num* + *mood* + *tense* + *Freq-diff* + *N* + *L* + *P*, the broad results (reported in Table 7) would be same as before: that the morphomic categories (N, L, P) are most important, followed by frequency, and inherent inflectional categories (mood, tense), while contextual inflectional categories (person and number) have no statistically significant effect.

**Table 7:** Results of an alternative linear regression model II.

| | Estimate | Std. error | *t* Value | Pr (>\|*t*\|) |
|---|---|---|---|---|
| (Intercept) | 7.77E-01 | 5.56E-03 | 139.831 | 2E-16*** |
| pers1 | 2.07E-03 | 4.20E-03 | 0.494 | 0.621 |
| num1 | 1.06E-03 | 3.91E-03 | 0.273 | 0.785 |
| mood1 | −2.19E-02 | 3.96E-03 | −5.539 | 4.05E-08*** |
| tense1 | −2.78E-02 | 6.19E-03 | −4.487 | 8.22E-06*** |
| freq_diff | 3.93E-08 | 3.79E-09 | 10.392 | 2E-16*** |
| N1 | −2.46E-01 | 6.14E-03 | −40.031 | 2E-16*** |
| L1 | −2.37E-01 | 5.61E-03 | −42.196 | 2E-16*** |
| P1 | −2.17E-01 | 4.54E-03 | −47.686 | 2E-16*** |

Other limitations are deeper, and should be addressed in future research. The first relates to the definition/formalization of semantic structure in the paradigm. A finer-grained approach could incorporate feature structures, that is, the fact that certain values are supposed to be closer than others (for instance, first and second person closer than first and third, 3SG and 3PL closer than 1SG and 1PL, see discussion around Table 3). This was not done here due to the lack of consensus on the "correct" feature structure of person and number. A finest-grained approach could make use of modern methods in corpus-based distributional semantics (for example, word2vec) to sidestep this issue and measure, directly, the relative (dis)similarity of different person-number and TAM values, thus incorporating semantic structure as a continuous rather than categorical predictor. This possibility, challenging in its own right, will be left for future research.

Another limitation is more ontological in nature and relates to the productivity of the different structural factors that I surveyed here. Inspecting all stem alternation patterns as I did here captures the synchrony of the family with abundant data but glosses over the different status of alternation patterns in different lexemes. As explained in Section 2, stem alternations in very many verbs in the database (for example, N of Spanish p**ie**rdo vs. p**e**rdemos 'lose', or the L of Portuguese di**g**o vs. di**z**es 'say') are simply inherited from regular sound changes in Proto-Romance. This provides arguably little evidence about whether morphomic structure has been actively involved in the presence and paradigmatic distribution of these alternations in the modern languages. A more qualitative approach to the influence of morphomic versus semantic structures in paradigmatic structure could decide to focus on innovative/analogical stem alternation patterns exclusively, and maybe even on morphological alternations different from the inherited ones (for instance, ue/o, g/z) that are characteristic of established morphomic templates (see Herce forthcoming for such an approach to quantify the productivity of Romance morphomes).

Last, but not least, the present research has explored the Romance family and lexicon as a whole, averaging across lexemes and varieties regarding the proportion of stem alternations, cell frequencies, etc. It is not the case, of course, or this should at least be subject to empirical test, that stem alternations across Romance languages are largely the same. It could well be that the relative weight of inherited morphomic structure, semantic structure, and frequency differ substantially from one variety, area, or branch of Romance to another. This would be a most interesting object of analysis to explore in conjunction with a philologically-informed account of concrete historical events (for example, semantic drift of tenses, language contact, other sound changes, etc.) taking place separately in different varieties. Because of its complexity in its own right and because it exceeds the goals of the present research, this has been glossed over here, although it could be the subject of a separate future investigation.

# 5 Conclusion

This paper constitutes the first attempt to quantify the relative importance of morphomic patterns, semantic values, and token frequency in the stem alternation patterns in Romance verbal inflection. Relying on very rich data (2,151 alternation patterns in 1,613 lexemes across 57 Romance varieties, see Figure 1), the relative stem-similarity of different paradigm cells was calculated (see the complete distance matrix in the appendix, and the complete dataset in the Supplementary Materials). This can then be used as a window into the morphological architecture of the Romance verbal paradigm.

The results of an explicit statistical model (linear regression) identify a scale with respect to the relative importance of different factors. Morphomic structure (that is the unnatural domains N, L, and P of stem predictability inherited from Proto-Romance) was found to be the most important factor to predict the stem (dis)similarity of two paradigm cells. Somewhat less important, but still highly statistically significant was the correlation between stem alternation and the token frequency of different cells. Less important still, but still highly significant, were inherent inflectional values tense, aspect and mood. By contrast, contextual inflectional values like person and number agreement were found not to play a significant effect in the structuring of stem alternation patterns in Romance. That is, sharing a person or number value was not associated with more stem similarity.

