Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09721-3

PROJECT NOTES

VeLeRo: an inflected verbal lexicon of standard Romanian and a quantitative analysis of morphological predictability

Borja Herce[1] · **Bogdan Pricop**[1]

Accepted: 10 January 2024

Abstract
This paper presents VeLeRo, an inflected lexicon of Standard Romanian which contains the full paradigm of 7297 verbs in phonological form. We explain the process by which the resource was compiled, and how stress, diphthongs and hiatus, consonant palatalization, and other relevant issues were handled in phonemization. On the basis of the most token-frequent verbs in VeLeRo, we also perform a quantitative analysis of morphological predictability in Romanian verbs, whose complexity patterns are presented within the broader Romance context.

Keywords Romanian · Paradigm · Verb · Inflected lexicon · Morphological predictability · Entropy

1 Introduction

Morphological predictability relations in paradigms have been explored for a long time. In classical language pedagogy, principal parts allowed learners of Latin, for example, to accurately predict any inflected form of a lexeme's paradigm from a small subset of word forms (see Finkel & Stump, 2009). Morphologists have long known, thus, that "one inflection tends to predict another" (Matthews, 1991: p. 97), so that Lat. NOM.SG *templum* predicts GEN.SG *templī*, NOM.PL *templa*, etc. Computational and theoretical advances in the last few decades, however, have enabled more exhaustive, systematic, and faster analyses of inflectional systems than ever before.

In recent years, the complexity of inflectional systems has started to be analyzed more and more often from the perspective of the Paradigm Cell Filling Problem (PCFP, Ackerman et al., 2009), which is the name given to the challenge that

✉ Borja Herce
borjaherce@gmail.com

Bogdan Pricop
bogdan.pricop@uzh.ch

[1] University of Zurich, Zurich, Switzerland

Published online: 23 March 2024

speakers of languages with inflection classes face to produce essentially any form of any lexeme, no matter how infrequent the lemma or paradigm cell. Although high frequency forms could be stored in memory and learned by rote, given the rank frequencies of words in natural languages (Zipf, 1932), with a long tail of extremely low or hapax (i.e. single-occurrence) words, most inflected words must be built online, with their forms predicted by language users on the basis of other more frequent forms in their paradigm, and/or on the basis of the equivalent forms or patterns as encountered in other more frequent lexemes.

The methods to explore these morphological predictability relations within inflectional or derivational paradigms have been expanding in recent years. Traditional philological- comparative qualitative methods (e.g. Malkiel, 1966, Maiden, 1992, 2018, O'Neill, 2018, Esher, 2022, Herce, 2022) continue to be pursued alongside quantitative computational ones involving Set Theory, Graph Theory, Machine learning, and Information Theory (e.g. Ackerman & Malouf, 2013; Beniamine, 2018; Elsner et al., 2019; Malouf, 2017; Sims, 2020; Stump & Finkel, 2013). Entropy (a measure of uncertainty, Shannon, 1948) is the key notion of Information Theory. Applied to paradigmatic forms and the PCFP, conditional entropy can be used to measure the (un)predictability of one form given another. Software has also been developed to calculate different complexity measures automatically over whole systems: Beniamine's Qumin (https://sacha.beniamine.net/software/qumin/), Finkel's Principal Parts Analyzer (https://www.cs.uky.edu/~raphael/linguistics/analyze.html), and Sims' Inflectional Networks scripts (https://github.com/sims120/inflectional-networks) among others.

Alongside these methodological, theoretical, and software solutions, progress has also been made on the side of resources and databases (see e.g. Kirov et al., 2018). Most crucial to the PCFP are inflected lexicons, where all inflected forms of hundreds or thousands of lemmas are listed, preferably in phonological form. This is the data that is required for the quantitative investigation of morphological predictability and complexity. Most underdocumented at present, and hence most urgent to describe, are inflectional systems of non-Indo-European and non-WEIRD (non-Western European Industrialized Rich and Democratic, see Henrich et al., 2010), low-resource languages (Malouf et al., 2020). Better (i.e. larger, phonemized, well-curated) resources, and comparable morphological predictability analyses, however, are also needed for many national standard Indo-European languages. Focusing on Romance, arguably the language family where more attention has been paid to paradigmatic morphology, we have some family-wide but comparatively small inflected lexicons (see Maiden et al., 2010; Beniamine et al., 2020), as well as large lexicons and PCFP-analyses for some of the major languages in the family [namely French (Bonami et al., 2014), Latin (Pellegrini & Passarotti, 2018), Italian (Pellegrini & Cignarella, 2020), Portuguese (Beniamine et al., 2021), and Spanish (Herce, 2023)]. No such inflected lexicon and analysis exists, however, for Romanian (see Diaconescu et al., 2015, and Lőrincz et al., 2022. for lexicons not specialized in the coverage of inflected forms). This is the purpose of the present paper. Section 2 will explain the creation of a verbal inflected lexicon for Romanian in phonological form (VeLeRo), annotated with lemma and cell frequencies. Section 3 presents a morphological predictability analysis of Romanian verbal inflection on the basis of this

resource, and discusses the results briefly, particularly how they compare to extant analyses of other Romance languages. Section 4 summarizes the main highlights of the paper and proposes avenues for future research.

2 Building VeLeRo

VeLeRo is an inflected lexicon of Romanian verbs in phonological form. It is based on Barbu's (2008) lexical database RoMorphoDict, which contains the orthographical inflected forms, lemmas and morpho-syntactic descriptions of around 700,000 Romanian words. The verbal lemmas and their inflected forms were extracted from this dataset and transcribed phonologically using Epitran (https://github.com/dmort27/epitran). Epitran is a massively multilingual, rule based G2P (grapheme to phoneme) system with support for 61 languages and distributed as open source software (a Python library) under an MIT license (Mortensen et al., 2018). Epitran's rule-based conversion provided a broad phonemic transcription, but certain aspects required further refinement. Thus, the initial conversion was only used as the starting point for later adjustments. Overall, 51% of the words were modified after the G2P conversion to represent the Romanian phonological system more accurately (e.g. regarding diphthongization, palatalization, and the rest of topics discussed in the upcoming sections). Using the Ratcliff/Obershelp string matching algorithm provided by the difflib Python library, there was a 93% match between the initial Epitran phoneme conversion and the final version of the lexicon.

