Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3027–3035
Marseille, 11–16 May 2020

Opening the Romance Verbal Inflection Dataset 2.0: a CLDF Lexicon

Sacha Beniamine*, Martin Maiden†, Erich Round*‡§
* Max Planck Institute for the Science of Human History, Jena, Germany,
† Faculty of Linguistics, Philology, and Phonetics, University of Oxford, Oxford, England,
‡ School of Languages and Cultures, University of Queensland, Brisbane, Australia,
§ Surrey Morphology Group, University of Surrey, Guildford, UK,
beniamine@shh.mpg.de, martin.maiden@mod-langs.ox.ac.uk, e.round@uq.edu.au

Abstract

We introduce the Romance Verbal Inflection Dataset 2.0, a multilingual lexicon of Romance inflection covering 74 varieties. The lexicon provides verbal paradigm forms in broad IPA phonemic notation. Both lexemes and paradigm cells are organized to reflect cognacy. Such multi-lingual inflected lexicons annotated for two dimensions of cognacy are necessary to study the evolution of inflectional paradigms, and test linguistic hypotheses systematically. However, these resources seldom exist, and when they do, they are not usually encoded in computationally usable ways. The Oxford Online Database of Romance Verb Morphology provides this kind of information, however, it is not maintained anymore and is only available as a web service without interfaces for machine-readability. We collect its data and clean and correct it for consistency using both heuristics and expert annotator judgements. Most resources used to study language evolution computationally rely strictly on multilingual contemporary information, and lack information about prior stages of the languages. To provide such information, we augmented the database with Latin paradigms from the LatInFlexi lexicon. Finally, to make it widely avalable, the resource is released under a GPLv3 license in CLDF format.

Keywords: Lexicon, Lexical Database, Morphology, Less-Resourced/Endangered Languages, Typological Databases, Phonetic Databases, Phonology, Multilinguality

1. Introduction

We[1] introduce the Romance Verbal Inflection Dataset 2.0. It is based on the Oxford Online Database of Romance Verb Morphology (Collective work coordinated by Martin Maiden, 2010, hereafter ODRVM), which constitutes the most comprehensive available unified representation of Romance verbal morphology. The resulting database is a collection of multilingual inflectional lexicons giving full paradigms in phonological form for 74 Romance varieties, including Latin. It is annotated for lexical cognate sets, and paradigm cells are organized in such a way as to reflect cognacy. It is intended as a resource for the study of paradigm evolution, as well as the comparative study of the inflection of the Romance verb.

Figure 1 synthesizes the family group and endangerment status of the varieties represented in the database, based on data from glottolog (Hammarström et al., 2018). Two varieties in our database are extinct: Latin and Dalmatian-Vegliote. Six varieties are considered *shifting* (in use, but not transmitted): Emilian from Travo, Cremonese Lombard, Aromanian, Istro-Romanian from Šušnjevica, Megleno-Romanian, Logudorese Sardinian. Five are considered threatened (Picard from Mesnil-Martinsart, Friulian, Piedmontese from Cairo Montenotte, Northern Veneto from Alpago, Campidanese Sardinian). The glottolog data on endangerment is an aggregation of three sources on endangerment (UNESCO, Ethnologue, ElCat) and is the most comprehensive we could find. Nonetheless, the endangerment status is not specified for 52 of our varieties: for very low resource languages, even knowing the endangerment status is not a given. Most of these varieties are not standard varieties or national languages, and are likely to be at least threatened.

The Oxford Online Database of Romance Verb Morphology, which was set up in 2010, is only available as a browsable website, and intended for manual inspection only. The main contribution of this paper is to make the database usable for computational work, from typological investigations of morphological complexity to evolutionary studies. We scraped the ODRVM, cleaned and corrected the data, completed it and reshaped it in order to conform to the Cross-Linguistic Data Formats Standard (Forkel et al., 2018). We merged in Latin paradigms from LatInFlexi (Pellegrini and Passarotti, 2018) to allow for richer evolutionary analyses. The resulting database is released under a share-alike open licence.

2. Related work and applications

Both the synchronic typology and diachronic evolution of language systems are increasingly being investigated using electronic datasets. However, existing datasets of grammatical traits typically contain on the order of one or two hundred, directly encoded variables (Dryer and Haspelmath, 2013; Carling et al., 2018), whereas datasets of actual word-forms in a commensurate representation make it possible to rapidly and automatically extract vastly more variables (Macklin-Cordes and Round, 2015). This increases the density of information per language that statistical analyses can examine, and can lead to significant, novel typological and evolutionary insights (Greenhill et al., 2017; Round, 2020). Inflectional systems in particular have always been a valuable source of information for linguists reconstructing language relationships, though there remains an active debate

[1] The three authors have contributed in the following ways: Martin Maiden led the original ODRVM project. Erich Round and Sacha Beniamine conceived the project for version 2.0. Sacha Beniamine scraped, sourced, corrected, enhanced, compiled, formatted and archived the data, and wrote associated code. Martin Maiden reviewed the data and provided expert linguistic judgments. All three authors wrote the paper.