These results constitute quantitative statistical confirmation of various claims in the literature. With respect to Romance verb stem alternations, it supports extant qualitative research highlighting the extraordinary importance of purely morphological domains and structures in the organization of Romance stem allomorphy, both synchronic and diachronic. Although the sound changes that generated N and L type alternations must have occurred nearly 2,000 years ago, and although the cells they applied to do not constitute semantic or syntactic natural classes, these structures have remained the most important organizing principle for Romance stem alternations.

In Figure 3 we find a hierarchical clustering arrived at via the hclust() function (method = 'single') of the R package 'cultevo' (Stadler 2018). We see that the clusters, based on the raw distances in the appendix, are very much in agreement with the morphomic domains identified in Table 2: From top to bottom we can see clusters for 0, PYTA, and L cells. The semantically core cells of N and NL also cluster together in Figure 3. This confirms the insights of Autonomous Morphology that morphology can have rules and principles of its own, and that these can be remarkably stable across time and space.

At the same time, Figure 3, and the present research, show that this is not the only structural principle at work in Romance verb stems. A higher token frequency is also strongly associated with a greater chance of stem alternations (see PRS and PRET cells). As argued by morphologists like Bybee, high token frequency provides a level of autonomy (in this case from inherited morphomic and semantic structure). Thus, if we observe which paradigm cells break continuity with the morphomic domains of Table 2, we see that it is the frequent cells 1SG.PRS.IND, and 3PL.PRS.IND that have become more dissimilar to their morphologically closest cells.

The 1SG.PRS.IND, for example, is the third most frequent cell in the corpus.[3] It is most stem-similar (see appendix) to the 1SG.PRS.SBJV, a cell with which it shares its morphomic domain (see Table 2). Despite this, a total of 503 stem alternations patterns (26%) have been found to include the former cell but not the latter or *vice versa*. The high frequency of the cell, as well as its different mood value relative to the other N/L cells, must be contributing to its relative autonomy. Mood and tense, that is so-called 'inherent' inflectional structure, is precisely the third highly significant factor in the structuring of stem alternation in contemporary Romance. Contextual inflectional values (that is, person and number), by contrast, have not been found to drive Romance

**3** It is probably just the idiosyncrasy of the extant Latin corpus (written, few conversations) that makes the 3SG.PRET.IND more frequent than the 1SG.PRS.IND.