Our inflected lexicon of Romanian verbs (which can be found freely available online at https://osf.io/kqrjg/?view_only=f1583503953d45028d3bb85a2e1c6d01) adopts a wide format, with each row corresponding to a lexeme (identified through the infinitive in orthographic form (e.g. *face* 'do'), and columns representing all 39 forms of every lexeme in phonological form (e.g. inf *fáʧe*, ind.prez.1sg *fák*, ind. prez.2sg *fáʧʲ*, etc.).[1] The following sections outline the procedures and decisions involved in the phonemization. A full inventory and description of phonemes can also be found online.

2.1 Stress assignment

In Romanian, stress can be oxytonic (final syllable), paroxytonic (penultimate syllable) or proparoxitonic (antepenultimate syllable) and is not marked orthographicaly. Some authors (Chitoran, 1996) have claimed that stress is highly predictable, depending, among others, on the part of speech of the lexical item, but some others (Dindelegan & Maiden, 2013), have claimed that stress is largely unpredictable and

[1] Two small deviations have been adopted from standard IPA usage to facilitate the computational analysis of forms. A widespread one in these resources is the indication of stress with an acute accent over the stressed vowel (i.e. *fáʧe*), rather than with a stress character before the stressed syllable (i.e. *'faʧe*). The other is the transcription of non-syllabic /e/ and /o/, generally transcribed as e̯ and o̯ respectively in standard IPA practice, as E and O instead. Both conventions are aimed at facilitating computational use.

highly mobile, especially in verbal inflections. RoMorphoDict was used as the point of departure for building the present resource because it indicates stress consistently on polysyllabic verbs. For consistency (i.e. to avoid spurious morphological contrasts), we added stress markers on monosyllabic verb forms as well (on the only vowel when a word had only one, or on the most open vowel in case the word contained a diphthong **or triphthong**).

2.2 Diphthongs vs hiatuses

Another important point when it comes to Romanian phonology is the representation of the tautosyllabic and heterosyllabic sequences of vowels (i.e. diphthongs/triphthongs and hiatuses). According to Chitoran (2002) Romanian has two non-controversial diphthongs, namely /e̯a/ and /o̯a/. Apart from these, the glides /j/ and /w/ can combine with most vowels to create additional ones. For the creation of this lexicon, we adopted Chitoran's (2002) treatment of diphthongs, where /j/ and /w/ have phonemic status and can be predicted by looking at syllable boundaries. Syllable boundary contrasts are not handled by Epitran, so these had to be encoded manually. For this, we used Barbu's (2008) RoSyllabiDict database as a reference for the syllabification of verbs and made the necessary adjustments by hand. For example, in words such as "a.ban.do.nea.ză", the sequence "ea" is homosyllabic, and was hence coded as a diphthong (e̯a). On the other hand, in words like "a.gre.a.se", were "e" and "a" are heterosyllabic, "ea" was transcribed as a sequence of vowels /ea/. The lexicon contains a total of 16 diphthong and 7 triphthong sequences comprising different glide—vowel combinations plus /e̯a/ and /o̯a/.

2.3 Diphthong reduction

Despite varying or inconsistent phonemic transcription practices elsewhere (e.g. in Maiden et al., 2010's Oxford Database of Romance Verb Morphology), sequences that involve postalveolars (e.g. /ʃ/) followed by /j/ or non-syllabic /e̯/ have been uniformly transcribed without this second segment here (e.g. /ziʧám/, rather than /ziʧe̯ám/, for *ziceam* 'say.1SG.IPF.IND'). This is justified by (i) the absence of an audible front vowel in these sequences, (ii) the absence of minimal pairs based on these sequences (for example /ʧe̯a/ vs /ʧja/ vs /ʧa/), and (iii) by the phonemic transcription conventions in parallel cases in related Romance languages, for example Italian /diʧámo/, rather than *diʧjámo for *diciamo* 'say.1PL.PRS.IND', or Spanish /riɲó/ rather than *riɲjó for *riñó* 'scold.3SG.PST.IND' (note that in this latter case the standard spelling reflects pronunciation accurately).

2.4 Palatalization

Palatalization is a prevalent feature of Romanian phonology and generally uncontested by most grammars. According to Chitoran (2002: p. 173), palatalization in Romanian can be phonologically or morphologically conditioned. The former refers to those instances where a change in articulation occurs automatically in a given

phonetic environment (i.e. is allophonic). The latter refers to the cases where this environment has disappeared, thus leaving the palatalized consonant not predictable from its phonetic environment (i.e. phonemic). An instance of the first type is the palatalization of velars like /k/ and /g/ before a front vowel, where they are realized as [c] and [ɟ] (Dindelegan & Maiden, 2013) or [kʲ] and [gʲ] (Chitoran, 2002) depending on the source. Because this inflected lexicon is aimed at capturing phonemic representations, this automatic allophonic palatalization has not been represented when morphologically inconsequential.[2] Only the latter type of palatalization (i.e. phonologized, unpredictable), thus, is relevant in the context of this inflected lexicon. The clearest case of this type is represented by the (orthographic) ending – i occurs in the second person (singular and plural) in verbs, and often indicates a palatalization of a word-final consonant rather than a full word-final vowel /i/. Epitran, again, does not account for this, so modifications were necessary (e.g. to transcribe the form *dați* 'give.ind.prez.2p' as /dáʦʲ/ rather than /dáʦi/). Regarding this palatalization, the choice was made to distinguish between post-alveolar (e.g. /ʃ/) and post-alveolar palatalized (/ʃʲ/) phonemes. Although there is some research suggesting that these are extremely close acoustically (see Spinu et al., 2012, 2019), most speakers appear to be able to distinguish the two sounds (Spinu, 2018). Thus, a form like *ziseși* "say.2SG.PST" was transcribed as /ziséʃʲ/. A full vowel /i/ (also /u/ in the 1SG.PRS) has been preserved in Romanian pronunciation, and hence transcribed as 'I' here, only after segment sequences of "*muta cum liquida*" (e.g. /íntru/ 'I enter', /íntri/ 'you enter', vs /egzíst/[3] 'I exist', /egzíʃtʲ/ 'you exist').