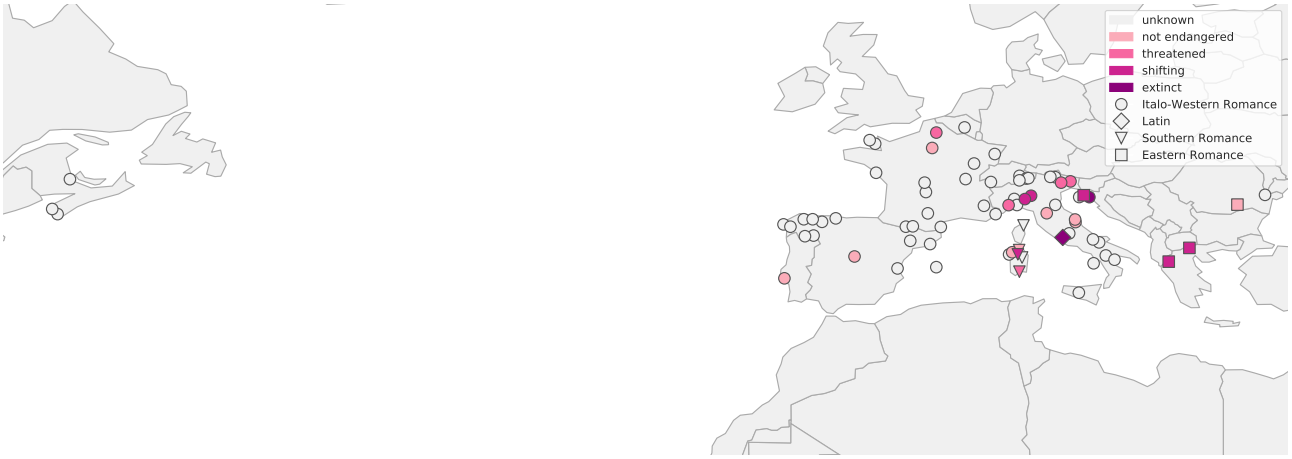

Figure 1: Location, Romance families and endangerment status for the 74 varieties documented in the Romance Verbal Inflection Dataset 2.0

over the mechanisms by which paradigms change (Fertig, 2013; Maiden, 2018). If such inferential tasks are to be automated, and if theories of inflectional change are to be tested in a rigorous, reproducible manner, then machine readable, cognacy-aligned inflectional datasets such as the one we present here must be priority for historical linguistics. This section discusses existing resources.

In the word and paradigm approach to morphology, and particularly inflection, authors have recently focused on the computational and typological study of inflectional complexity (Finkel and Stump, 2009; Ackerman and Malouf, 2013; Blevins, 2013, among others). The ideal data for this purpose is large lexicons of fully inflected surface forms in phonological notation. Several lexicons have been elaborated and distributed for some of the largest Romance languages: Portuguese (Veiga et al., 2013), French (Bonami et al., 2014), Italian (Calderone et al., 2017), as well as for Latin (Pellegrini and Passarotti, 2018). In Natural Language Processing, orthographic lexicons are useful for lemmatization and as auxiliary resources to other tasks. Some inflectional Romance lexicons have been created for that purpose, for example by Barbu (2008) for Romanian. Such individual large lexicons, maintained independently, constitute high quality data, but exist only for very few varieties.

A large number of inflected lexicons were extracted from the Wiktionary as part of the Unimorph project (Kirov et al., 2016), and were used in the context of reinflection tasks (Cotterell et al., 2016), for which such datasets are also necessary. Among Romance languages, these include Asturian, Catalan, French, Friulian, Galician, Italian, Ladin, Neapolitan, Occitan, Portuguese, Romanian, Spanish, Venetian, as well as Middle French and Old French. Latin is also available as a Unimorph lexicon. These datasets can however not be used in their raw form for linguistic investigations(Malouf et al., 2019), due to their orthographic nature and to the lack of homogeneity in the original wiktionnary.

The above-mentioned lexicons are not intended for historical comparison, and as such, contain no cognacy annotation. Historical lexical databases such as the Indo-European Lexicon Cognacy Database (Evolutionary Processes in Language and Culture research group, 2015) are available which provide wordlists in orthographic and phonological notation, for a very large number of languages, organized by meanings and annotated for cognacy. However, the word-forms are not inflected, and each lexeme figures only as a single word-form. As a result, this type of resource is not suitable for the computational study of paradigm evolution. While we know of no comparable resource, there does exist a standard for tabular multilingual linguistic datasets aimed towards the study of language evolution: The Cross-Linguistic Data Formats (Forkel et al., 2018, hereafter CLDF). A CLDF dataset is composed of a set of utf-8 encoded csv tables described by a `TableGroup` description[2] serialized as JSON file, as well as a bibliographic file in Bibtex format. The tables need to implement one of the CLDF modules (in our case, a *WordList*).

To ensure the re-usability and long term preservation of the database, we formatted version 2 in order to be CLDF compliant. This standard has three main advantages for our goals: First, it ensures a clear and self-documented data organization, as well as a degree of data cleanliness. Second, we hope that using this format will facilitate its use for evolutionary studies, where several tools already are CLDF compliant. Third, since it is mostly composed of csv tables, it can be easily read and used for other linguistic investigations.

3. The Oxford Online Database of Romance Verb Morphology

The Oxford Online Database of Romance Verb Morphology (Collective work coordinated by Martin Maiden, 2010) displays a representation of the inflectional paradigms of the verb for the 73 Romance varieties[3] listed in Table 1 (exclud-

[2] TableGroup are described as part of the Metadata Vocabulary for Tabular Data W3c recommendation.

