Proceedings of the Society for Computation in Linguistics

Volume 4 Article 11

2021

Formalizing Inflectional Paradigm Shape with Information Theory

Grace LeFevre
The Ohio State University, lefevre.33@osu.edu

Micha Elsner
The Ohio State University, elsner.14@osu.edu

Andrea D. Sims
The Ohio State University, sims.120@osu.edu

Formalizing Inflectional Paradigm Shape with Information Theory

Grace LeFevre
The Ohio State University
lefevre.33@osu.edu

Micha Elsner*
The Ohio State University
elsner.14@osu.edu

Andrea D. Sims*
The Ohio State University
sims.120@osu.edu

Abstract

"Paradigm shape," our term for the morphological structure formed by implicative relations between inflected forms, has not been formally quantified in a gradient manner. We develop a method to formalize paradigm shape by modeling the joint effect of stem alternations and affixes. Applied to Spanish verbs, our model successfully captures aspects of both allomorphic and distributional classes. These results are replicable and extendable to other languages.

1 Introduction

In this paper, we explore what we call "paradigm shape,"[1] which is a type of morphological structure characterized by the implicative relations holding among inflected forms. This structure reflects the predictable, patterned ways in which stem alternants and even fully suppletive allomorphs occur in parallel paradigm cells across inflection classes in some languages (see the Spanish verbs SENTIR 'feel', PENSAR 'think', and MOVER 'move' in Table 1). Historically, some Romance verbs shifted to better conform to existing paradigm shapes, indicating that this is a cognitively real organizing principle for speakers (Maiden, 2005). As such, it has important implications for language learning and change.

We develop a computational method to precisely quantify similarity in paradigm shape. Building on previous work measuring the interpredictability of word forms (Ackerman et al., 2009; Bonami and Beniamine, 2016), we apply information-theoretic entropy to these forms to compute sets of values characterizing the shapes of the inflection classes. In contrast to previous work focusing exclusively on stem shape organization (e.g. Maiden, 2005; Boyé and Cabredo Hofherr, 2006), our method focuses equally on stems and affixes. Furthermore, our results enable direct analysis of both allomorphic and distributional classes (Baerman et al., 2017), where most previous shape-based analysis has been purely distributional. As such, this work provides a unified, computational approach to phenomena relating to paradigm shape that have predominantly been treated separately in the past. We implement our method on Spanish in this paper as a test case.[2]

2 Phenomenon to be modeled

Morphological structure is characterized by both syntagmatic relations and paradigmatic relations. Syntagmatic relations involve the combinatorial properties of morphemes, such as the relationship between the Spanish verb stem *cant-* 'sing' and 1SG suffix *-o*. Paradigmatic relations involve substitutional relationships, such as the relationship between 1SG *canto* and other inflected forms of the same lexeme, or between *canto* and the 1SG of other verbs (see Table 1). In this paper we seek to model a kind of paradigmatic relation that we call "paradigm shape." A lexeme's paradigm shape is defined by the implicative relations holding among its inflected forms, for instance how well an unobserved form of some lexeme is predicted by an observed form. Implicative relations of this sort bind the forms in a paradigm together (Wurzel, 1989) and conceptually, two lexemes have the same paradigm shape to the extent that they exhibit the same paradigmatic implications.

Spanish offers an interesting test case because, along with other Romance languages, it is well known for having paradigmatically-

*Joint second authors.

[1]This use of the term should not be confused with the same term used elsewhere to refer to the number of morphosyntactic property sets a lexeme expresses (e.g. Corbett, 2009).

[2]The code and data for our analysis are available at github.com/gracelefevre/paradigm-shape.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 102-115.
Held on-line February 14-19, 2021

LEXEME	GLOSS	PRS.1SG	PRS.2SG	PRS.3SG	PRS.1PL	PRS.2PL	PRS.3PL
CANTAR	'sing'	canto	cantas	canta	cantamos	cantáis	cantan
SUBIR	'rise'	subo	subes	sube	subimos	subís	suben
SENTIR	'feel'	siento	sientes	siente	sentimos	sentís	sienten
PENSAR	'think'	pienso	piensas	piensa	pensamos	pensáis	piensan
MOVER	'move'	muevo	mueves	mueve	movemos	movéis	mueven

Table 1: Present indicative forms of verbs from a few Spanish microclasses; stem alternations highlighted.

structured stem alternants in verbs. Spanish verbs are traditionally grouped into inflectional macroclasses (terminology from Beniamine et al., 2017) based on the theme vowel that shows up in the infinitive: *-a* vs. *-e* vs. *-i* (Butt et al., 2019). For the words in Table 1, this would group CANTAR and PENSAR together, define a second group for MOVER, and delineate a final group for SUBIR and SENTIR. However, there is clearly more to say about the morphological structure of these words. No verb in Table 1 has exactly the same exponence as any other, which is to say, they each represent a distinct inflectional *microclass* in the sense of Beniamine et al. (2017). SENTIR, PENSAR and MOVER have stem alternations whereas CANTAR and SUBIR do not. Moreover, the distribution of stem alternants is the same for each lexeme (highlighted by purple shaded cells), despite the alternations not involving the same phonology (*e~ie* vs. *o~ue*).[3]

The history of Romance is replete with examples of morphological change motivated by paradigmatic stem distributions of this sort. In Old Spanish, the present indicative forms of IRE 'go' were *vo, vas, va, imos, ides, van* (Maiden, 2005; O'Neill, 2018b), showing full stem suppletion with the same distribution of alternants as in Table 1.[4] The Old Spanish forms arose from incursion, in which two separate lexemes merge to become a single lexeme with suppletive forms. The fact that the result reproduced an existing distribution of alternants attests that the distribution was (and presumably still is) cognitively real for speakers.