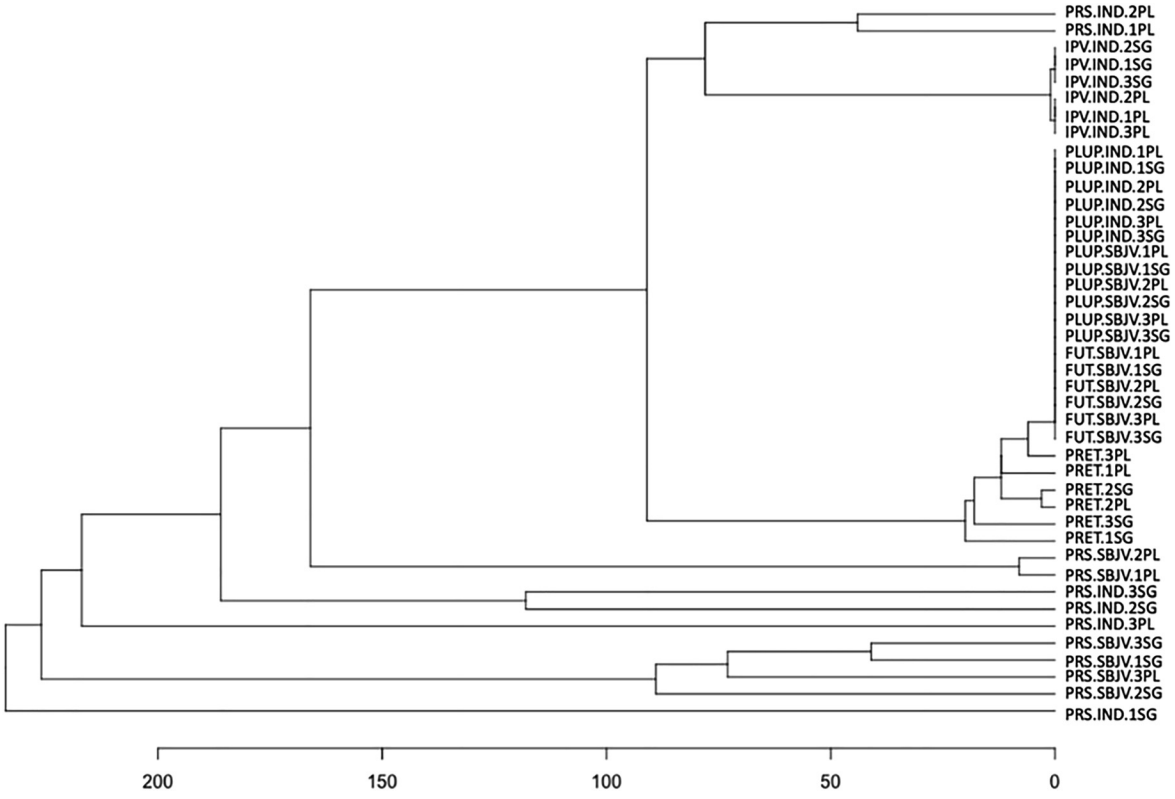


**Figure 3:** Hierarchical clustering of Romance paradigm cells based on stem similarity.

stem alternation in any systematic way. This provides quantitative confirmation of traditional qualitative observations in the literature on stem alternation and suppletion (Bybee 1985: 57; Corbett 2007) that inherent inflectional categories are the ones that tend to control them.

Despite their apparent exceptionality[4] with respect to the large role of inherited morphomic patterns, Romance verb stem alternations have been found here to be much less exotic in all other respects. Established knowledge and empirical insights on the crucial role of frequency and semantic (TAM) structure in paradigmatic architecture, thus, remain valid even here. A complete picture of Romance stem alternation patterns (and probably most other "highly morphomic" inflectional systems) needs to take into account not only inherited morphological predictability relations, but also frequency and inherent-inflectional semantic structure. Morphological Autonomy, thus, should be understood, not as the complete independence of morphology from other components of grammar (à la Blevins), but merely as the possibility for historical morphological accidents and idiosyncrasies to outrank other more universal structural biases (for example, semantic natural classes) in concrete systems. Further research could be aimed at assessing whether/to what extent this is so cross-linguistically, or if Romance should be understood as an exotic outlier.

4 The role of inherited purely morphological structures in other families should be contrasted by replicating the present research with a different dataset. Impressionistically, stem alternations seem better aligned to TAM in most other languages and families (for example, Germanic and other Indo-European), but there might also be some (for example, Saami, Tol, Chinantec, Wubuy, etc. see Herce 2020b, 2023), where morphomic structures are even "more morphomic" than in Romance.

# Appendix

Cell-to-cell stem distance matrix (number of patterns above diagonal, percentage under).