[3] While the site indicates the existence of data for 80 varieties, only 73 are publicly accessible. These are the ones which were included in our release.

ing Latin, which was not part of the original database). The site is organized by language varieties, then by lexemes (labelled by etymon). For each lexeme, it describes the entire verb's inflectional paradigm in phonemic notation. Comments and source are given at the language level, and additional comments and information are given for each lexeme.

3.1. Aims

The Database is only an account of a wide range of published descriptions of Romance verb morphology, and does not purport to reproduce or supplant such descriptions. The act of interpretation is inherently problematic, and users who wish to pursue points of particular interest in the data for any given variety should of course consult the sources cited for that variety. The Database sets out to document only *synthetic* word forms, that is composed of a single word, excluding analytic constructions, which make use of auxiliaries or periphrasis. However, the boundary between *synthetic* and *analytic* is not always clear (Ledgeway, 2011). In addition, Romance languages show varying degrees of fusion of verb-forms with (clitic) subject pronouns: in Romansh and Ladin varieties, for example, the verb system has a set of special forms with phonologically reduced word-final endings when used in interrogative constructions or in certain other types of construction usually requiring syntactic inversion of subject and verb (Alton and Vittur, 1968; Minach and Gruber, 1972); and for many Gallo-Romance and northern Italo-Romance varieties it could be argued (Rizzi, 1986) that what are conventionally regarded as 'obligatory subject clitics' are analyzable as part of the inflectional morphology of the verb. The limitation to 'synthetic' word-forms is dictated not by any particular theoretical stance, but both by tradition and by practical limitations.
The aim of the database is to offer to Romance linguists (and morphologists in general) a tool for the comparative analysis of the inflectional morphology of the Romance verb.

3.2. Sources

The database was constituted by hand, selecting sources with strict criteria. To be included, sources needed to:

- be a detailed and authoritative description of the variety;
- offer a comprehensive description of the inflectional morphology of the verb, including both regular and irregular verbs;
- present data obtained from interrogation of one or more native speakers;
- document at least full inflectional paradigms for the continuants of Latin AMBULARE, DARE, ESSE, FACERE, HABERE, IRE, POSSE, SAPERE, STARE, TENERE, UADERE, UELLE, UENIRE and UIDERE. These (semantically basic and highly frequent) verbs are almost all present in all Romance languages, and are the locus of some major and idiosyncratic types of morphological structure
- document the continuants of at least one member of each of the four major Latin conjugational classes (in addition to those given above);
- if possible, describe a variety spoken in a particular village or town.

Three types of exceptions were made to these rules. First, for some dialects, the compilers of the ODRVM directly interrogated native speakers to acquire data (for example the Italo-Romance varieties of Mussomeli and Macerata). Second, in order to be able to include the last remnant of the Dalmatian branch of the Romance languages, Vegliote, the compilers drew on Bartoli (1906) whose grammatical analysis is mainly extracted from texts and particularly from the elicited narratives of the person believed to be the last (near-) native speaker. Due to the Zipfian distribution of paradigm cells (Blevins et al., 2016; Bonami and Beniamine, 2016), the resulting paradigms are incomplete (of 2510 forms, 636 are marked as defective, and 1450 as missing). Third, in the case of the major standard languages, no source is specified since the forms are extensively and uncontroversially established.

4. The Romance Verbal Inflection Database 2.0

The ODRVM was created manually, and the current site is no longer maintained. Our goal for version 2.0 was two-fold: first, normalize and organize the data to make it exploitable computationally; second, ensure the long term preservation of the data, and make it easier to correct and expand.
The datasets which are used to study language evolution are often large multilingual databases. They rarely represent known earlier stages of the documented languages. To provide diachronic information, we augmented the database with Latin paradigms from the LatInFlexi lexicon.
This section discusses the steps taken to produce a computationally usable language resource in CLDF (Forkel et al., 2018) from the ODRVM, the addition of Latin paradigms, and the format of the resulting database[4].

4.1. Overview of the changes

The database of the original website held the only copy of the data. Because of this, it proved difficult to obtain a usable database dump, and impossible to obtain any documentation on its structure. Instead, with the compilers' permission, we resorted to scraping the website to extract the public database contents. This allowed us to quickly obtain all publicly available data.
Most of the changes made to the data are formatting and normalizations, where we examined unique values for each column manually, and made changes to ensure a homogeneous coding of all categorical variables and identifiers, such as inflection classes, etyma, language names, and to ensure the same notational conventions, in particular regarding phonemic notation. Systematic changes were made to normalize separators everywhere, to ensure that composite values

[4]The full scripts used for that purpose, including scraping, cleaning, corrections, adding Latin paradigms, and producing a CLDF dataset, can be found on a gitlab repository `https://gitlab.com/sbeniamine/scraping_and_formatting_the_ODRVM`