2.5 Resource overview

After all these steps, we arrived at a consistent phonemization of the complete paradigms of 7297 Romanian verbs, for a total of 284,583 word forms. This is comparable (see Table 1) to the size of extant inflected lexicons in phonological form from other national standard Romance languages. Although this feature will not be used in the second part of this paper, our resource also includes overabundant (Thornton, 2012) or substandard/dialectal forms (e.g. /fúrəm/ occurs, alongside standard /fusérəm/ as the ind.perf.1pl form of the verb *fi* 'be'). Given the importance of usage frequency for morphological learnability, predictability, and the PCFP, this information has also been supplied. The frequencies of lemmas in the corpus CoRoLa (Tufiș et al., 2019), a digital lemmatised corpus with over 1 billion words, have been added to VeLeRo and pattern in the way illustrated in the left-hand panel of Fig. 1, with 9 verbs above 1 million tokens, 152 above 100,000, 727 above 10,000, 2146

[2] Automatic phonological processes can certainly affect morphological predictability relations. If two stressed vowels (e.g. /o/ and /u/) merge into the same realization when unstressed (e.g. both become /u/), this does not allow speakers to predict the stress vowel from the unstressed one. These neutralizations need to be, and have been, represented in phonemic transcription.

[3] The letter 'x' corresponds usually to /ks/ but sometimes to /gz/. This is not predictable from the phonological environment (see Dindelegan and Maiden 2013: p. 49), so 14 verbs have been manually amended as having the /gz/ pronunciation: *exista, exercita, executa, examina, exersa, exemplifica, exila, coexista, reexamina, exaspera, exulta, exubera, preexista,* and *exotiza*.

Table 1 Size of VeLeRo and comparable inflected lexicons in other Romance varieties

Language	Reference	Lemmas	Word forms
Latin	Pellegrini and Passarotti (2018)	3348	850,392
French	Bonami et al., (2014)	4991	253,174
Italian	Pellegrini and Cignarella (2020)	2053	108,809
Portuguese	Beniamine et al., (2021)	4987	324,214
Spanish	VeLeSpa	6554	412,839
Romanian	VeLeRo, this paper	7297	284,583

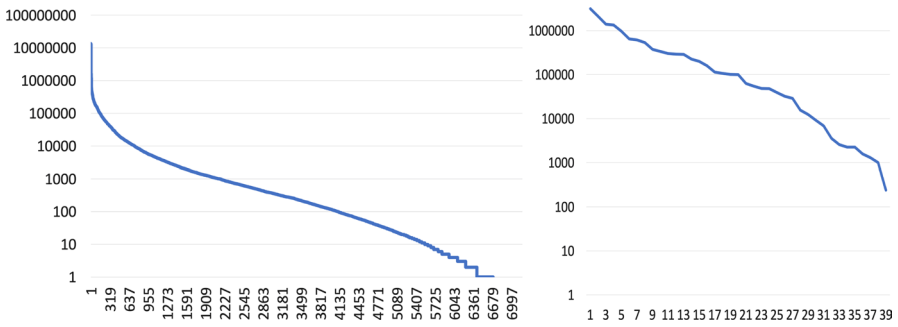

Fig. 1 Rank frequency of the 7296 verbs in VeLeRo (left) and of its 39 cells (right)

above 1000, 4109 above 100, 5601 above 10, and 6688 verbs with at least 1 token in CoRoLa, while 609 verbs from our dataset, are completely unattested in that corpus. The frequency of paradigms cells (based also on CoRoLa)[4] is shown through the rank order plot in the right-hand panel of Fig. 1. These also vary widely between the 3,152,827 tokens of the ind.prez.3sg (the most frequent value) and the 240 of the ind.perf.2pl (the least frequent one).

3 A quantitative analysis of the PCFP in Romanian verbal inflection

To keep computational times within reasonable limits, and to allow for comparability with extant quantitative analyses of the PCFP in other Romance languages, we decided to select for further analysis those (3564) verbs with 200 or more tokens in CoRoLa. This measure/threshold is justified due to two main reasons. The first is that many of the lower frequency verbs are unknown to even highly-educated native

[4] Note that the frequencies of syncretic forms (e.g. 1SG.IND.IPF and 1PL.IND.IPF, as in /publikám/) are not distinguished in CoRoLa. Because we want to obtain a separate measure for each, we estimated individual paradigm-cell frequencies from the proportions observed between comparable person-number values in other tenses where these are not syncretic (e.g. in the PRS.IND: públik vs publikám).

Table 2 Extracted morphological alternations between two cells in eight verbs (A)

Lemma	gloss	ind.prez.2sg	imper.2sg	(ind.prez.2sg, imper.2sg)
da	'give'	dáj	dă	áj ⇌ ă
lua	'take'	jéj	já	éj ⇌ á
forma	'form'	formézʲ	forme̯ázə	ézʲ ⇌ e̯ázə
facilita	'ease'	fatʃilitézʲ	fatʃilite̯ázə	ézʲ ⇌ e̯ázə
duce	'carry/lead'	dútʃʲ	dú	tʃʲ ⇌
comunica	'comunicate'	komúnitʃʲ	komúnikə	tʃʲ ⇌ kə
împărți	'share'	împártsʲ	împárte	tsʲ ⇌ te
păcăli	'fool'	pəkəléʃtʲ	pəkəléʃte	tʲ ⇌ te

speakers of Romanian, and hence probably do not really form part of the average acquired inflectional system of the language, despite the presence of these verbs (and many others) in dictionaries and grammars that are understandably aimed at exhaustivity. The second is that this number of verbs is closer to the average number of items analyzed with identical methods in the extant literature on other Romance languages (see Bonami et al., 2014, Pellegrini & Passarotti, 2018, Pellegrini & Cignarella, 2020, Beniamine et al., 2021, and Herce, 2023), which will enable us to draw more meaningful cross-linguistic comparisons.