| | PRS.IND.1PL | PRS.IND.1SG | PRS.IND.2PL | PRS.IND.2SG | PRS.IND.3PL | PRS.IND.3SG | PRS.SBJV.1PL | PRS.SBJV.1SG | PRS.SBJV.2PL | PRS.SBJV.2SG | PRS.SBJV.3PL | PRS.SBJV.3SG | IPV.IND.1PL | IPV.IND.1SG | IPV.IND.2PL | IPV.IND.2SG | IPV.IND.3PL | IPV.IND.3SG | PLUP.IND.1PL | PLUP.IND.1SG | PLUP.IND.2PL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRS.IND.1PL | – | 1,102 | 44 | 832 | 763 | 862 | 627 | 1,226 | 631 | 1,167 | 1,182 | 1,225 | 85 | 84 | 85 | 84 | 85 | 84 | 174 | 174 | 174 |
| PRS.IND.1SG | 51% | – | 1,138 | 618 | 601 | 614 | 914 | 503 | 920 | 592 | 537 | 542 | 1,129 | 1,128 | 1,129 | 1,128 | 1,129 | 1,128 | 349 | 349 | 349 |
| PRS.IND.2PL | 2% | 53% | – | 820 | 797 | 848 | 661 | 1,264 | 661 | 1,205 | 1,218 | 1,263 | 79 | 78 | 79 | 78 | 79 | 78 | 170 | 170 | 170 |
| PRS.IND.2SG | 39% | 29% | 38% | – | 351 | 118 | 1,153 | 1,032 | 1,161 | 943 | 1,048 | 1,041 | 881 | 880 | 881 | 880 | 881 | 880 | 295 | 295 | 295 |
| PRS.IND.3PL | 35% | 28% | 37% | 16% | – | 321 | 1,086 | 837 | 1,088 | 896 | 799 | 860 | 796 | 795 | 796 | 795 | 796 | 795 | 303 | 303 | 303 |
| PRS.IND.3SG | 40% | 29% | 39% | 5% | 15% | – | 1,166 | 1,029 | 1,172 | 992 | 1,025 | 1,006 | 913 | 912 | 913 | 912 | 913 | 912 | 312 | 312 | 312 |
| PRS.SBJV.1PL | 32% | 47% | 34% | 59% | 56% | 60% | – | 621 | 8 | 562 | 575 | 620 | 660 | 661 | 660 | 661 | 660 | 661 | 328 | 328 | 328 |
| PRS.SBJV.1SG | 63% | 26% | 65% | 53% | 43% | 53% | 32% | – | 623 | 89 | 90 | 41 | 1,241 | 1,242 | 1,241 | 1,242 | 1,241 | 1,242 | 368 | 368 | 368 |
| PRS.SBJV.2PL | 32% | 47% | 34% | 59% | 56% | 60% | 0% | 32% | – | 564 | 573 | 622 | 660 | 661 | 660 | 661 | 660 | 661 | 328 | 328 | 328 |
| PRS.SBJV.2SG | 60% | 30% | 62% | 48% | 46% | 51% | 29% | 5% | 29% | – | 171 | 98 | 1,192 | 1,193 | 1,192 | 1,193 | 1,192 | 1,193 | 368 | 368 | 368 |
| PRS.SBJV.3PL | 61% | 28% | 62% | 54% | 41% | 53% | 29% | 5% | 29% | 9% | – | 73 | 1,201 | 1,202 | 1,201 | 1,202 | 1,201 | 1,202 | 368 | 368 | 368 |
| PRS.SBJV.3SG | 63% | 28% | 65% | 53% | 44% | 52% | 32% | 2% | 32% | 5% | 4% | – | 1,256 | 1,257 | 1,256 | 1,257 | 1,256 | 1,257 | 368 | 368 | 368 |
| IPV.IND.1PL | 4% | 52% | 4% | 41% | 37% | 42% | 34% | 64% | 34% | 61% | 62% | 64% | – | 1 | 0 | 1 | 0 | 1 | 161 | 161 | 161 |
| IPV.IND.1SG | 4% | 52% | 4% | 41% | 37% | 42% | 34% | 64% | 34% | 61% | 62% | 64% | 0% | – | 1 | 0 | 1 | 0 | 160 | 160 | 160 |
| IPV.IND.2PL | 4% | 52% | 4% | 41% | 37% | 42% | 34% | 64% | 34% | 61% | 62% | 64% | 0% | 0% | – | 1 | 0 | 1 | 161 | 161 | 161 |
| IPV.IND.2SG | 4% | 52% | 4% | 41% | 37% | 42% | 34% | 64% | 34% | 61% | 62% | 64% | 0% | 0% | 0% | – | 1 | 0 | 160 | 160 | 160 |