Family	Variety	Town	Forms	Lexemes
Catalan	Eastern > Alguerès	L'Alguer /Alghero	1081	17
	Eastern > Balear > Mallorquí	Palma	1397	19
	Eastern > Central > Barceloní	Barcelona	1190	19
	Eastern > Rossellonès	Canet de Rosselló / Canet-en-Roussillon	1126	18
	Western	Lleida	1249	18
	Western	València	1221	18
Dalmatian	Vegliote	Veglia/Krk	2510	40
Francoprovençal	Lyonnais	Vaux	3423	56
	Valaisan > Val dIlliez	Val d'Illiez	4014	65
French	Acadian	Pubnico	675	10
	Acadian	Baie Sainte-Marie	420	7
	Acadian > South-East New Brunswick	Moncton & environs	667	11
	Franc-Comtois	Pierrecourt	3757	62
	Lorrain	Ranrupt	3645	55
	Modern Standard		3994	66
	Norman	Jersey	1674	27
	Norman > Guernsey		1388	23
	Picard	Mesnil-Martinsart	2957	49
	Poitevin-Saintongeais > Vendéen	Beauvoir-sur-Mer	3231	52
	Wallon	Namur	3482	57
Friulian	Friulian		1772	29
	Western > Maniago	Greci	1421	23
Galego-Portuguese	Galician	Xermade	918	15
	Galician	Lubián	912	15
	Galician	Fisterra / Finisterra	1029	15
	Galician	Dodro	951	15
	Galician	Vilanova de Oscos	937	15
	Galician	Cualedro	925	15
	Portuguese		3103	47
Italian	Central > Laziale > North-central	Ascrea	3134	51
	Central > Marchigiano	Servigliano	1028	17
	Central > Marchigiano	Macerata	1698	25
	Central > Modern Standard		2829	46
	Central > Tuscan > Corsican	Sisco	1544	25
	Northern I > Emilian	Travo	1623	27
	Northern I > Emilian > Romagnol	Lugo	1291	21
	Northern I > Ligurian > Genoese	Genova	1642	25
	Northern I > Lombard > Alpine	Val Calanca	622	10
	Northern I > Lombard > Cremonese	Cremona	772	10
	Northern I > Piedmontese	Cairo Montenotte	1300	21
	Northern I > Piedmontese > Basso	Cascinagrossa	2070	34
	Northern II > Veneto > Istrioto > Valle dIstria	Valle d'Istria	1202	20
	Northern II > Veneto > Northern	Alpago	1160	18
	Southern I > Lucano > Archaic	Nova Siri	1687	26
	Southern I > Lucano > Calabria	Papasidero	1026	17
	Southern I > Lucano > Central	Calvello	1433	22
	Southern I > Molisano	Casacalenda	2108	35
	Southern I > Pugliese > Dauno	Lucera	603	10
	Southern II > Sicilian > Central	Mussomeli	1139	18
Italic > Latino-Faliscan	Latin	Rome	22407	231
Ladin- Dolomitic	Atesino > Val Badia		2998	49
	Atesino > Val Gardena		3387	54
Occitan	Northern > Auvergnat > Gartempe	Gartempe (Creuse)	2681	44
	Northern > Limousin	Saint-Augustin	1702	27
	Northern > Vivaro-Alpin	Seyne	2449	39
	Southern > Languedocien	Graulhet	2055	34
	Southern > Provençal	Nice	2996	46
Romanian	Aromanian		2985	48
	Istro-Romanian	Šušnjevica	1952	32
	Megleno-Romanian		2293	37
	Modern Standard		3580	58
Sardinian	Campidanese		555	8
	Gallurese		302	5
	Logudorese		387	5
	Nuorese		1146	16
	Sassarese		532	8
Spanish	Aragonese	Panticosa	1563	24
	Aragonese > Ansotano	Ansó, Fago	1239	19
	Asturo-Leonese > Asturian	Somiedo	1216	16
	Asturo-Leonese > Asturian > Sudeste de	Parres	1780	22
	Modern Standard		2589	43
Swiss Ræto-Romance	Puter > Upper Engadine		4285	64
	Surmiran	Bivio-Stalla	1558	22
	Surselvan		7695	103

Table 1: List of all varieties in the Romance Verbal Inflection Dataset 2.0.

ID	Latitude	Longitude	Closest_Glottocode	Town	Comment	Variety
Friulian	46.2480	13.0955	friu1240		Source of data: The paucity of adequate descri...	Friulian
French_Picard_from_Mesnil-Martinsart	50.0536	2.6475	pica1241	Mesnil-Martinsart		Picard
Italian_Northern_I_Emilian_Romagnol_from_Lugo	44.4166	11.9166	nort2607	Lugo		Northern I > Emilian > Romagnol

Linguistic_Date	Sources	Country	Family	Region	Variety_in_ODRVM
mid to late 20th century	Zof2000	Italy	Friulian	Friuli	Friulian
late 19th / early twentieth century	Flutre1955	France	French	Picardy	French - Picard – Mesnil-Martinsart
mid to late 20th century	Pelliciardi1977	Italy	Italian	Emilia Romagna	Italian - Northern I - Emilian - Romagnol - Lugo

(a) Language table

ID	LemLat_ID	Language_of_the_etymon	Etymon	Latin_Conjugation	Part_Of_Speech	Derived_from
akkattare		Romance	*akkattare		V	
cubare	cubo	Latin	cubare	I	V	
de-aperire		Latin	de-aperire	IV	V	aperire

(b) Cognate Sets table sample

ID	Etymon_in_ODRVM	Meaning	Comment	Cognateset_ID	Language_ID
lex_1340	PLACERE	please		placere	French_Wallon_from_Namur
lex_2039	UENIRE	come		uenire	Italian_Northern_I_Emilian_from_Travo
lex_127	AMBULARE / IRE / UADERE	go		ambulare~ire~uadere	French_Acadian_South-East_New_Brunswick_from_M...