After the mentioned exclusions (which included nonstandard overabundant forms), all remaining verbs and forms were analyzed in Qumin (Quantitative Modelling of Inflection, Beniamine, 2018). This is a set of Python scripts that automatically extracts morphological alternations between all possible pairs of word forms in all lemmas. Due to the sheer number of combinations, 1482 (= 39 × 38) per verb, this is a task that can only be performed computationally. The algorithm finds maximally generalizable morphological alternations,[5] and derives a knowledge of which verbs show the same contrasts (i.e. belong to the same inflection class) and which have different ones (i.e. belong to different classes).

Consider, as an illustrative example in Table 2, some of the ways in which the 2SG present indicative can differ in Romanian from the 2SG imperative. While in some verbs (*forma* and *facilita*) these forms differ in identical ways (/ézʲ/ in the former value has to be replaced by /e̯ázə/ to form the latter and vice versa), these forms contrast in different ways in other verbs (e.g. *duce* adds tʃʲ to the 2SG imperative to form the 2SG present indicative, but the same rule does not apply to derive the same form in *comunica*). For the purposes of this pair of cells, thus, the former verbs (i.e. *forma* and *facilita*) belong to the same class, while the latter verbs (i.e. *duce* and *comunica*) belong to different morphological classes. To aid with sequence-to-sequence alignment, and the interpretability of morphological alternations, the Qumin algorithm also makes use of distinctive phonological features. A separate

[5] For more detailed explanation of how the alternations are identified, for example when multiple descriptions are possible, (see Beniamine et al., 2021).

Table 3 Extracted morphological alternations between two cells in eight verbs (B)

Lemma	gloss	ind.imperf.1sg	ind.imperf.2sg	(‘ind. imperf.1sg, ind. imperf.2sg)
da	‘give’	dədẹ́ám	dədẹ́áj	m ⇌ j
lua	‘take’	luám	luáj	m ⇌ j
forma	‘form’	formám	formáj	m ⇌ j
facilita	‘ease’	faʧilitám	faʧilitáj	m ⇌ j
duce	‘carry/lead’	duʧám	duʧáj	m ⇌ j
comunica	‘communicate’	komunikám	komunikáj	m ⇌ j
împărți	‘share’	ɨmpərʦẹ́ám	ɨmpərʦẹ́áj	m ⇌ j
păcăli	‘fool’	pəkəlẹ́ám	pəkəlẹ́áj	m ⇌ j

file needs to be supplied which defines the language’s phonemes and their features (e.g. + or − voiced, + or − nasal, + or − velar, etc.). This file can be found online, along with VeLeRo itself.

As in the eight illustrative word-form pairs in Table 2, patterns of morphological alternations are extracted for all word-form pairs of all verbs in our sample (1482 × 3564 = 5.3 million alternations). This is a demanding computational process that can last several hours. After they are extracted, the patterns can be inspected for quality control (making sure, for example, that infrequent or exceptional patterns are not due to mistakes or inconsistencies in either the original inflected lexicon or its subsequent phonemization). The extracted patterns also constitute the basis for various other scripts and functions within Qumin that allow to calculate further measures like conditional entropies, group inflection classes, etc.

In the extant literature on the PCFP in Romance and beyond, the analysis of morphological predictability within paradigms often starts with a presentation of which values or word forms are mutually interpredictable with complete certainty. Thus, although the morphological relationship between some cells (e.g. ind.prez.2sg and imper.2sg in Table 2) is a heterogeneous one, in the sense that it varies unpredictably from verb to verb, the morphological difference between other cells is the same across all verbs. This is the case, for example, of the ind.imperf.1sg and ind. imperf.2sg. As Table 3 shows, the former can be reliably transformed into the latter by replacing a word-final /m/ with /j/, and, conversely, the latter can be transformed into the former by changing this final /j/ into /m/.

These paradigmatic domains of interpredictability, comparable to the notions of ‘stem space’ (see Montermini & Bonami, 2013) or ‘distillations’ (Stump & Finkel, 2013), provide a first measure of the complexity of an inflectional system. From the 1482 pairs of cells in a Romanian verbal paradigm, 236 (15.9%) involve no uncertainty (i.e. they have a conditional entropy of zero). In terms of concrete paradigm cells (see Fig. 2), the 39 cells of the Romanian verbal paradigm can be classified into 14 areas of interpredictability.

As Fig. 2 shows, many of these areas (Z1, Z5, Z6, Z12, Z13, Z14) are single-cell ones (e.g. Z1 is the imper.2sg) and hence trivial "areas" to some extent, since

	1SG	2SG	3SG	1PL	2PL	3PL
imper.	-	Z1	-	-	Z2	-
ind.prez.	Z3	Z4	Z5	Z2	Z2	Z6
conj.prez.	Z3	Z4	Z7	Z2	Z2	Z7
ind.imperf.	Z8	Z8	Z8	Z8	Z8	Z8
ind.perf.	Z9	Z9	Z10	Z11	Z11	Z11
ind.mmperf.	Z10	Z10	Z10	Z10	Z10	Z10
inf	Z12					
ger	Z13					
part	Z14					

Fig. 2 Areas of morphological interpredictability in Romanian verbal inflection

they merely indicate that some cells are not interpredictable with any other cell. A similar case is the one represented by those forms which contrast in values (e.g. ind.prez.1sg and conj.prez.1sg) but never in their form (*dáw dáw, jáw jáw, forméz forméz, fafilitéz fafilitéz, dúk dúk*, etc. for the verbs in Table 2). Systematic syncretisms like this account for areas Z3, Z4, and Z7. Remaining areas (i.e. Z2, Z8, Z9, Z10, and Z11) are the ones based on predictable morphological alternations like the one in Table 2).