| IPV.IND.3PL | 4% | 52% | 4% | 41% | 37% | 42% | 34% | 64% | 34% | 61% | 62% | 64% | 0% | 0% | 0% | 0% | – | 1 | 161 | 161 | 161 |
| IPV.IND.3SG | 4% | 52% | 4% | 41% | 37% | 42% | 34% | 64% | 34% | 61% | 62% | 64% | 0% | 0% | 0% | 0% | 0% | – | 160 | 160 | 160 |
| PLUP.IND.1PL | 34% | 68% | 33% | 58% | 59% | 61% | 70% | 79% | 70% | 79% | 79% | 79% | 32% | 31% | 32% | 31% | 32% | 31% | – | 0 | 0 |
| PLUP.IND.1SG | 34% | 68% | 33% | 58% | 59% | 61% | 70% | 79% | 70% | 79% | 79% | 79% | 32% | 31% | 32% | 31% | 32% | 31% | 0% | – | 0 |
| PLUP.IND.2PL | 34% | 68% | 33% | 58% | 59% | 61% | 70% | 79% | 70% | 79% | 79% | 79% | 32% | 31% | 32% | 31% | 32% | 31% | 0% | 0% | – |
| PLUP.IND.2SG | 34% | 68% | 33% | 58% | 59% | 61% | 70% | 79% | 70% | 79% | 79% | 79% | 32% | 31% | 32% | 31% | 32% | 31% | 0% | 0% | 0% |
| PLUP.IND.3PL | 34% | 68% | 33% | 58% | 59% | 61% | 70% | 79% | 70% | 79% | 79% | 79% | 32% | 31% | 32% | 31% | 32% | 31% | 0% | 0% | 0% |
| PLUP.IND.3SG | 34% | 68% | 33% | 58% | 59% | 61% | 70% | 79% | 70% | 79% | 79% | 79% | 32% | 31% | 32% | 31% | 32% | 31% | 0% | 0% | 0% |
| PLUP.SBJV.1PL | 23% | 68% | 23% | 62% | 58% | 63% | 42% | 72% | 42% | 71% | 71% | 72% | 21% | 21% | 21% | 21% | 21% | 21% | 0% | 0% | 0% |
| PLUP.SBJV.1SG | 23% | 68% | 23% | 62% | 58% | 63% | 42% | 72% | 42% | 71% | 71% | 72% | 21% | 21% | 21% | 21% | 21% | 21% | 0% | 0% | 0% |
| PLUP.SBJV.2PL | 23% | 68% | 23% | 62% | 58% | 63% | 42% | 72% | 42% | 71% | 71% | 72% | 21% | 21% | 21% | 21% | 21% | 21% | 0% | 0% | 0% |
| PLUP.SBJV.2SG | 23% | 68% | 23% | 62% | 58% | 63% | 42% | 72% | 42% | 71% | 71% | 72% | 21% | 21% | 21% | 21% | 21% | 21% | 0% | 0% | 0% |
| PLUP.SBJV.3PL | 23% | 68% | 23% | 62% | 58% | 63% | 42% | 72% | 42% | 71% | 71% | 72% | 21% | 21% | 21% | 21% | 21% | 21% | 0% | 0% | 0% |
| PLUP.SBJV.3SG | 23% | 68% | 23% | 62% | 58% | 63% | 42% | 72% | 42% | 71% | 71% | 72% | 21% | 21% | 21% | 21% | 21% | 21% | 0% | 0% | 0% |
| FUT.SBJV.1PL | 33% | 69% | 33% | 55% | 64% | 58% | 49% | 74% | 49% | 66% | 79% | 80% | 27% | 27% | 27% | 27% | 27% | 27% | 0% | 0% | 0% |
| FUT.SBJV.1SG | 33% | 69% | 33% | 55% | 64% | 58% | 49% | 74% | 49% | 66% | 79% | 80% | 27% | 27% | 27% | 27% | 27% | 27% | 0% | 0% | 0% |
| FUT.SBJV.2PL | 33% | 69% | 33% | 55% | 64% | 58% | 49% | 74% | 49% | 66% | 79% | 80% | 27% | 27% | 27% | 27% | 27% | 27% | 0% | 0% | 0% |
| FUT.SBJV.2SG | 33% | 69% | 33% | 55% | 64% | 58% | 49% | 74% | 49% | 66% | 79% | 80% | 27% | 27% | 27% | 27% | 27% | 27% | 0% | 0% | 0% |
| FUT.SBJV.3PL | 33% | 69% | 33% | 55% | 64% | 58% | 49% | 74% | 49% | 66% | 79% | 80% | 27% | 27% | 27% | 27% | 27% | 27% | 0% | 0% | 0% |