More broadly, Maiden (2004, 2005, 2009) identifies three major distributions of stem alternants in Romance verbs: the "L-pattern" (shared alternation in 1SG present indicative and all present subjunctive), the "N-pattern" (shared alternation in 1SG, 2SG, 3SG, and 3PL of the present indicative), and the "U-pattern" (shared alternation in 1SG and 3PL present indicative and all present subjunctive).[5] In Table 1, SENTIR, PENSAR, and MOVER belong to the N-pattern. Maiden (2005, p. 169) observes that while details differ from one language to another, in the history of Romance speakers "... actually pass up golden opportunities to align allomorphs with morphosyntactic properties...", instead choosing to maintain these distributions, reinforce them, and extend them to new verbs. We are interested in the role this abstract distribution of alternants plays in facilitating or inhibiting inferences about the inflected forms of lexemes.

At the same time, as can be observed in Table 1, inflectional suffixes— in particular the theme vowels that show up in many inflected forms (e.g. PRS.1PL *-amos* vs. *-emos* vs. *-imos*)— have their own distribution. As noted above, the theme vowels group verbs into macroclasses differently than the stem alternations do. Theme vowels also impact how predictable other inflected forms of the same lexeme are. While the *a* theme vowel appears relatively consistently across the paradigm, the *e* and *i* classes sometimes collapse (compare PRS.1PL and PRS.2SG). As a result, these two macroclasses are more confusable.[6] Moreover, the theme vowel does not surface in some cells (e.g. PRS.1SG), making these cells poorly informative about other inflected forms of the lexeme. Cells/inflected form thus differ in how informative

[3] In Spanish, alternation is related to lexical stress placement; in the relevant verbs the diphthong alternants appear when the vowel is stressed and *e* and *o* appear when unstressed. However, for our purposes this is not material. We are interested in the effect of the resulting stem distributions on the implicative structure of the paradigm.

[4] This alternation has been leveled in Modern Spanish, which has present indicative forms *voy, vas, va, vamos, vais, van*.

[5] The U-pattern does not occur in Modern Spanish, having been replaced with the L-pattern (Maiden, 2005, p. 146). Among the modern languages, it is restricted to some Italian varieties and Romanian (Maiden, 2009, p. 47). We therefore do not consider it further in this paper.

[6] See Penny (1972) for changes in the history of Spanish that were motivated by this confusability.

they are about the inflected forms of other cells as a function of their suffixes.

We take "paradigm shape" in Spanish to encompass both the stem alternations and the suffixes. In this paper we develop methods for modeling their joint effect on the implicative relations holding among inflected forms and use this to quantify similarity in paradigm shape across lexemes.

3 Previous Work

Related work can be roughly divided into two lines of investigation. The first models the distribution of stem alternants within a paradigm. The second consists of work on inflectional complexity which is interested in the predictability of inflected forms. These have points of intersection, since both focus on the distribution of implicative relations within the paradigm. However, to the best of our knowledge our work is among the first seeking to integrate the insights of each.[7]

Work modeling the paradigmatic distribution of stem alternants (in Romance and elsewhere) has tended to approach it either from diachronic perspective (Aski, 1995; Herce, 2020; Hippisley et al., 2004; Juge, 1999; Maiden, 2004, 2005, 2009; O'Neill, 2018b; Wheeler, 2011), as noted above, or from the perspective of formal linguistic theory (Bonami and Boyé, 2002; Boyé and Cabredo Hofherr, 2006, 2010; Hippisley, 1998; Maiden, 2011; Montermini and Bonami, 2013; O'Neill, 2018a).[8] As an example of the latter approach, Boyé and Cabredo Hofherr (2006) identify eleven stem 'zones' for Spanish verbs— sets of paradigm cells which always exhibit the same stem form. No verb has a different stem for each zone[9] and Boyé and Cabredo Hofherr argue that the distribution of stems alternants is not random, but rather, systematically constrained by the organization of the stem space, which can be represented as an acyclic graph of implicative relations. Core insights of this and other formal work on stem organization are thus that (in Romance) certain cells predictably have the same stem form, that cells with different stem forms enter into predictable implicative relations, and that these relations are often parallel across classes.

At the same time, the formal theory approach has a number of limitations in the context of trying to quantify the extent to which lexemes have similar paradigm shapes. Two are important here. First, there is a limited ability to express *partial* similarity in the implicative relations holding among inflected forms. For example, Boyé and Cabredo Hofherr's method encodes implicative relations holding among stems, but not the extent to which words are similar in their implicative relations. In Maiden's classification into L-, N- and U-patterns, the patterns are discrete and any notion of similarity among patterns is left informal at best. Yet intuitively, some paradigms are more similar in shape than others, without being exactly the same. For example, VENIR 'come', DECIR 'say' and TENER 'have' follow the L-pattern but *additionally* have the N-pattern alternation (except in the 1SG indicative present, where the N- and L-patterns overlap). These verbs thus have a "modified" L-pattern. We want to quantify this and other distributional variation in fully gradient terms.

Second, formal analyses have tended to abstract away from affixes, in order to focus on stem organization. Yet as we note above, inferences about the inflected forms of a Spanish verb depend on both stem distributions and affix distributions, which are partly independent. So in order to understand paradigm shape as an organizing principle of inflectional systems, we want to model stem and affix distributions jointly.

The second line of research reflects complementary insights and complementary problems. Coming from the literature on inflectional complexity, it consists of work that uses information-theoretic measures, predominantly conditional entropy, to measure the average uncertainty associated with the unobserved form realizing some paradigm cell, given one or more observed forms of the same lexeme (Ackerman et al., 2009; Ackerman and Malouf, 2013; Bonami and Beniamine, 2016; Cotterell et al., 2019; Mansfield, 2016; Parker and Sims, 2020; Sims and Parker, 2016;

[7] The work of Stump and Finkel (Finkel and Stump, 2007, 2009; Stump and Finkel, 2013) is also notable for bridging these two lines of research. They extensively examine how principal parts structure inflectional systems. Defining different notions of principal parts— static, dynamic, and adaptive— allows them to investigate questions of distributional parallelism across lexemes and classes. However, since principal parts are defined set-theoretically, their approach encounters difficulty capturing *partial* parallelism. Ultimately, we take their work as inspiration but we think that our approach has a number of advantages.