(c) Lexemes table sample

ID	Name	Description	Continuants
PLUP-IND	Latin pluperfect indicative		CONT_LAT_PLUP-IND
IMPERF-SBJV	Latin imperfect subjunctive		CONT_LAT_IMPERF-SBJV
3PL	3pl	third person plural	

(d) Parameters table sample

ID	Language_ID	Cell	Form	Cognateset_ID
form_723896	Romanian_Modern_Standard	PRS-SBJV~1PL	ˈnaʃtem	nasci
form_2203100	Galego-Portuguese_Portuguese	INFL_INF~3PL	prɐˈzerɐ̃ĩ	placere
form_2262553	Italian_Central_Marchigiano_from_Macerata	ROM_FUT~3PL	aˈvra	habere

(e) Forms table sample

Table 2: Three-row samples for each table in the database

could be split back programatically. We also made manual corrections to mistypes in each table. Heuristics were used to find inconsistencies, then changes were made semi-automatically according to expert annotator judgments.

We added forms for Latin paradigms, taken from from LatInFlexi (Pellegrini and Passarotti, 2018). We selected all lexemes in LatInFlexi which were either one of the Latin etyma in our resource, or from which one of the Latin etyma was marked as derived in the ODRVM. We manually selected the Latin paradigm cells corresponding to cognate cells in the ODRVM. The International Phonetic Alphabet (hereafter IPA) notation used was mostly compatible with ours, requiring minimal change.

The CLDF guidelines recommend splitting data into tables for each type of information, referencing across tables using identifiers. We formatted the ODRVM data into five tables: Languages, Parameters, Cognate Sets, Lexemes and Forms. Each table is given in long form. It has a row identifier, as well as a series of specific columns, see Table 2 (a-e). The resource is comprised of these tables, as well as a JSON meta description and a (.bib) bibliography file created semi-automatically from references on the variety pages.

4.2. Languages

The language table contains one row for each variety documented in the resource. In terms of the CLDF ontology, this table is a `LanguageTable`. We populate it by parsing the variety pages of the ODRVM website, and produce a table with the following columns: `ID`, `Latitude`, `Longitude`, `Closest_Glottocode`, `Town`, `Comment`, `Variety`, `Linguistic_Date`, `Sources`, `Country`, `Family`, `Region`, `Variety_in_ODRVM`. A three row sample is shown in Table 2.a.

The original variety names (retained in the column `Variety_in_ODRVM`) provided some hierarchical information on their linguistic lineage, separated by dashes. We cleaned these names by automatically removing the first part, and moving it to the `family` column, removing the city names (redundant with the `town` field), and making manual corrections. The result can be found in the column `Variety`. The 'linguistic classification' field on the website was too noisy to be usable. Rather than dashes, which led to ambiguity with multi-word language names, we joined hierarchical information by `" > "` (see Table 1).

The CLDF guidelines stress the need to provide references to common resources, in particular glottocodes (Hammarström et al., 2019) for language varieties. We added manually the closest glottocodes for each variety in column `Closest_Glottocode`. The granularity of glottolog and of our database are different, and, we sometimes had to provide the same code for several varieties. For example, three Acadian varieties were all attributed to Acadian. More concerning, we had to give the two Marchigiano varieties the same glottocode as Modern Standard Italian, for a lack of a better node. As a result, the codes should be used with caution.

The columns `Country, Region, Town, Latitude` and `Longitude` used to situate an entire variety are also simplifications. They are meant to designate the location where the primary data was collected. For data from single localities (villages or towns) the chosen coordinates are the corresponding latitude and longitude. In the case of national or regional languages which cannot readily be assigned to any one geographical point, the coordinates are either those of the capital city or chief town of the country or region in which the variety originates. In the case of non-national varieties spoken over a continuous area in more than one country (e.g., Megleno-Romanian), or scattered over several countries (e.g., Aromanian), the chosen coordinate was that of some locality in the area, which is either central or where the largest concentration of speakers is found. For historical reasons, the database indicates Florence for Italian. For Latin, we gave the coordinates of Rome. For the Friulian variety 'Western > Maniago', the source (Iliescu, 1972) pertains to dialects transplanted by emigrants into Romania, whence the rather surprising geographic assignment of this variety to Romania. The `Country, Region` and `Town` correspond to the chosen coordinates.

The `Linguistic_Date` have been copied directly from the ODRVM. They indicate the approximate date at which the data were obtained, divided into late 19th to early 20th century, mid to late 20th century and late 20th to early 21st century.

The `Source` column provides the bibtex key reference used in the associated bib file.

Finally, a `Comment` column mention any problematic or especially noteworthy characteristics of the variety described.

4.3. Cognate sets

The Cognate Sets table indicates *presumed* etyma from which verbs in the database are descended. In terms of the CLDF ontology, this table is a `CognatesetTable`. The table was created from each individual lexeme page for each variety. Because some cognate level information was duplicated in the database, a lot of normalization was needed to obtain this table. The resulting columns are: `ID, LemLat_ID, Language_of_the_etymon, Etymon, Latin_Conjugation, Part_Of_Speech, Derived_from`, as shown in Table 2.b.

The etyma given in the `Etymon` column are categorically not intended as a definitive statement about the etymology of that lexeme, but simply as an identifier which can serve to facilitate cross-linguistic comparison of lexically cognate verbs. They are, in effect, a means of signaling cognacy.