(continued)

| | PRS.IND.1PL | PRS.IND.1SG | PRS.IND.2PL | PRS.IND.2SG | PRS.IND.3PL | PRS.IND.3SG | PRS.SBJV.1PL | PRS.SBJV.1SG | PRS.SBJV.2PL | PRS.SBJV.2SG | PRS.SBJV.3PL | PRS.SBJV.3SG | IPV.IND.1PL | IPV.IND.1SG | IPV.IND.2PL | IPV.IND.2SG | IPV.IND.3PL | IPV.IND.3SG | PLUP.IND.1PL | PLUP.IND.1SG | PLUP.IND.2PL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FUT.SBJV.3SG | 33% | 69% | 33% | 55% | 64% | 58% | 49% | 74% | 49% | 66% | 79% | 80% | 27% | 27% | 27% | 27% | 27% | 27% | 0% | 0% | 0% |
| PRET.IND.1PL | 31% | 73% | 29% | 59% | 60% | 62% | 53% | 75% | 53% | 70% | 75% | 76% | 28% | 28% | 28% | 28% | 28% | 28% | 3% | 3% | 3% |
| PRET.IND.1SG | 34% | 78% | 33% | 62% | 64% | 65% | 58% | 80% | 58% | 75% | 80% | 81% | 32% | 32% | 32% | 32% | 32% | 32% | 5% | 5% | 5% |
| PRET.IND.2PL | 29% | 72% | 28% | 58% | 59% | 61% | 53% | 74% | 53% | 70% | 74% | 76% | 27% | 27% | 27% | 27% | 27% | 27% | 3% | 3% | 3% |
| PRET.IND.2SG | 29% | 72% | 28% | 58% | 59% | 60% | 53% | 74% | 53% | 70% | 74% | 76% | 27% | 27% | 27% | 27% | 27% | 27% | 3% | 3% | 3% |
| PRET.IND.3PL | 35% | 77% | 34% | 64% | 64% | 66% | 57% | 79% | 57% | 75% | 79% | 81% | 33% | 32% | 33% | 32% | 33% | 32% | 1% | 1% | 1% |
| PRET.IND.3SG | 36% | 79% | 34% | 63% | 65% | 66% | 58% | 81% | 58% | 76% | 81% | 82% | 34% | 34% | 34% | 34% | 34% | 34% | 4% | 4% | 4% |