[8] Much of this work engages with the concept of a *morphome* (Aronoff, 1994), meaning structure that is irreducibly morphological in nature and autonomous of both syntax and phonology. This issue is beyond the scope of the present paper.

[9] SER 'be' has the largest number, with six distinct stems.

Stump and Finkel, 2013). This work tends to be typological in focus.

The information-theoretic approach has proven popular for quantifying paradigmatic relations in a gradient way. At the same time, this literature has tended to abstract away stem alternations in order to focus on affixes and other 'primary' inflectional exponence (e.g. Ackerman and Malouf, 2013), although there are exceptions (Parker and Sims, 2020). This reflects in part a tendency to rely on hand segmentation of words into morphological exponents, a problematic issue (Beniamine and Guzmán Naranjo, 2021) that we return to below.

More importantly, extending the information-theoretic approach to the task of quantifying paradigm shape turns out to be a challenge because conditional entropy is insufficient by itself to fully capture the regularities that we are interested in. Specifically, since entropy is calculated over surface exponents, it treats identical distributions instantiated by different phonological material (as with the stem alternations in SENTIR and PENSAR vs. MOVER in Table 1) as formally independent. Conditional entropy thus misses abstract generalizations of the sort embodied by Maiden's L- and N-patterns. Ultimately, entropy is appropriate to quantifying what Baerman et al. (2017) call 'allomorphic' inflection class systems, but it does not automatically capture the kinds of generalizations that instantiate 'distributional' systems.

Allomorphic systems are the type of inflection class system familiar to most linguists: classes are defined by inflectional exponents and two lexemes belong to different classes if they are realized by different exponents. In contrast, distributional systems are ones in which two lexemes are realized by the same set of exponents, but these are distributed differently among paradigm cells (Baerman et al., 2017).[10] Class divisions are thus defined by the distribution of exponents, rather than the form of the exponents. Baerman et al. are primarily interested in how inflection class distinctions are instantiated, but from a converse perspective, we observe that the idea of classes based on the distribution of exponents, rather than the phonological form of exponents, also serves to group classes that have different exponents in the same distribution.[11] The insight behind Maiden's L- and N-patterns (and other work on stem organization) is that stem classes are distributional in nature. Conditional entropy as it has been employed in the inflectional complexity literature cannot capture such classes unless the input data to entropy calculations is transformed into a purely distributional representation (a process we refer to below as 'deidentification').

In this paper we seek to bridge the gap between the historical/formal literature on stem space organization and the information-theoretic literature on inflectional complexity and improve on both. We draw on information-theoretic measures developed in the inflectional complexity literature and apply them to investigating the extent to which implicative relations exhibit distributional parallelism across lexemes and classes. As we show below, by doing so we are able to precisely quantify similarity in paradigm shape in a way that is replicable and extendable to new languages. Our methods also take into account both stem and affix distributions. This allows us to capture insights that have predominantly been treated separately in previous work.

4 Methods

We quantify the strength of implicative relations between cells by identifying sets of cells that "confuse" two microclasses in that system—that is, sets among which internal comparisons do not allow precise assignment of a verb to a single microclass. Using entropy, we then compute the degree to which each such set of cells helps to identify the exact inflectional microclass of each verb. We structure these values in a matrix of m microclasses $\times$ n sets of cells, where each entry corresponds to the entropy value associated with a set of cells for a particular microclass. These entropy values provide a quantitative basis for analyzing the inflectional system along multiple organizational dimensions.

To make precise our definition of "confusion," our method relies on segmentation. Given a set of forms of a single lexeme, we identify a *theme*, stem-like material that remains invariant for every form in the set, and a set of *distinguishers*,

[10] One of Baerman et al.'s canonical examples is from Azazulco Otomí, an Oto-Manguean language of Mexico. In verbs, one class is defined by having the suffix *-di* in the first, second, and third person realis incompletive, and another class is defined by having the suffix *-di* in the first and second person realis completive [pp.13,112]. Which cells *-di* shows up in is the only thing distinguishing these two classes.

[11] This is also reflected in the concept of a *distillation*, as found in Stump and Finkel (2013).

set of forms	theme	distinguishers
piensas, pensamos	pensas	i, mo
pienso, piensas, piensa	piens	o, as, a
	piens	o, sa, a

Table 2: Local segmentation examples.

form-specific, affix-like material that vary across the set.[12] See Table 2 for examples using forms of PENSAR.

Although segmentation-based analysis of morphological systems is common in the computational literature, in the morphological literature there is no accepted standard for what constitutes a 'correct' segmentation (Spencer, 2012). The assignment of phonological material to stems vs. to affixes often reflects language-specific traditions of analysis, unclear analytic criteria, and/or theoretical considerations (Taylor, 2008). Furthermore, different segmentation strategies can result in different analyses of inflection class structure (Beniamine et al., 2017). Our use of "local" segmentation (potentially identifying a different theme and distinguishers for each set of forms) rather than "global" segmentation (identifying a single theme/stem for each lexeme) follows Beniamine et al. (2017). They show that this method yields better descriptions of lexemes with stem alternations. In such cases, the alternating characters are included in the theme when only one stem allomorph is included in the set, but in the distinguisher when multiple allomorphs are present (compare rows 1 and 2 in Table 2). Therefore, the analysis of different sets can show both the regularity of the affixes (PENSAR takes the *-as* suffix in 2.SG) and the presence of the alternation.

We can now define a confusion between microclasses: a set of cells confuses two classes if, when the inflected forms for each class are locally segmented, they have identical distinguishers. For instance, [PRS.1SG, PRS.2SG] confuses CANTAR and PENSAR. Local segmentation yields the distinguishers *o*, *as* for both. However, if we added PRS.1PL to the set, it would no longer confuse these two classes. The *-i-* from the stem alternation in PENSAR is now forced into the distinguisher (since it is not shared with the 1PL).