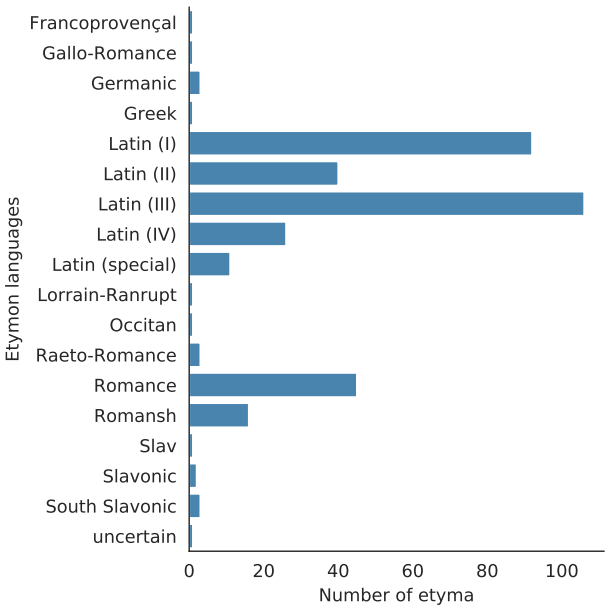

Figure 2: Number of etyma per etymon language and Latin inflection class.

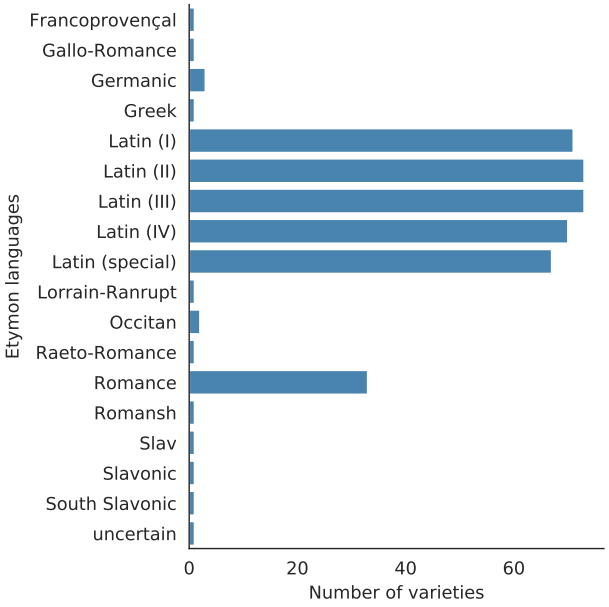

Figure 3: Number of Romance varieties per etymon language and Latin inflection class.

Overwhelmingly, the `Language_of_the_etymon` is Latin, as can be seen in Figures 2 and 3. This information, present in the ODRVM, was collected from etymological dictionaries for each Romance language, as well as from (Meyer-Lübke, 1935). 'Romance' is used to indicate the origin of lexemes which, while not attested in Classical Latin, are present in all or most Romance languages and have no external origin. A case is *passare, which appears to have no precedent in Latin, but is clearly derivationally created from Latin PASSUS 'step', and extensively attested across the Romance languages. Etyma preceded by an asterisk are hypo-

thetical and unattested; a few preceded by '?' are hypothetical, unattested, and open to serious question. In cases of major doubt, the database provides a very recent and local form (e.g. Romansh forms).
For Latin etyma, the database provides the Latin inflectional class membership (one of the four traditional classes), which we report in the `Latin_Conjugation` column. The designation 'special' means that the Latin verb was idiosyncratic and could not be easily classified as belonging to any one of the major inflectional classes. For each Latin etymon where we added a Latin lexeme to the database, we indicate the LemLat identifier in the `LemLat_ID` to preserve the mapping to LatinFlexi (Pellegrini and Passarotti, 2018).
In some cases, the cognacy of verbs is partly masked by the fact that in some varieties the same etymon appears in a derived form, preceded (historically) by prepositions or other prefixes. In such instances, the ODRVM specified the preposition or prefix in parentheses, after the basic etymon (e.g., 'SEDERE (AD+)' for forms derivable from ADSEDERE). We extracted these values to fill a separate `Derived_from` column, and restored the etymon names with a dash, (e.g. AD-SEDERE).
We added a `Part_Of_Speech` column, to allow for the future addition of other paradigms. Currently, all rows are marked as 'V'.

4.4. Lexemes

The Lexeme table provides information for lexemes in each variety, specified by a cognate set identifier and a language identifier. In terms of the CLDF ontology, this table is a `FormTable`. Each row documents one lexeme, using the following columns: `ID, Etymon_in_ODRVM, Meaning, Comment, Cognateset_ID, Language_ID` (see Table 2.c). This table is not declared as any pre-defined component from the CLDF specification. The main reason for its existence is the presence of lexeme level language specific comments and meaning in the ODRVM.
The `Cognateset_ID` column links to identifiers from the cognateset table. The `Etymon_in_ODRVM` is kept for backward compatibility, and shows the etymon exactly as it is given on the ODRVM website, before any cleaning or corrections.
Certain verbs have more than one etymon — and are therefore suppletive by incursion in the terminology of (Corbett, 2007). In these cases, the `Cognateset_ID` column lists all the etyma, joined by a tilde '~'. The most common cases involve the verbs 'to go' (e.g., ambulare~ire~uadere for French), and 'to be' (e.g., esse~stare for French).
The `Meaning` specified for each lexeme is intended to be broadly indicative of its meaning in each particular language. The meaning for auxiliaries is specified as 'auxiliary'. LatInFlexi does not specify English glosses for lexemes. To populate the `Meaning` column for Latin verbs, we added the glosses from Maiden (2018, pp.319–322), and entered glosses manually for the remaining lexemes.
The `Cognateset_ID` and `Language_ID` link back to the `ID` columns of respectively Cognate and Language tables, in the manner of relational databases.
Finally, the `Comment` column indicates any additional comment which was present on the lexeme's page in the ODRVM.