Cell-to-cell stem distance matrix, continued.

| | PLUP.IND.2SG | PLUP.IND.3PL | PLUP.IND.3SG | PLUP.SBJV.1PL | PLUP.SBJV.1SG | PLUP.SBJV.2PL | PLUP.SBJV.2SG | PLUP.SBJV.3PL | PLUP.SBJV.3SG | FUT.SBJV.1PL | FUT.SBJV.1SG | FUT.SBJV.2PL | FUT.SBJV.2SG | FUT.SBJV.3PL | FUT.SBJV.3SG | PRET.IND.1PL | PRET.IND.1SG | PRET.IND.2PL | PRET.IND.2SG | PRET.IND.3PL | PRET.IND.3SG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRS.IND.1PL | 174 | 174 | 174 | 401 | 401 | 401 | 401 | 401 | 401 | 113 | 113 | 113 | 113 | 113 | 113 | 437 | 490 | 419 | 420 | 504 | 508 |
| PRS.IND.1SG | 349 | 349 | 349 | 1,173 | 1,173 | 1,173 | 1,173 | 1,173 | 1,173 | 234 | 234 | 234 | 234 | 234 | 234 | 1,044 | 1,109 | 1,026 | 1,027 | 1,107 | 1,123 |
| PRS.IND.2PL | 170 | 170 | 170 | 400 | 400 | 400 | 400 | 400 | 400 | 113 | 113 | 113 | 113 | 113 | 113 | 421 | 474 | 403 | 404 | 488 | 492 |
| PRS.IND.2SG | 295 | 295 | 295 | 1,054 | 1,054 | 1,054 | 1,054 | 1,054 | 1,054 | 186 | 186 | 186 | 186 | 186 | 186 | 846 | 893 | 828 | 825 | 909 | 905 |
| PRS.IND.3PL | 303 | 303 | 303 | 993 | 993 | 993 | 993 | 993 | 993 | 217 | 217 | 217 | 217 | 217 | 217 | 855 | 914 | 837 | 838 | 918 | 926 |
| PRS.IND.3SG | 312 | 312 | 312 | 1,073 | 1,073 | 1,073 | 1,073 | 1,073 | 1,073 | 197 | 197 | 197 | 197 | 197 | 197 | 885 | 932 | 867 | 864 | 948 | 944 |
| PRS.SBJV.1PL | 328 | 328 | 328 | 637 | 637 | 637 | 637 | 637 | 637 | 166 | 166 | 166 | 166 | 166 | 166 | 673 | 737 | 670 | 671 | 726 | 741 |
| PRS.SBJV.1SG | 368 | 368 | 368 | 1,092 | 1,092 | 1,092 | 1,092 | 1,092 | 1,092 | 252 | 252 | 252 | 252 | 252 | 252 | 949 | 1,013 | 946 | 947 | 1,010 | 1,025 |
| PRS.SBJV.2PL | 328 | 328 | 328 | 636 | 636 | 636 | 636 | 636 | 636 | 166 | 166 | 166 | 166 | 166 | 166 | 673 | 737 | 670 | 671 | 726 | 741 |
| PRS.SBJV.2SG | 368 | 368 | 368 | 1,075 | 1,075 | 1,075 | 1,075 | 1,075 | 1,075 | 226 | 226 | 226 | 226 | 226 | 226 | 889 | 953 | 886 | 887 | 950 | 965 |
| PRS.SBJV.3PL | 368 | 368 | 368 | 1,071 | 1,071 | 1,071 | 1,071 | 1,071 | 1,071 | 269 | 269 | 269 | 269 | 269 | 269 | 950 | 1,014 | 947 | 948 | 1,011 | 1,026 |
| PRS.SBJV.3SG | 368 | 368 | 368 | 1,092 | 1,092 | 1,092 | 1,092 | 1,092 | 1,092 | 271 | 271 | 271 | 271 | 271 | 271 | 969 | 1,033 | 966 | 967 | 1,030 | 1,045 |
| IPV.IND.1PL | 161 | 161 | 161 | 366 | 366 | 366 | 366 | 366 | 366 | 91 | 91 | 91 | 91 | 91 | 91 | 398 | 463 | 380 | 381 | 465 | 481 |
| IPV.IND.1SG | 160 | 160 | 160 | 366 | 366 | 366 | 366 | 366 | 366 | 91 | 91 | 91 | 91 | 91 | 91 | 397 | 462 | 379 | 380 | 464 | 480 |
| IPV.IND.2PL | 161 | 161 | 161 | 366 | 366 | 366 | 366 | 366 | 366 | 91 | 91 | 91 | 91 | 91 | 91 | 398 | 463 | 380 | 381 | 465 | 481 |
| IPV.IND.2SG | 160 | 160 | 160 | 366 | 366 | 366 | 366 | 366 | 366 | 91 | 91 | 91 | 91 | 91 | 91 | 397 | 462 | 379 | 380 | 464 | 480 |
| IPV.IND.3PL | 161 | 161 | 161 | 366 | 366 | 366 | 366 | 366 | 366 | 91 | 91 | 91 | 91 | 91 | 91 | 398 | 463 | 380 | 381 | 465 | 481 |
| IPV.IND.3SG | 160 | 160 | 160 | 366 | 366 | 366 | 366 | 366 | 366 | 91 | 91 | 91 | 91 | 91 | 91 | 397 | 462 | 379 | 380 | 464 | 480 |
| PLUP.IND.1PL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 24 | 12 | 14 | 6 | 20 |
| PLUP.IND.1SG | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 24 | 12 | 14 | 6 | 20 |
| PLUP.IND.2PL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 24 | 12 | 14 | 6 | 20 |
| PLUP.IND.2SG | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 24 | 12 | 14 | 6 | 20 |
| PLUP.IND.3PL | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 24 | 12 | 14 | 6 | 20 |
| PLUP.IND.3SG | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 24 | 12 | 14 | 6 | 20 |
| PLUP.SBJV.1PL | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 111 | 30 | 30 | 103 | 117 |
| PLUP.SBJV.1SG | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 111 | 30 | 30 | 103 | 117 |
| PLUP.SBJV.2PL | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 111 | 30 | 30 | 103 | 117 |
| PLUP.SBJV.2SG | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 111 | 30 | 30 | 103 | 117 |
| PLUP.SBJV.3PL | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 111 | 30 | 30 | 103 | 117 |
| PLUP.SBJV.3SG | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 0 | 48 | 111 | 30 | 30 | 103 | 117 |
| FUT.SBJV.1PL | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 0 | 27 | 27 | 27 | 26 | 21 | 21 |
| FUT.SBJV.1SG | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 0 | 27 | 27 | 27 | 26 | 21 | 21 |
| FUT.SBJV.2PL | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 0 | 27 | 27 | 27 | 26 | 21 | 21 |
| FUT.SBJV.2SG | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 0 | 27 | 27 | 27 | 26 | 21 | 21 |
| FUT.SBJV.3PL | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 0 | 27 | 27 | 27 | 26 | 21 | 21 |
| FUT.SBJV.3SG | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | – | 27 | 27 | 27 | 26 | 21 | 21 |
| PRET.IND.1PL | 3% | 3% | 3% | 4% | 4% | 4% | 4% | 4% | 4% | 9% | 9% | 9% | 9% | 9% | 9% | – | 65 | 18 | 21 | 67 | 85 |
| PRET.IND.1SG | 5% | 5% | 5% | 10% | 10% | 10% | 10% | 10% | 10% | 9% | 9% | 9% | 9% | 9% | 9% | 5% | – | 83 | 82 | 32 | 20 |
| PRET.IND.2PL | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 9% | 9% | 9% | 9% | 9% | 9% | 1% | 6% | – | 3 | 85 | 103 |
| PRET.IND.2SG | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 9% | 9% | 9% | 9% | 9% | 9% | 1% | 6% | 0% | – | 88 | 102 |
| PRET.IND.3PL | 1% | 1% | 1% | 9% | 9% | 9% | 9% | 9% | 9% | 7% | 7% | 7% | 7% | 7% | 7% | 5% | 2% | 6% | 6% | – | 18 |
| PRET.IND.3SG | 4% | 4% | 4% | 10% | 10% | 10% | 10% | 10% | 10% | 7% | 7% | 7% | 7% | 7% | 7% | 6% | 1% | 7% | 7% | 1% | – |