This notion of confusion is important because it enables us to view the classes as *locally* similar even when they are globally different. As shown above, PENSAR and CANTAR are confused by sets which do not show both stem alternants of PENSAR. On the other hand, PENSAR and SENTIR are confused by sets which do not vary the expression of the theme vowel (such as 3SG, 1PL). We can apply the same definition even when the stem is entirely suppletive. Because our definition is based on internal contrasts within sets of cells, it can recognize (for example) that SER 'be' is anomalous among Spanish verbs by virtue of its suppletive preterite (1SG present indicative *soy*, preterite *fui*), but also that, within the set of preterite forms, its conjugation is relatively regular.

Enumerating every set of cells which confuses two microclasses is difficult, since if a large set S confuses two microclasses, each of its exponentially many subsets does as well. We restrict our attention to the *maximal* confusion sets for each pair of microclasses,[13] which, we show below, can be efficiently computed. A set of cells S (size >1) is a maximal confusion set for microclasses A and B if no superset of S also confuses A and B.

Once all maximal confusion sets have been identified, we evaluate each set's predictive power by determining how it groups all the microclasses in the system into mutually confusable partitions. We compute how well the set narrows down the identity of each microclass. If a particular microclass is confusable with k classes on the basis of some set of cells, the remaining uncertainty is $-\log_2 k$ bits. Entropy's usefulness as a quantitative standard is particularly clear in the case of no confusability: if the set uniquely identifies a particular class, the entropy value is zero, indicating that there is no remaining uncertainty about which class the set belongs to.

By applying this process to each maximal confusion set and each microclass, we compute a matrix of entropy values quantifying the distribution of predictive relationships across the inflectional system. For this paper, we analyzed the Spanish verbal inflectional system, using 60 morphosyntactic property sets of 40 verb microclasses (drawn

[12] Ideally, the theme is the longest common subsequence of all the forms; we approximate this NP-hard computation by aligning the forms one at a time using dynamic programming. Each character has an identical insertion cost of 1. Once the theme is identified, we realign each form to the theme and designate the unaligned characters as the distinguisher. Positions of discontinuities are not noted in either the themes or the distinguishers.

[13] Our method computes maximal confusion sets only for pairs of classes. We believe that sets can be computed for larger numbers of classes as well, but leave this for future work.

from Brodsky (2005)).[14] Our method generates 290 maximal sets, for a resulting 40 x 290 matrix of entropy values. Further details of the algorithm are described below.

4.1 Deidentification

The procedure just described highlights differences between microclasses based on both the affixes they take and the distribution of different stem allomorphs within the paradigm. To focus on purely distributional information, we also develop a "deidentified" analysis which abstracts away from the forms of the distinguishers. In this analysis, we replace the individual characters within the distinguishers (which represent affixes, stem alternations and other local variation within the set) with abstract identifiers indicating the positions of identical characters. For instance, the distinguishers *o, as, a* would be represented as α, $\beta\gamma$, β. This enables them to match *o, es, e*, in which the theme vowel has changed but its distribution has not.

To replace the characters with identifiers, we perform a multi-string alignment of the distinguishers within each step. We search for multiway alignments using the A^* algorithm (Russell and Norvig, 2021, p85). Having obtained strings of abstract identifiers, we want to identify confusion sets between microclasses. This requires a slight modification of our confusion definition for the deidentified case: a set of cells confuses two classes if, when the inflected forms for each class are locally segmented, they have deidentified distinguishers with a perfect one-to-one correspondence. This enables us to identify matches between distinguisher sets with different abstract identifiers. For example, α, $\beta\gamma$, β and γ, $\delta\epsilon$, δ do not have identical identifiers but do have a perfect one-to-one correspondence (α:γ, β:γ, γ:ϵ) and therefore comprise a confusion set. We again use the A^* algorithm to search for maximal one-to-one correspondences between sets.

Using the previously described Spanish data, our deidentified method generates 25,239 maximal sets, for a resulting 40 x 25239 matrix of entropy values.

4.2 Algorithmic efficiency

We state above that the maximal confusion sets for each pair of microclasses can be efficiently computed, although the search space contains exponentially many sets. Here, we explain how this can be done. We begin with the intuition that every maximal confusion set must be associated with some theme for each row involved. That is, knowing that a given set of morphosyntactic properties can yield an identical distinguisher set for class 1 and class 2 presupposes the existence of some theme A for class 1 and some theme B for class 2 that produce these distinguishers. Therefore, determining all the possible themes for class 1 and class 2 and finding the largest confusion set for each cross-class pairing of themes will necessarily yield all the maximal confusion sets (along with some non-maximal confusion sets).

Next, we note that themes (longest common subsequences of sets of forms) grow monotonically shorter as more forms are added. Thus, all possible themes for a given microclass can be computed by aligning every pair of forms within it, then aligning the resulting themes until no further themes can be generated. Once all themes for a class have been determined, we compare each theme against each form in the class and find the largest possible set of cells for which that theme is valid; denote this set $S(T)$ for a theme T.

Now, to find confusion sets for a pair of classes I, J, we test every pair of themes A_i and B_j. We take set $S(A_i) \cap S(B_j)$ and test whether it has at least two members, and whether local segmentation of those members actually produces the themes A_i, B_j.[15] All sets that meet this specification are output as potential confusion sets. After all confusion sets are generated for a pair of rows, any confusion sets that are subsets of any others are removed; this ensures that only maximal confusion sets are retained. This method of identifying maximal confusion sets is followed for every pair of classes in the system.

[14] We only look at verbs with full paradigms in this work. Real language learners may not observe every form for every verb, due to their Zipfian distribution (Lignos and Yang, 2018); we do not address the question of learning shapes from this kind of partial data. We also do not address the issue of inflectional defectiveness (paradigmatic gaps) (Albright, 2003), which causes problems for our method even when all forms of a verb are available. Sims (2015) and others argue that gaps are sometimes irreducible morphological objects, including in Spanish verbs (Boyé and Cabredo Hofherr, 2010; Maiden and O'Neill, 2010; O'Neill, 2018a). In this case, it makes sense to treat defective verbs as defining additional microclasses.

[15] Because the intersection may be smaller than the original sets, local segmentation might produce themes which are longer than A_i, B_j, in which case the set is not valid *for this pair of themes*, although it might be output for another pair.