Notable commonalities can be identified between Romanian verbal inflection and that of the other major Romance languages analyzed with this same methodology to date (Beniamine, 2018, Pellegrini & Cignarella, 2020, Beniamine et al., 2021, and Herce, 2023). The number of interpredictability areas (14 in Romanian, vs 15 in Italian [and Latin], 14 in Spanish and French, and 12 in Portuguese), and their distribution (e.g. most areas in the present indicative) are very similar to those found in other Romance languages. As in all other Romance languages analyzed so far, the 1SG present indicative, and the past participle(s) constitute areas of their own. Some other aspects are shared with most but not all other Romance languages, for example, the fact that the 2SG imperative is also a one-cell area of its own (shared with all of Romance except Portuguese), or the fact that all the imperfective indicative cells constitute another area together, to the exclusion of all other cells (shared with all except French).

Our results also show, of course, some differences to the patterns found in other Romance languages. A somewhat trivial one concerns the raw number of values a verb can inflect for, which is less in Romanian than in the other major Romance languages, mainly due to the absence of the synthetic future and conditional tenses, which did not emerge outside Western Romance (i.e. Portuguese, Spanish, French, Italian). In the cognate persons and tenses, another well-known difference is the one derived from the fact that the morphological subjunctives of the first and second person have been replaced by the corresponding indicative forms. This has generated a morphological overlap between present indicative and subjunctive not seen generally in other Romance languages.

Table 4 Presence (1) or absence (0) of morphological properties across Romance languages and Latin (left), and between-language Hamming distances (right)

	2SG.PRS.IND=3SG.PRS.IND	2SG.IMP=2SG.PRS.IND	3SG.PRS.IND=3PL.PRS.IND	1PL.PRS.IND=INF	1PL.PRS.IND=2PL.PRS.IND	1SG.PST.IND=2SG.PST.IND	3PL.PST.IND=3PL.PLUP.SBJV	1PL.PRS.IND=1PL.PRS.SBJV	1SG..PRS.IND=1SG.PRS.SBJV	3SG.IPF.IND=1PL.IPF.IND
Romanian	0	0	0	0	1	1	0	1	1	1
Spanish	1	0	1	1	1	0	1	0	0	1
French	1	0	0	0	1	1	1	0	0	0
Italian	0	0	0	0	0	0	0	1	0	1
Portuguese	1	1	0	1	0	0	1	0	0	1
Latin	0	0	0	1	1	1	1	0	0	1

	Romanian	Spanish	French	Italian	Portuguese	Latin	Average
Romanian	-	7	5	3	8	4	5.4
Spanish	7	-	4	6	3	3	4.6
French	5	4	-	6	5	3	4.6
Italian	3	6	6	-	5	5	5
Portuguese	8	3	5	5	-	4	5
Latin	4	3	3	5	4	-	3.8

Regarding more aspects where there is within-Romance variation, Romanian patterns like Spanish and French (also like Latin), but unlike Italian or Portuguese, concerning the interpredictability of 1PL and 2PL present indicative. Regarding the absence of morphological interpredictability between the 2SG and the 3SG present indicative, and between the 1PL present indicative and the infinitive, Romanian patterns like Italian, and unlike Spanish and Portuguese.

Although this goes beyond the goals of the present paper, these properties (e.g. whether any pair of cells is (1) or is not (0) mutually predictable) could be represented as vectors of (binary) values for each language, and between-language similarity could be explored via Hamming distances or similar (see Table 4) to check if paradigmatic structural distance corresponds to phylogenetic distance or is instead more similar to others like orthographic distance or mutual intelligibility (see Ciobanu & Liviu).

A finer-grained approach to predictability reveals that some of these differences are not so categorical. Although 2SG (Z4) and 3SG present indicative (Z5), and 1PL present indicative (Z2) and the infinitive (Z12), are not perfectly interpredictable in Romanian as they are in other Romance languages, these forms are still very close to perfect predictability (i.e. conditional entropy = 0 in both directions in Table 5) but ultimately fall short of it.

The average implicative entropy between Romanian cells is overall 0.1467 (0.1804 between distillations), which is slightly lower than in the other Romance languages that have been analyzed in a comparable way except Spanish: 0.28 for Latin (Pellegrini & Passarotti, 2018), 0.18 for French, and 0.17 for Portuguese (Beniamine, 2018), 0.07 for Spanish (Herce, 2023). Conditional entropies between Romanian verb distillations differ widely, between 1.172 as the highest [the uncertainty involved in predicting Z6 (ind.prez.3pl) from Z3 (the ind.prez.1sg)] and 0 as the lowest. Closest to perfect interpredictability are Z5 and Z6, i.e. 3SG and 3PL present, and the different areas within the former perfectum tenses (aka. PYTA in the literature on Romance stem alternations, see Maiden, 2001), e.g. Z9 and Z10, as well as these areas and the participle (i.e. Z14).

Table 5 Conditional entropies (column given row) between the distillations in Fig. 2

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	Z11	Z12	Z13	Z14
Z1		0.081	0.051	0.009	0.018	0.014	0.107	0.132	0.055	0.051	0.056	0.065	0.076	0.056
Z2	0.478		0.544	0.202	0.484	0.434	0.478	0.1	0.003	0.002	0.005	0	0.005	0.002
Z3	0.507	0.282		0.009	0.514	1.172	0.539	0.223	0.127	0.247	0.137	0.189	0.215	0.14
Z4	0.314	0.217	0.117		0.282	0.289	0.218	0.118	0.046	0.194	0.053	0.069	0.165	0.054
Z5	0.041	0.099	0.063	0.006		0	0.122	0.144	0.063	0.059	0.061	0.073	0.083	0.063
Z6	0.036	0.08	0.063	0.006	0.001		0.117	0.121	0.068	0.068	0.068	0.066	0.088	0.07
Z7	0.368	0.189	0.046	0.005	0.371	0.381		0.069	0.052	0.17	0.055	0.051	0.175	0.053
Z8	0.649	0.593	0.772	0.433	0.692	0.576	0.863		0.333	0.295	0.306	0.411	0.438	0.303
Z9	0.517	0.008	0.554	0.197	0.519	0.518	0.505	0.091		0	0.002	0.008	0.042	0
Z10	0.452	0.03	0.549	0.209	0.457	0.477	0.505	0.075	0.002		0.002	0.03	0.012	0.008
Z11	0.531	0.03	0.559	0.2	0.531	0.519	0.51	0.075	0.002	0		0.03	0.013	0.008
Z12	0.493	0.187	0.519	0.197	0.485	0.457	0.484	0.086	0.003	0.002	0.005		0.006	0.002
Z13	0.455	0.25	0.649	0.278	0.471	0.505	0.577	0.344	0.302	0.301	0.251	0.184		0.241
Z14	0.495	0.02	0.546	0.218	0.487	0.491	0.49	0.08	0.005	0.003	0.008	0.02	0.012	