4.5. Parameters

The Parameter table provides the list of all paradigm cells documented in the database. In terms of the CLDF ontology, this table is a `ParameterTable`. The columns are: `ID, Name, Description, Continuants` (see Table 2.d). The paradigm cell names are those extracted from the database website.
The parameters are given for each specific paradigmatic dimension. The `Name` and `Description` are taken from the ODRVM's documentation, and give a short name and an explicative comment for the feature. Some names are explicitly historical (e.g., 'Continuant of Latin X'). These are appropriate where cognate forms have diverged so widely in respect of their functions that any label suggesting some shared cross-linguistic function would be misleading. In these cases, the cell 'Latin X' is given as a separate row, with the identifier of its continuant in the 'Continuant' column. In other cases, the same label (e.g., 'present indicative') is conventionally used to describe the relevant, and historically cognate, set of forms across Romance. We added extra paradigm cells to fully describe Latin paradigms. The `Continuants` column links Latin paradigm cells to the `ID` of their continuant in the same table.

4.6. Forms

The Forms table (see Table 2.e) provides a phonemic representation of each individual form. For each form, the table indicates its identifier (`ID`), the form itself (`Form`), the corresponding language identifier (`Language_ID`), the paradigm cell (`Cell`), and one or more (in case of suppletion) cognateset identifiers (`Cognateset_ID`). The `Cell` is a combination of identifiers from the Parameters table, separated with a tilde '~'.
We homogenized the notation of missing forms. We write 'Ø' to indicate defective forms (the given verb does not have a form for this cell in this language) and '?' to indicate missing information (we do not have the form for this verb and cell in this language). We preserved the same paradigmatic structure in a given language even when a lexeme is missing information or defective for an entire tense/mood.
The sources range in their representation of phonetic detail from quite 'narrow' phonetic transcriptions to rather approximate representations using conventional orthography for the language concerned, or modifications of conventional orthography. There is no denying that the phonetic accuracy of the ODRVM varies considerably from variety to variety, depending on the nature of its sources. To facilitate comparison, the ODRVM provides forms in a broad IPA notation. These were transcribed with careful attention to all indications in the material consulted about the value of the symbols there used. Comments in the language or lexemes table document uncertainties in the transcription.
For version 2.0, we corrected many mistypes and character substitution. The set of IPA characters used in version 2.0 is given in Table 3. In some forms, parentheses indicate phonological material which is optionally or variably present in pronunciation. For example, in the Acadian

	bilab.	lab-dent.	dent.	alv.	post-alv.	ret.	pal.	lab-pal.	lab-vel.	vel.	uvul.	glot.
stop	p b			t d		ɖ	c ɟ			k g		
nasal	m			n			ɲ			ŋ		
trill				r							ʀ	
tap				ɾ								
fricative	ɸ β	f v		θ ð	s z	ʃ ʒ		ç ʝ			x ɣ	χ ʁ	h
affricate					ʦ ʣ	ʧ ʤ						
approximant					ɹ			j	ɥ	w			
lateral approximant					l			ʎ					

	diacritic	value
consonant	◌ʲ	palatalized
consonant	◌̩	syllabic
vowel	◌ː	long
vowel	◌̯	non-syllabic
vowel	◌̞	lowered
vowel	◌̃	nasalized
vowel	◌̝	raised
syllable	ˈ	stress

	front	central	back
close	i y	ɨ	u
	ɪ	ʊ	
	e ø	ɘ	o
mid		ə	
	ɛ œ		ʌ ɔ
	æ	ɐ	
open	a ɶ		ɑ ɒ

Table 3: IPA characters used.

variety from South-East New Brunswick, the first person plural of the indicative present for the cognate FACERE is given as /f(ə)zɔ̃/.

We use brackets to mark clitic pronouns, when we do not have an example of that form without the clitic and it is possible that without the clitic the form of the verb would be slightly different. For example in Vegliote, the imperative second person singular of the cognate TENERE is given as /ˈtjan[te]/.

For the most part, we are confident that the data presented are at least a reliable phonemic representation of the data. In cases of substantial doubt, we have made a note to this effect. It bears repetition, however, that what is presented in the ODRVM, and thus here, is an *interpretation* of other material, and that the references provided here should always be regarded as the sole authority on phonetic representation.

5. Conclusion

We scraped the Oxford Online Database of Romance Verb Morphology. We normalized notations across the database and corrected erroneous values using heuristic and expert annotator judgment. To provide diachronic information, we added Latin paradigms from LatInFlexi. We formatted the resulting dataset into a CLDF compliant format, and made it available under the GPLv3 license.[5]

The resource is archived in Zenodo under DOI 10.5281/zenodo.3611076.

