Isogloss
Open Journal of Romance Linguistics

Received: 21-05-2024
Accepted: 04-10-2024
Published: 20-12-2024

VeLeCa: A verbal lexicon of Catalan with PCFP analysis

Borja Herce
University of Zurich
borja.hercecalleja@uzh.ch

Bogdan Pricop
University of Zurich
bogdan.pricop@uzh.ch

How to cite: Herce, Borja & Pricop, Bogdan. 2024. VeLeCa: A verbal léxicon of Catalan with PCFP analysis. *Isogloss. Open Journal of Romance Linguistics* 10(1)/17, 1-17.
DOI: https://doi.org/10.5565/rev/isogloss.457

Abstract

This paper presents VeLeCa, a new resource on Catalan verbal inflection containing the phonological form of 174,200 word forms from 3,484 lexemes and their respective lexical and morphosyntactic values and frequencies. We describe the challenges and procedure we followed in the compilation and phonemization of this resource, and conduct a computational analysis of the Paradigm Cell Filling Problem (i.e. morphological predictive complexity) in the system to contrast it with those from related Romance languages.

Keywords: Catalan, verb, paradigm, PCFP, morphology

1. Introduction

Romance languages are well-known for their complex verbal inflectional system, with multiple conjugations and irregulars, and dozens of inflected forms. The existence of a well-documented direct ancestor (Latin) and the global influence of some of its modern descendants have contributed to establishing the Romance verb as one of the most studied inflectional systems around the world (see e.g. Maiden 2018). Modern national standard Romance languages (from West to East: Portuguese, Spanish,

French, Italian, and Romanian) have the resources to study them in detail and for their use in Natural Language Processing (NLP) and the digital world more generally. Focusing on the verb, for example, inflected lexicons and quantitative analyses of morphological complexity exist for all of these (see Beniamine et al. 2021, Herce 2023, Bonami et al. 2014, Pellegrini & Cignarella 2020, and Herce & Pricop 2024 respectively). Smaller languages, by contrast, have a risk of falling behind.

Catalan, for example, despite being the sixth largest Romance language by number of speakers (Vila, 2020:636) after the aforementioned national standard ones, lacks a comparable inflected lexicon and quantitative morphological assessment of its verbal inflectional system. Producing these is the goal of this paper. This is intended to contribute to our understanding of morphological-predictive relations within paradigms. Inflectional morphological paradigms are a micro-cosmos of human language, made up of both memorized items and rules to complex words in productive ways. The ability to produce inflected forms that have not been heard before is a central one for the productive use of language. This challenge has come to be known as the Paradigm Cell Filling Problem (PCFP, Ackerman et al. 2009). The measurement and analysis of differences between languages and different aspects of complexity has become a prominent topic of analysis in the last decade (see e.g. Ackerman & Malouf 2013, Stump & Finkel 2013, Blevins et al. 2017, Erdman et al. 2020). Universal principles and generalizations (also Romance-specific observations, see Maiden 2018, Herce 2022) have been proposed in the literature that need to be tested against larger, richer, and more diverse datasets. This is the area where our present paper will contribute.

2. Introduction to Catalan

Catalan is a (Western-)Romance language, spoken natively by around 5.6 million speakers (Vila, 2020:636) mainly in North-Eastern Spain (Catalonia, Valencia, and the Balearic Islands). Catalan can be divided into two geographic varieties, Eastern Catalan and Western Catalan, each with further subdivisions (see e.g. Perea & Ueda 2010). The present work will focus on the Central variety of Eastern Catalan, which is the one spoken in and around Barcelona (also northern Tarragona and Girona, see Nogués-Graell 2019) and functions as the standard one.