The final step of the algorithm is applying entropy. After we have collected all maximally confusable sets for all microclasses, we generate a matrix of m microclasses $\times$ n maximally confusable sets. For each maximal set, we iterate through the classes; for each class, we determine how many classes it can be confused with based on the forms in the maximal set. Two classes are confusable by a set of forms if it is possible for both classes to generate an identical distinguisher set for those forms; similar logic applies to confusability of three or more classes. The corresponding cells in the matrix are filled with the resulting count values. The entropy is $-\log_2$ of the counts.

5 Results

We visualize our matrices of entropy values using hierarchical clustering and t-SNE analyses (van der Maaten and Hinton, 2008). Though the underlying clusters we discuss in this section are present in both analyses, we focus on the t-SNE results, providing the dendrograms as Appendix A. Figure 1 shows t-SNE visualizations for the maximally confusable sets generated by our primary method, and Figure 2 shows the same for the deidentified case. For our traditional class categorizations, we grouped all the classes Brodsky (2005) deemed "fundamentally irregular" together and then organized the remaining "basically regular" classes into *-ar*, *-er*, and *-ir* groups based on their infinitive forms. We also categorized our classes based on Maiden's alternation patterns, identifying the L-pattern, the N-pattern, and a "modified L-pattern" (a mixed N-pattern and L-pattern) in our data.

These visualizations highlight several key components of paradigm shape in Spanish verbs. Traditional allomorphic class groupings are readily distinguishable in Figure 1, most clearly in the cluster of red *-ar* verbs. By contrast, the *-er* and *-ir* verbs are somewhat interspersed. This comports with the fact that the *-ar* classes have a fairly consistent *a* inflectional suffix across their paradigms, while the *-er* and *-ir* classes demonstrate inconsistency in the realization of an *i* vs. *e* suffix. These clustering structures show that our method is capturing aspects of allomorphic classes delineated by inflectional exponents.

Distributional class groupings are also present. For example, the large swath of classes at the top of Figure 1 has two main clusters. Each consists of both *-er* and *-ir* verbs, so it is clear that the clustering is not driven by inflectional affixes alone. In fact, Maiden's stem alternations explain most of the clustering distinction, as most verbs in the left-hand cluster exhibit the N-pattern (SENTIR, PEDIR, DORMIR, CONSTRUIR, ARGÜIR, OÍR, PERDER, MOVER, DISCERNIR, and ADQUIRIR) while most in the right-hand cluster have the L-pattern (LUCIR, ASIR, CONOCER, COMER, SUBIR, VALER, and SALIR).

Our deidentified approach is able to take this one step further and draw even finer distributional distinctions between classes. The previously mentioned upper-left-hand cluster in Figure 1 splits into two smaller clusters under the deidentified approach in Figure 2. Though the large cluster is united by all of its members having Maiden's N-pattern (except for OÍR, which has the mixed N-pattern and L-pattern), the split into smaller clusters can be explained by another alternation in the preterite. As shown in Table 3, the verbs in the first group (SENTIR, PEDIR, DORMIR, CONSTRUIR, ARGÜIR, and OÍR) all have an alternation in their third person singular and third person plural preterite indicative forms, whereas those in the second group (PERDER, MOVER, DISCERNIR, and ADQUIRIR) have no alternations in the preterite.

It is important to note that the verbs in the first group do not all exhibit the same alternation: SENTIR and PEDIR have *e~i*; DORMIR has *o~u*; and CONSTRUIR, ARGÜIR, and OÍR have *i~y*. This shows a strength of our local segmentation strategy. The *i~y* alternation appears at the boundary between stem and affix, but we are not forced to commit to placing it in one or the other *a priori*; instead, it can be grouped with the other two alternations which occur in comparable positions. This illustrates our method's utility in identifying distributional class structure.

Our method provides a gradient, numerical characterization of structural similarities between the paradigms of Spanish verbs. In doing so, it captures several pre-existing intuitions about the implicative structure of Spanish verbal inflections, including the traditional inflection classes as well as Maiden's distributional classes. Moreover, it also makes finer distinctions which were not explicitly listed in prior work, but which follow from their principles of analysis.

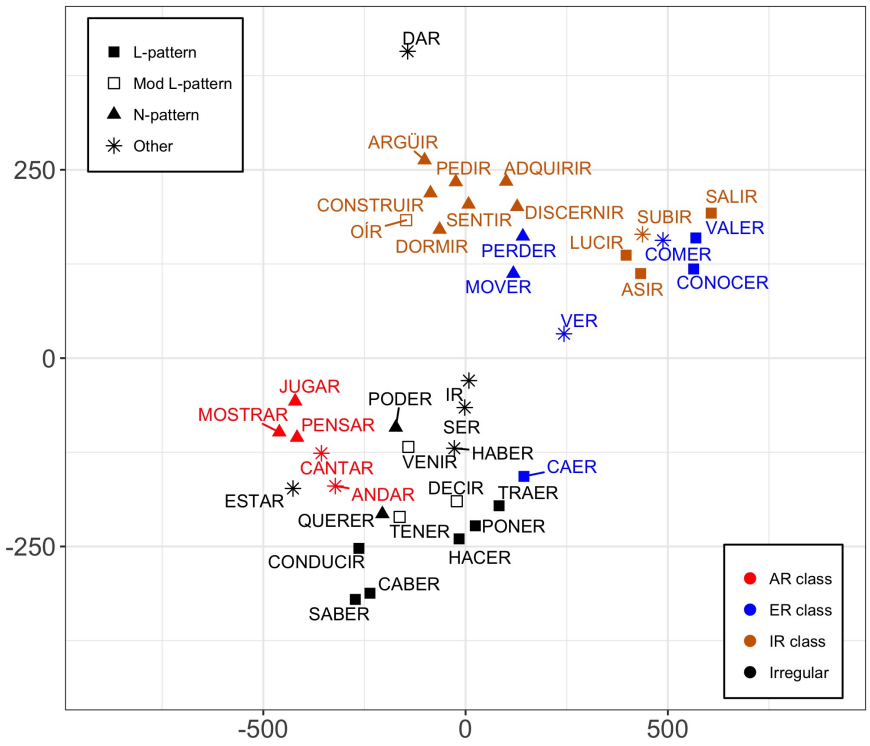

Figure 1: Results of t-SNE analysis based on entropy of maximally confusable sets. Colors show the traditional classes; symbols show the Maiden alternation patterns.