Table 6 Some present indicative inflected forms and alternations in Romanian

verb	ind.prez.1sg	ind.prez.3sg	ind.prez.3pl	'ind.prez.1sg', 'ind.prez.3sg'	'ind.prez.1sg', 'ind.prez.3pl'	'ind.prez.3sg', 'ind.prez.3pl'
fi	súnt	jéste	súnt	_ún_ ⇌ jé__e	⇌ (1177)	jé__e ⇌ _ún_
avea	ám	áre	áw	m ⇌ re	m ⇌ w	re ⇌ w
vrea	vrEáw	vrEá	vór	w ⇌ (5)	_Eáw ⇌ ó_	_Eá ⇌ ó_
face	fák	fátʃe	fák	k ⇌ tʃe (33)	⇌ (1177)	tʃe ⇌ k (33)
privi	privésk	privéʃte	privésk	sk ⇌ ʃte (865)	⇌ (1177)	ʃte ⇌ sk (865)
vedea	vɐ́d	véde	vɐ́d	é_ ⇌ é_e (4)	⇌ (1177)	é_e ⇌ é_ (4)
considera	konsíder	konsíderə	konsíderə	⇌ ə (537)	⇌ ə (536)	⇌ (2367)
publica	públik	públikə	públikə	⇌ ə (537)	⇌ ə (536)	⇌ (2367)
încheia	ɨnkéj	ɨnkéje	ɨnkéje	⇌ e (256)	⇌ e (88)	⇌ (2367)
pune	pún	púne	pún	⇌ e (256)	⇌ (1177)	e ⇌ (168)
ajunge	aʒúng	aʒúndʒe	aʒúng	g ⇌ dʒe (59)	⇌ (1177)	dʒe ⇌ g (59)

Table 7 Average predictability and predictiveness of Romanian distillations

Predictability of distillations		Predictiveness of distillations	
Z6	0.449	Z8	0.513
Z7	0.424	Z13	0.37
Z1	0.41	Z3	0.331
Z5	0.409	Z11	0.231
Z3	0.387	Z9	0.228
Z2	0.159	Z12	0.225
Z4	0.151	Z14	0.221
Z8	0.127	Z10	0.216
Z10	0.107	Z2	0.21
Z13	0.102	Z4	0.164
Z12	0.092	Z7	0.153
Z9	0.082	Z5	0.068
Z11	0.077	Z6	0.066
Z14	0.077	Z1	0.059

Table 6 illustrates some of the high-entropy (red) and low-entropy (green) alternations just mentioned, with the numbers in parentheses indicating the number of verbs (if > 1) for which a particular alternation holds. Predicting the 3PL present indicative from the 1SG of the same tense is difficult (in this direction) because there are various frequent unpredictable ways in which the latter form can be turned into the former, most importantly ⇌ e (i.e. add /e/), ⇌ ə (i.e. add /ə/), and ⇌ (i.e. leave unchanged). Predictions in the opposite direction (i.e. predicting the 1SG form from the 3PL) are easier, and the same applies to predictions between 3SG and 3PL present indicative, because most forms (e.g. ʤe ⇌ g, ʧe ⇌ k, ʃte ⇌ sk) allow a speaker to deduce one from the other. Note that, among the alternations shown, only ⇌ and e ⇌ overlap in forms, which means a 'zero' 3PL does not fully diagnose the 3SG, which could be formed either with 'zero' (the most frequent option), or by adding − e. This is the reason why, as shown in Table 5, the conditional entropy of the 3PL (Z6) given the 3SG (Z5) is 0 (no uncertainty) but that of the 3SG given the 3PL is not zero (although still very low).

Table 7 shows the average predictability (average of columns in Table 5) and predictiveness (average of rows in Table 5) across all distillations. Unlike in other Romance languages, it can be observed that differences in predictiveness are similar in range to differences in predictability, ranging roughly between 0.5 and almost zero. In common with other Romance languages, however, we can see quite a sharp boundary between the most and least predictable distillations. The zones Z1, Z3, Z5, Z6 and Z7 are all quite difficult to predict from other forms (note that higher numbers indicate less predictability). This is due to unpredictable stem alternations that separate rhyzotonic forms (the so-called N-morphome cells, see Maiden, 2018) from

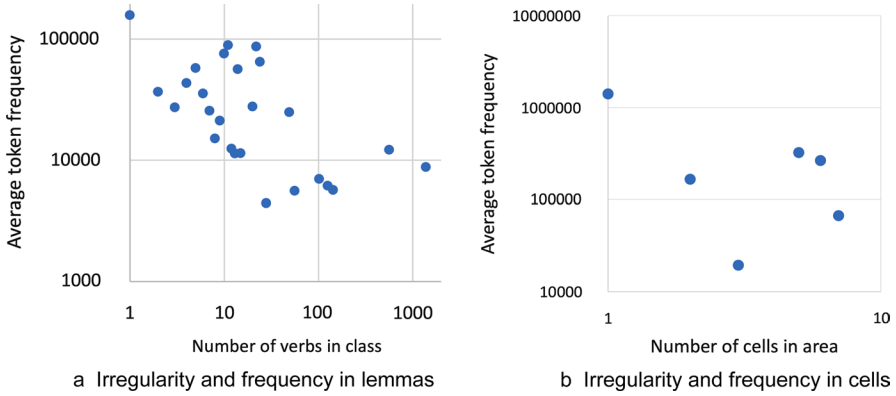

Fig. 3 **a** Irregularity and frequency in lemmas. **b** Irregularity and frequency in cells

arhyzotonic ones. Consider for example 1PL.PRS.IND.beg rugə́m > 1SG.PRS.IND.beg róg, but 'occupy' okupə́m > okúp and 'calculate' kalkulə́m > kalkuléz. Looking at predictiveness, in turn, we find that the imperfect indicatives (Z8) are clearly the forms which are least informative about the morphology of paradigm cells from other distillations. This is due to the very low allomorphic diversity of these forms compared to others, since the only lexical-class distinction that applies to them is the contrast between a suffix − a (for first conjugation verbs and a few from the fourth) and a suffix − e̯a (for all others).