We provide this resource with the intent of furthering the study of how inflectional paradigms evolve. Current studies on inflection change either rely entirely on the knowledge of individual linguists, or on synchronic multilingual datasets encoding categorical features and measurements. Only multilingual lexicons providing full paradigms at different time depths, annotated for cognacy across lexemes and paradigm cells, can allow for richer modeling and comparisons. This is precisely what our resource offers. Our aim in making this dataset available to computational linguists is to enable and encourage the study of paradigm evolution.

In the future, we hope to enhance the database with more paradigms and languages. We would also like to enrich the characterization of paradigm cells by specifying, in addition to the cognate cells, the morphosyntactic value of each cell in each language. A longer term project would consist in gathering similar datasets across other language families.

[5]The CLDF dataset is available as a gitlab repository at: `https://gitlab.com/sbeniamine/Romance_Verbal_Inflection_Dataset/`

6. Acknowledgements

The Oxford Online Database of Romance Verb Morphology was constructed in connection with the Arts and Humanities Research Council-funded research project Autonomous Morphology in Diachrony: comparative evidence from the Romance languages (AH/D503396/1), carried out at Oxford University first in the Faculty of Medieval and Modern Languages, and latterly in the Faculty of Linguistics, Philology and Phonetics, between October 2006 and December 2010.

7. Bibliographical References

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

Alton, J. and Vittur, F. (1968). *L ladin dla val Badia: Beitrag zu einer Grammatik des Dolomitenladinischen*. Weger, Bressanone.

Barbu, A.-M. (2008). Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *Proceedings*

of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, May. European Language Resources Association (ELRA).

Bartoli, M. (1906). *Das Dalmatische. Altromanische Sprachreste von Veglia bis Ragusa und ihre Stellung in der Apennino-Balkanischen Romania*, volume 1 and 2. Hölder, Vienna.

Blevins, J. P., Milin, P., and Ramscar, M. (2016). The Zipfian Paradigm Cell Filling Problem. In Ferenc Kiefer, et al., editors, *Morphological paradigms and functions*. Brill, Leiden.

Blevins, J. P. (2013). The information-theoretic turn. *Psihologija*, 46(4):355–375.

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

Bonami, O., Caron, G., and Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, et al., editors, *Actes du quatrième Congrès Mondial de Linguistique Française*, pages 2583–2596.

Calderone, B., Pascoli, M., Sajous, F., and Hathout, N. (2017). Hybrid method for stress prediction applied to glaff-it, a large-scale italian lexicon. In *Proceedings of the Language, Data and Knowledge Conference (LDK 2017)*, Galway, Ireland.

Carling, G., Larsson, F., Cathcart, C. A., Johansson, N., Holmer, A., Round, E., and Verhoeven, R. (2018). Diachronic Atlas of Comparative Linguistics (DiACL)—A database for ancient language typology. *PloS one*, 13(10):e0205313.

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

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The sigmorphon 2016 shared task – morphological reinflection. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany, August. Association for Computational Linguistics.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Fertig, D. L. (2013). *Analogy and morphological change*. Edinburgh University Press, Edinburgh.

Finkel, R. and Stump, G. T. (2009). Principal parts and degrees of paradigmatic transparency. In James P. Blevins et al., editors, *Analogy in Grammar*, pages 13–54. Cambridge University Press, Cambridge.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205, October.

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

Hammarström, H., Castermans, T., Forkel, R., Verbeek, K., Westenberg, M., and Speckmann, B. (2018). Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392, 9.

Iliescu, M. (1972). *Le Frioulan à partir des dialectes parlés en Roumanie*. Mouton, The Hague.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Ledgeway, A. (2011). Morphosyntactic persistence from Latin into Romance. In M. Maiden, et al., editors, *The Cambridge History of the Romance Languages*, pages 382–471. CUP, Cambridge.

Macklin-Cordes, J. and Round, E. (2015). High-definition phonotactics reflect linguistic pasts. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, Tübingen.

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

Malouf, R., Ackerman, F., and Semenuks, A. (2019). Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics*, 11.

Meyer-Lübke, W. (1935). *Romanisches etymologisches Wörterbuch*. Winter, Heidelberg.

Minach, F. and Gruber, T. (1972). *La rujneda de Gherdëina. Saggio per una grammatica ladina*. Typak, Urtijëi.

Pellegrini, M. and Passarotti, M. (2018). Latinflexi: an inflected lexicon of latin verbs. In Elena Cabrio, et al., editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253 of *CEUR Workshop Proceedings*, page December, Aachen.

Rizzi, L. (1986). On the status of subject clitics in Romance. In O. Jaeggli et al., editors, *Studies in Romance Linguistics*, pages 137–52. Foris, Dordrecht.

Round, E. R. (2020). Phonotactics in Australian languages. In Claire Bowern, editor, *Oxford Guide to Australian languages*. Oxford University Press, Oxford.

Veiga, A., Candeias, S., and Perdigão, F. (2013). Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment. *Journal of the Brazilian Computer Society*, 19(2):127–134.

8. Language Resource References

Collective work coordinated by Martin Maiden. (2010). *Oxford Online Database of Romance Verb Morphology*.

Evolutionary Processes in Language and Culture research group. (2015). *Indo-European lexical cognacy database*.

Harald Hammarström and Robert Forkel and Martin Haspelmath. (2019). *Glottolog database 4.0*. Zenodo.