LEXEME	GLOSS	PRET.1SG	PRET.2SG	PRET.3SG	PRET.1PL	PRET.2PL	PRET.3PL
SENTIR	'feel'	sentí	sentiste	sintió	sentimos	sentisteis	sintieron
PEDIR	'ask for'	pedí	pediste	pidió	pedimos	pedisteis	pidieron
DORMIR	'sleep'	dormí	dormiste	durmió	dormimos	dormistes	durmieron
CONSTRUIR	'build'	construí	construiste	construyó	construimos	construisteis	construyeron
ARGÜIR	'argue'	argüí	argüiste	arguyó	argüimos	argüisteis	arguyeron
OÍR	'hear'	oí	oíste	oyó	oímos	oísteis	oyeron
PERDER	'lose'	perdí	perdiste	perdió	perdimos	perdisteis	perdieron
MOVER	'move'	moví	moviste	movió	movimos	movistes	movieron
DISCERNIR	'discern'	discerní	discerniste	discernió	discernimos	discernisteis	discernieron
ADQUIRIR	'acquire'	adquirí	adquiriste	adquirió	adquirimos	adquiristeis	adquirieron

Table 3: Preterite alternation that leads to the cluster split observable in Figure 2

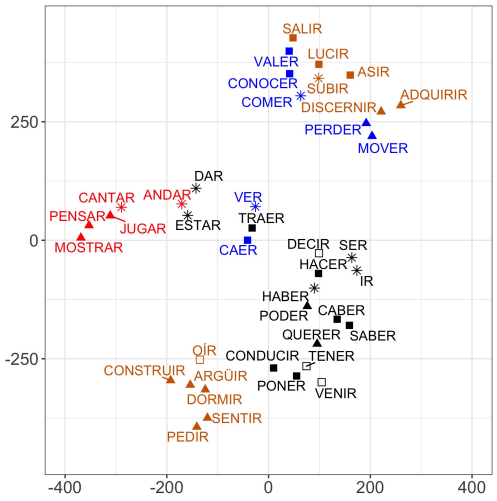

Figure 2: Results of t-SNE analysis based on entropy of maximally confusable sets in the deidentified condition. Coding of colors and shapes is the same as in Figure 1.

6 Conclusion and future work

In this paper, we present a method for precisely quantifying paradigm shape in a replicable, extendable way. Bridging the gap between formal literature on stem space organization and information-theoretic literature on inflectional complexity, this work models the joint effect of stem alternations and suffixes on the implicative relations holding among inflected forms. We show that our model captures important allomorphic and distributional class insights in Spanish verbs. In the future, we would like to extend our notion of confusion sets beyond the pairwise case. We would also like to develop a systematic way of substantiating the precise structures captured by our approach; though our analysis goes some length toward showing what particular aspects of structure the method is sensitive to, a rigorous substantiation is beyond the current scope of our work. Finally, we aim to use our method to analyze other Romance languages and to trace how shape has impacted the historical development of Romance verbs.

Acknowledgements

We thank Bob Levine and the members of the Autumn 2019 Undergraduate Research Seminar at Ohio State for their feedback on a preliminary version of this work, as well as three anonymous reviewers.

References

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

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The Low Conditional Entropy Conjecture. *Language*, 89(3):429–464.

Adam Albright. 2003. A quantitative study of spanish paradigm gaps. *Proceedings of the West Coast Conference on Formal Linguistics*, 22:1–14.

Mark Aronoff. 1994. *Morphology by itself: Stems and inflectional classes*. MIT Press.

Janice Aski. 1995. Verbal suppletion: An analysis of Italian, French, and Spanish *to go*. *Linguistics*, 33:403–432.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2017. *Morphological complexity*. Cambridge University Press.

Sacha Beniamine, Olivier Bonami, and Benoît Sagôt. 2017. Inferring inflection classes with description length. *Journal of Language Modelling*, 5:465–525.

Sacha Beniamine and Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. *Proceedings of the Society for Computation in Linguistics*, 4.

Olivier Bonami and Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Olivier Bonami and Gilles Boyé. 2002. Suppletion and dependency in inflectional morphology. In Frank van Eynde, Lars Hellan, and Dorothee Beermann, editors, *Proceedings of the 8th International Conference on Head-Driven Phrase Structure Grammar*, pages 51–70. CSLI.

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

Gilles Boyé and Patricia Cabredo Hofherr. 2010. Defectiveness as stem suppletion in French and Spanish verbs. In Matthew Baerman, Greville G. Corbett, and Dunstan Brown, editors, *Defective paradigms: Missing forms and what they tell us*, pages 35–52. Oxford University Press, in coordination with British Academy Press.

David Brodsky. 2005. *Spanish verbs made simple(r)*. University of Texas Press.

John Butt, Carmen Benjamin, and Antonia Moreira Rodríguez. 2019. *A new reference grammar of modern Spanish, 6th edition*. Routledge.

Greville G. Corbett. 2009. Canonical inflection classes. In Fabio Montermini, Gilles Boyé, and Jesse Tseng, editors, *Selected proceedings of the 6th Décembrettes*, pages 1–11. Cascadilla.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Raphael Finkel and Gregory T. Stump. 2007. Principal parts and morphological typology. *Morphology*, 17:39–75.

Raphael Finkel and Gregory T. Stump. 2009. Principal parts and degrees of paradigmatic transparency. In James P. Blevins and Juliette Blevins, editors, *Analogy in grammar: Form and acquisition*, pages 13–53. Oxford University Press.

Borja Herce. 2020. 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*, 30:90–115.