Beyond Romance, Romanian behaves like most (or maybe all) languages concerning the association between high frequency and irregularity (Herce, 2019; Wu et al., 2019). This can be observed (see Fig. 3) both at the lexeme level, where verbs from smaller classes tend to be more frequent (see Fig. 3a) and at the paradigm cell level, where cells that belong to small or single-cell interpredictability domains (e.g. Z1, Z5, Z6, etc. in Fig. 2) tend to be more frequent than more regular cells (see e.g. the Z8 cells in Table 3), i.e. cells that can be predicted from multiple other ones in the paradigm.

4 Conclusion

This paper has presented a new resource VeLeRo (Verbal Lexicon of Romanian), containing the full paradigms in phonological form of 7297 verbs, as well as the frequencies of these lemmas and of their different inflectional values in CoRoLa (Tufiș et al., 2019). The lexicon is made openly available for further research into quantitative morphology and the PCFP.

After outlining the overall interest of this resource and what role it can have within the general morphological-predictability literature in Section one, we proceeded to explain in Sect. 2 all steps and challenges that were involved into the resource's compilation: the addition of stress (in a language whose orthography does not indicate it), the phonemization of problematic cases like palatalizations

and diphthongs, etc. In Sect. 3, in turn, we used our resource and extant freely-available software [Beniamine's (2018) Qumin], to conduct an initial assessment on morphological predictive complexity in the system. The measures and results have been contextualized within the overall Romance landscape. Romanian is widely regarded as the most divergent Romance national standard language due to its earlier phylogenetic split (Balkan Romance split off from Western Romance [Port, Sp, Fr, It] before this group of languages started to break up), the lack of geographic contiguity to the rest of the Romance world, and strong language contact with Slavic languages. Although paradigm-structural differences between Romanian and its Romance sisters might correlate with geographic and phylogenetic distances (see Table 4), our results illustrate a great degree of similarity overall to the other Romance languages, with a comparable level of complexity (similar number and pattern of distillations, and average conditional entropies). Differences are found mostly on those aspects where variation exists already within Western Romance (e.g. regarding the morphological-predictive allegiance of the 1PL and 2PL present indicative, infinitive, of 2SG and 3SG or the present indicative, etc.) and tend to be a matter of degree.

Overall, hence, the results point to the diachronic stability of morphological predictive relations within inflectional paradigms. This goes in line with the conservative nature that has been often claimed for inflectional morphology (Meillet, 1958, Nichols, 1996), which is considered to be less prone to borrowing/contact than other components of language (Matras, 2015). This should be particularly the case for structures like inflection classes and stem alternations which do not bear a direct relationship with meaning/function (Maiden, 2018). If the conservativeness of paradigmatic predictability structure is confirmed, this would open the door to the possibility of using the paradigmatic complexity and patterns themselves for phylogenetic purposes (see Herce and Bickel forthcoming), i.e. to diagnose genetic relations between languages or inflectional systems and categories. This should be the focus of future research, along with the creation of large and well-curated inflected lexicons for other languages, particularly minoritized non-WEIRD languages (see e.g. Cruz et al., 2020), which might differ importantly with respect to the larger languages we are more familiar with (see e.g. Trudgill, 2011).

Acknowledgements We would like to thank the editors of LRE, and three reviewers for their insightful comments.

Author contributions BH had the original idea and wrote the manuscript, BP led phonemization and computational analysis. All authors reviewed the manuscript.

Funding Open access funding provided by University of Zurich. This research has been partially funded by the NCCR Evolving Language, Swiss National Science Foundation (Agreement Nr. 51NF40_180888).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In J. P. Blevins & J. Blevins (Eds.), *Analogy in Grammar: Form and Acquisition* (pp. 54–82). Oxford University Press.

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language, 89*(3), 429–464.

Barbu, A.-M. (2008). Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. In *LREC*. Marrakech: European Language Resources Association (ELRA), pp. 1937–1941.

Beniamine, S. (2018). Typologie quantitative des systèmes de classes flexionnelles: Université Paris Diderot dissertation.

Beniamine, S., Bonami, O., & Luís, A. R. (2021). The fine implicative structure of European Portuguese conjugation. *Isogloss. Open Journal of Romance Linguistics, 7*, 1–35.

Beniamine, S., Maiden, M., & Round, E. (2020). Opening the romance verbal inflection dataset 2.0: A CLDF lexicon. In *12th Conference on Language Resources and Evaluation [postponed due to Corona]*, 3027–3035. European Language Resources Association (ELRA).

Bonami, O., Caron, G., & Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In *SHS Web of Conferences*, vol. 8, 2583–2596. EDP Sciences.

Chitoran, I. (1996). Prominence vs. rhythm: The predictability of stress in Romanian. *Amsterdam Studies in the Theory and History of Linguistic Science, Series, 4*, 47–58.

Chitoran, I. (2002). The phonology of Romanian: a constraint-based approach. *Studies in Generative Grammar*, Vol. 56, 105–113. Mouton de Gruyter.

Ciobanu, A. M., & Dinu, L. P. (2014a). On the Romance languages mutual intelligibility. In *Proceedings of the 9th international conference on language resources and evaluation, LREC*, pp. 3313–3318.