# References


Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 54–82. Oxford: Oxford University Press.

Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 29(3). 429–464.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Blevins, James P. forthcoming. Two frameworks of morphological analysis. *Linguistic Analysis*.

Bloomfield, Leonard. 1926. A set of postulates for the science of language. *Language* 2(3). 153–164.

Bolinger, Dwight L. 1948. On defining the morpheme. *Word* 4(1). 18–23.

Booij, Geert. 1996. Inherent versus contextual inflection and the split morphology hypothesis. In *Yearbook of morphology 1995*, 1–16. Dordrecht: Springer.

Boyé, Gilles & Patricia Cabredo-Hofherr. 2006. The structure of allomorphy in Spanish verbal inflection. *Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset* 13. 9–24.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form*. Philadelphia: John Benjamins.

Bybee, Joan. 2006. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Carstairs-McCarthy, Andrew. 2010. *The evolution of morphology*. Oxford: Oxford University Press.

Corbett, Greville G. 2007. Canonical typology, suppletion, and possible words. *Language* 83(1). 8–42.

Coseriu, Eugenio. 1976. *Das romanische verbalsystem*, vol. 66. Tübinger: Gunter Narr.

Cysouw, Michael. 2009. *The paradigmatic structure of person marking*. Oxford: Oxford University Press.

Delatte, Louis, Étienne Evrard, Suzanne Govaerts & Joseph Denooz. 1981. *Dictionnaire fréquentiel et index inverse de la langue latine*. Liege: L.A.S.L.A.

Embick, David. 2015. *The morpheme*. Amsterdam: De Gruyter.

Esher, Louise. 2015. Morphomes and predictability in the history of Romance perfects. *Diachronica* 32(4). 494–529.

Halle, Morris & Alec Marantz. 1994. Some key features of distributed morphology. *MIT Working Papers in Linguistics* 21. 275–288.

Herce, Borja. 2016. Why frequency and morphological irregularity are not independent variables in Spanish: A response to Fratini et al. (2014). *Corpus Linguistics and Linguistic Theory* 12(2). 389–406.

Herce, Borja. 2020a. Alignment of forms in Spanish verbal inflection: The gang poner, tener, venir, salir, valer as a window into the nature of paradigmatic analogy and predictability. *Morphology* 30(2). 91–115.

Herce, Borja. 2020b. *A typological approach to the morphome*. Guildford, UK: University of the Basque Country and University of Surrey PhD dissertation.

Herce, Borja. 2023. *The typological diversity of morphomes: A cross-linguistic study of unnatural morphology*. Oxford: Oxford University Press.

Herce, Borja. forthcoming. Morphological autonomy and the long-term vitality of morphomes: CVC- to C(V)- stems in Romance verbs, hiatus avoidance, and paradigmatic analogy.

Herce, Borja, Carmen Saldana, John Mansfield & Balthasar Bickel. forthcoming. Positional splits in person-number agreement paradigms reflect a naturalness gradient: Typological and experimental evidence.

Maiden, Martin. 1992. Irregularity as a determinant of morphological change. *Journal of Linguistics* 28(2). 285–312.

Maiden, Martin. 2016. Morphomes in diachrony. In Ana R. Luís & Ricardo Bermúdez-Otero (eds.), *The morphome debate*, 33–63. Oxford: Oxford University Press.

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

Maiden, Martin, John Charles Smith, Silvio Cruschina, Marc-Olivier Hinzelin & Maria Goldbach. 2010. Oxford online database of Romance verb morphology. Available at: http://romverbmorph.clp.ox.ac.uk/.

Matthews, Peter H. 1965. The inflectional component of a word-and-paradigm grammar. *Journal of Linguistics* 1(2). 139–171.

Montermini, Fabio & Olivier Bonami. 2013. Stem spaces and predictability in verbal inflection. *Lingue e Linguaggio* 12(2). 171–190.

Nevins, Andrew, Cilene Rodrigues & Kevin Tang. 2015. The rise and fall of the L-shaped morphome: Diachronic and experimental studies. *Probus* 27(1). 101–155.

Pinker, Steven. 1999. *Words and rules*. New York: HarperCollins.

R Core Team. 2014. *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing Vienna. Available at: http://www.R-project.org/.

R Studio Team. 2020. *Rstudio: Integrated development environment for R*. Boston, MA: RStudio, PBC. Available at: http://www.rstudio.com/.

Saldana, Carmen, Borja Herce & Balthasar Bickel. 2022. More or less unnatural: Semantic similarity shapes the learnability and cross-linguistic distribution of syncretism in morphological paradigms. *Open Mind*. https://doi.org/10.17605/OSF.IO/JPUM6.

Stadler, Kevin. 2018. Cultevo: Tools, measures and statistical tests for cultural evolution. Available at: https://kevinstadler.github.io/cultevo/.

Stump, Gregory & Raphael A. Finkel. 2013. *Morphological typology: From word to paradigm*, vol. 138. Cambridge: Cambridge University Press.

Wu, Shijie, Ryan Cotterell & Timothy J. O'Donnell. 2019. Morphological irregularity correlates with frequency. *arXiv preprint arXiv: 1906.11483*.

Zipf, George Kingsley. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Boston, MA: Houghton Mifflin Company.

**Supplementary Material:** The online version of this article offers supplementary material (https://doi.org/10.1515/lingvan-2022-0028).