Andrew Hippisley. 1998. Indexed stems and Russian word formation: A network-morphology account of Russian personal nouns. *Linguistics*, 36:1093–1124.

Andrew Hippisley, Marina Chumakina, Greville Corbett, and Dunstan Brown. 2004. Suppletion: Frequency, categories and distribution of stems. *Studies in Language*, 28:387–418.

Matthew Juge. 1999. On the rise of suppletion in verbal paradigms. In *25th Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, pages 183–194. Berkeley Linguistics Society.

Constantine Lignos and Charles Yang. 2018. Morphology and language acquisition. In Andrew Hippisley and Gregory T. Stump, editors, *Cambridge handbook of morphology*, pages 765–791. Cambridge University Press.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Martin Maiden. 2004. When lexemes become allomorphs: On the genesis of suppletion. *Folia Linguistica*, 38:227–256.

Martin Maiden. 2005. Morphological autonomy and diachrony. In Geert Booij and Jaap van Marle, editors, *Yearbook of morphology 2004*, pages 137–175. Springer.

Martin Maiden. 2009. From pure phonology to pure morphology: The reshaping of the Romance verb. *Recherches linguistiques de Vincennes*, 38:45–82.

Martin Maiden. 2011. Morphomes and 'stress-conditioned allomorphy' in Romansh. In Martin Maiden, John Charles Smith, Maria Goldbach, and Marc-Olivier Hinzelin, editors, *Morphological autonomy: Perspectives from Romance inflectional morphology*, pages 36–50. Oxford University Press.

Martin Maiden and Paul O'Neill. 2010. On morphomic defectiveness: Evidence from the Romance languages of the Iberian Peninsula. In Matthew Baerman, Greville G. Corbett, and Dunstan Brown, editors, *Defective paradigms: Missing forms and what they tell us*, pages 103–124. Oxford University Press, in coordination with British Academy Press.

John Mansfield. 2016. Intersecting formatives and inflectional predictability: How do speakers and learners predict the correct form of Murrinhpatha verbs? *Word Structure*, 9:183–214.

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

Paul O'Neill. 2018a. Near-synonymy in morphological structures: Why Catalans can abolish constitutions but Portuguese and Spanish speakers can't. *Languages in Contrast*, 18:6–34.

Paul O'Neill. 2018b. Velar allomorphy in Ibero-Romance: Roots, endings and clashes of morphomes. In Miriam Bouzouita, Ioanna Sitaridou, and Enrique Pato, editors, *Studies in historical Ibero-Romance morpho-syntax*, pages 13–46. John Benjamins.

Jeff Parker and Andrea D. Sims. 2020. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of Russian nouns. In Francesco Gardani and Peter Arkadiev, editors, *The complexities of morphology*, pages 23–51. Oxford University Press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ralph Penny. 1972. Verb-class as a determiner of stem-vowel in the historical morphology of Spanish verbs. *Revue de Linguistique Romane*, 36:342–359.

Stuart Russell and Peter Norvig. 2021. *Artificial intelligence: A modern approach*. Pearson.

Andrea D. Sims. 2015. *Inflectional defectiveness*. Cambridge University Press.

Andrea D. Sims and Jeff Parker. 2016. How inflection class systems work: On the informativity of implicative structure. *Word Structure*, 9(2):215–239.

Andrew Spencer. 2012. Identifying stems. *Word Structure*, 5:88–108.

Gregory T. Stump and Raphael A. Finkel. 2013. *Morphological typology: From word to paradigm*. Cambridge University Press.

Catherine Taylor. 2008. Maximising stems. In Miltiadis Kokkonidis, editor, *Proceedings of LingO 2007*, pages 228–235. Faculty of Linguistics, Philology and Phonetics, University of Oxford.

Max Wheeler. 2011. The evolution of a morphome in Catalan verb inflection. In Martin Maiden, John Charles Smith, Maria Goldbach, and Marc-Olivier Hinzelin, editors, *Morphological autonomy: Perspectives from Romance inflectional morphology*, pages 183–209. Oxford University Press.

Wolfgang U. Wurzel. 1989. *Inflectional morphology and naturalness*. Kluwer.

Appendix A: Dendrograms

Although we find our similarity measurements are most interpretable via the T-SNE visualizations in the main paper, T-SNE is non-deterministic and can sometimes erroneously group points that are not similar in the underlying space. Thus, we also present dendrograms in which the points are clustered using the complete method from Scikit Learn (Pedregosa et al., 2011). These show the same clustering structures discussed in the main paper. Figure 3 shows the main results and Figure 4 the deidentified results.