Ciobanu, A. M., & Dinu, L. P. (2014b). An etymological approach to cross-language orthographic similarity. Application on Romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1047–1058.

Cruz, H., Stump, G., & Anastasopoulos, A. (2020). A resource for studying chatino verbal morphology. https://arxiv.org/2004.02083

Diaconescu, Ş. S., Rizea, M. M., Codîrlaşu, F. C., Ionescu, M., Rădulescu, M., Mincă, A., & Fulea, Ş. (2015). *Fonetica limbii române*, vol. I–IV, SOFTWIN.

Dindelegan, G. P., & Maiden, M. (2013). *The grammar of Romanian*. Oxford University Press.

Elsner, M., Sims, A. D., Erdmann, A., Hernandez, A., Jaffe, E., Jin, L., ... & Stevens-Guille, S. (2019). Modelling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute?. *Journal of Language Modelling*, 53–98.

Esher, L. (2022). Overlapping subjunctive forms in Gallo- and Ibero-Romance verb paradigms. *Revue Romane, 57*(1), 86–115.

Finkel, R., & Stump, G. (2009). What your teacher told you is true: Latin verbs have four principal parts. *Digital Humanities Quarterly, 3*(1), 10–31826.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29–29.

Herce, B. (2023). VeLeSpa: An inflected verbal lexicon of Peninsular Spanish and a quantitative analysis of paradigmatic predictability. https://www.researchsquare.com/article/rs-2877209/v1

Herce, B. (2019). Deconstructing (ir)regularity. *Studies in Language, 43*(1), 44–91.

Herce, B. (2022). Stress and stem allomorphy in the Romance perfectum: Emergence, typology, and motivations of a symbiotic relation. *Linguistics, 60*(4), 1103–1147.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., ... & Hulden, M. (2018). UniMorph 2.0: universal morphology. https://unimorph.github.io/publications/

Lőrincz, B., Irimia, E., Stan, A., and Barbu Mititelu, V. (2022). RoLEX: The development of an extended Romanian lexical dataset and its evaluation at predicting concurrent lexical information. Natural Language Engineering, 1–26.
Maiden, M. (1992). Irregularity as a determinant of morphological change. *Journal of Linguistics, 28*(2), 285–312.
Maiden, M. (2001). A strange affinity: 'perfecto y tiempos afines.' *Bulletin of Hispanic Studies, 78*(4), 441–464.
Maiden, M. (2018). *The Romance verb: Morphomic structure and diachrony*. Oxford University Press.
Maiden, M., Smith, J. C., Cruschina, S., Hinzelin, M.-O., & Goldbach, M. (2010). *Oxford online database of romance verb morphology*. University of Oxford.
Malkiel, Y. (1966). Diphthongization, monophthongization, metaphony: Studies in their interaction in the paradigm of the Old Spanish-ir verbs. *Language, 42*(2), 430–472.
Malouf, R. (2017). Abstractive morphological learning with a recurrent neural network. *Morphology, 27*, 431–458.
Malouf, R., Ackerman, F., & Semenuks, A. (2020). Lexical databases for computational analyses: A linguistic perspective. *Proceedings of the Society for Computation in Linguistics, 3*(1), 297–307.
Matras, Y. (2015). Why is the borrowing of inflectional morphology dis-preferred? In F. Gardani, P. Arkadiev, & N. Amiridze (Eds.), *Borrowed Morphology* (pp. 47–80). De Gruyter.
Matthews, P. H. (1991). *Morphology*. Cambridge University Press.
Meillet, A. (1958). Linguistique historique et linguistique générale. Société Linguistique de Paris, Collection Linguistique, 8. Librairie Honoré Champion, Paris.
Montermini, F., & Bonami, O. (2013). Stem spaces and predictability in verbal inflection. *Lingue E Linguaggio, 12*(2), 171–190. https://doi.org/10.1418/75040
Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 2710–2714.
Nichols, J. (1996). The Comparative Method as heuristic. In *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, edited by Mark Durie and Malcolm Ross, 39–71.
O'Neill, P. (2018). Velar allomorphy in Ibero-Romance. *Studies in Historical Ibero-Romance Morpho-Syntax, 16*, 13–43.
Pellegrini, M., & Passarotti, M. (2018). LatInfLexi: an inflected lexicon of Latin verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, 324–329. Accademia University Press.
Pellegrini, M., & Cignarella, A. T. (2020). (Stem and Word) Predictability in Italian verb paradigms: An Entropy-Based Study Exploiting the New Resource LeFFI. In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it 2020)*: 1–6. CEUR.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.
Sims, A. D. (2020). Inflectional networks: Graph-theoretic tools for inflectional typology. *Proceedings of the Society for Computation in Linguistics, 3*(1), 88–98.
Spinu, L. (2018). Investigating the status of a rare cross-linguistic contrast: The case of Romanian palatalized postalveolars. *The Journal of the Acoustical Society of America, 143*(3), 1235–1251.
Spinu, L., Percival, M., & Kochetov, A. (2019). Articulatory Characteristics of Secondary Palatalization in Romanian Fricatives. In *Interspeech*, 3307–3311.
Spinu, L., Vogel, I., & Timothy Bunnell, H. (2012). Palatalization in Romanian—Acoustic properties and perception. *Journal of Phonetics, 40*(1), 54–66.
Stump, G., & Finkel, R. A. (2013). *Morphological typology: From word to paradigm* (Vol. 138). Cambridge University Press.
Thornton, A. M. (2012). Reduction and maintenance of overabundance. A case study on Italian verb paradigms. *Word Structure, 5*(2), 183–207.
Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.
Tufiș, D., Mititelu, V. B., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., & Onofrei, M. (2019). Little strokes fell great oaks: Creating CoRoLa, the reference corpus of contemporary Romanian. *Revue Romane De Linguistique, 64*(3), 227–240.
Wu, S., Cotterell, R., & O'Donnell, T. J. (2019). Morphological irregularity correlates with frequency. https://arxiv.org/1906.11483

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press.