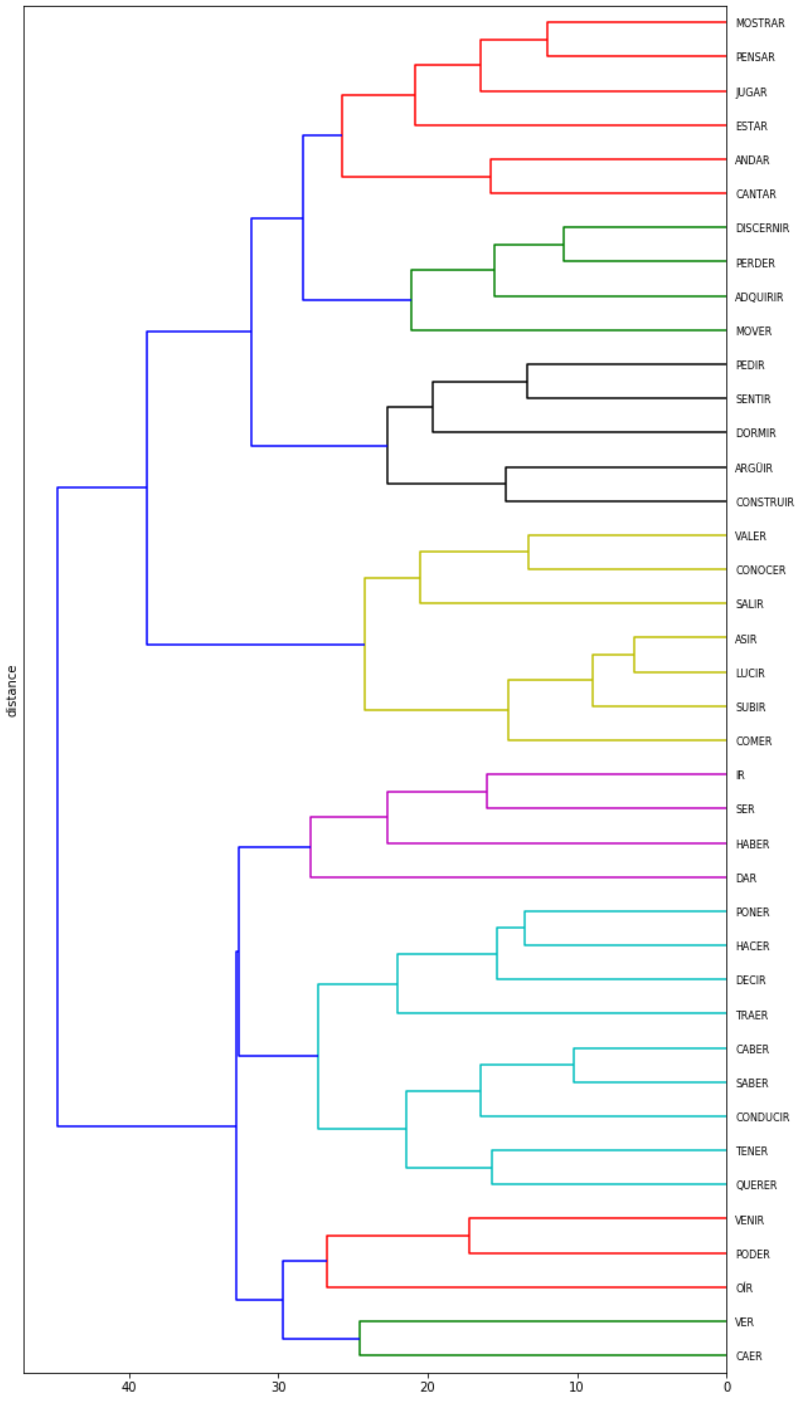

Figure 3: Results of hierarchical clustering analysis based on entropy of maximally confusable sets.

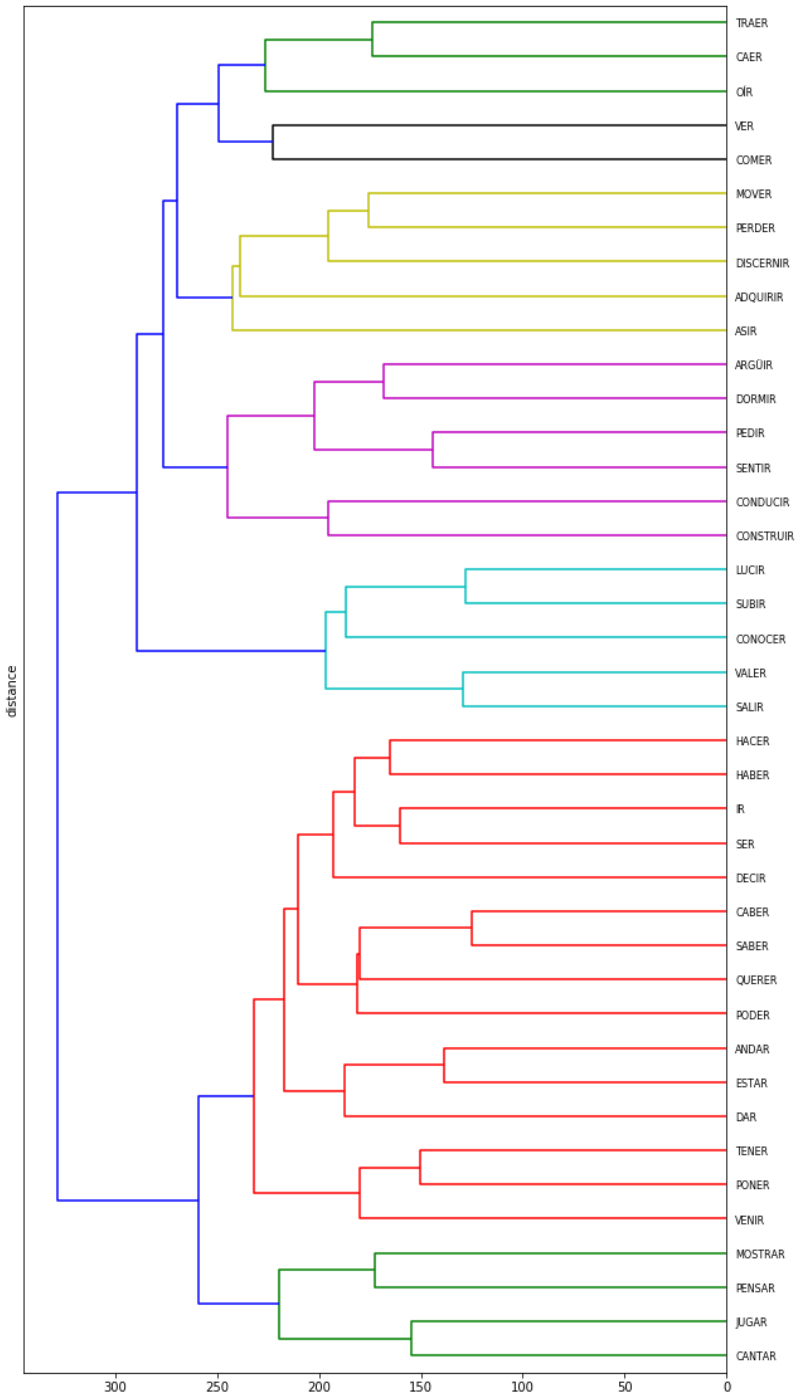

Figure 4: Results of hierarchical clustering analysis based on entropy of maximally confusable sets in the deidentified condition.