Phonologically (Wheeler 2005, Dols 2020), Central Catalan has a system of 25 consonants (see Table 1) and seven vowels (/i/, /e/, /ɛ/, /a/, /o/, /ɔ/, /u/) with an additional central vowel (ə) that only occurs in unstressed syllables[1]. Stress in the language is free (i.e. has to be learnt as part of the pronunciation of a word) but is generally on the last or second-to-last syllable. Quite pronounced vowel reduction linked to stress is highly characteristic of Central Catalan. The vowels /e/, /ɛ/ and /a/ are reduced to /ə/ in unstressed syllables, while /o/, /ɔ/ and /u/ are all realized as /u/, and only /i/ remains distinct (Carbonell & Llisterri 1999, Herrick 2003). The 7-vowel system is thus reduced to a 3-vowel system in unstressed syllables

[1] Hiatus sequences escape this reduction (e.g. *teatre* ‘theatre’ is pronounced /teˈatrə/ rather than */təˈatrə/, and *teatral* ‘theatrical’ is pronounced /teəˈtral/ rather than */təəˈtral/).

Table 1. Catalan consonant inventory

	Bilabial	**Labio-Dent.**	**Dental**	**Alveolar**	**Post-Alv.**	**Palatal**	**Velar**
Plosive	p b		t d				k g
Affricate[2]				t͡s d͡z	t͡ʃ d͡ʒ		
Nasal	m		n			ɲ	(ŋ)[3]
Trill			r				
Tap or Flap			ɾ				
Fricative		f		s z	ʃ ʒ		
Central-Approximant[4]						j	w
Lateral-Approximant			l			ʎ	

Catalan reliably encodes stress through its orthography[5] (see Lamuela 2020), but does not reliably encode vowel qualities. If a word bears an orthographic accent for stress, all 7 vowel qualities are distinguished (í, é, è, à, ò, ó, ú), however if it does not, which can be the case of both stressed and unstressed syllables, only 5 vowel graphemes are distinguished (i, e, a, o, u). In stressed but unaccented syllables, the contrast between mid-open and mid-closed vowels (i.e. /e/ vs /ɛ/, and /o/ vs /ɔ/) is therefore not expressed. Given its orthography, for example, *arrendes* '2SG.lease' could be be pronounced /ərˈendəs/ or /ərˈɛndəs/. The root vowel in this case is /e/ so only the former is correct. Conversely, *arrenques* '2SG.root out' could be pronounced as /ərˈenkəs/ or /ərˈɛnkəs/. In this case /ɛ/ is the root vowel. In unstressed syllables, the opposite is the case. Because the standard orthography does not reflect the aforementioned vowel reductions of Central Catalan, it shows five vowel graphemes for what are only three phonemic distinctions. This means that /ə/ can be represented as either <e> or <a> orthographically, and /u/ can be represented as <o> or <u>.

[2] Although they contrast with fricatives in some positions, the status of affricates as either single phonemes or segment sequences is controversial (see Wheeler 2005:11-13) and subject to variation.

[3] While /ŋ/ is usually considered an allophone of /n/ before velars, this conditioning environment is sometimes lost in Central Catalan, which gives rise to minimal pairs like *aprenc* '1SG.learn' vs *aprèn* '3SG.learn' (/əpɾˈɛŋ/ vs /əpɾˈɛn/). This is the reason we consider it a phoneme in this paper.

[4] The phonemic status of the two central-approximants /j/ and /w/ is disputed, as these are said to derive from /i/ and /u/. Certain metrical specifications might block or promote the emergence of glides from these underlying vowels. For instance, a word initial position and the proximity of the high vowels to the stressed syllable blocks the emergence of glides (e.g. d[iə]ˈlecte 'dialect') while more distance from the stress promotes it (e.g. d[jə]lectoloˈgia 'dialectology') (Lloret & Prieto 2022: 5-11).

[5] Monosyllabic words do not need (and do not have) any indication of their stress. Polysyllabic words that are stressed on the last syllable receive an orthographic accent if they end in a vowel, vowel+*s*, *-en*, or *-in*. Polysyllabic words that are stressed on their second-to-last syllable receive an orthographic accent if they do *not* end in a vowel, vowel+*s*, *-en*, or *-in*. Polysyllabic words stressed on other syllables always bear an orthographic accent. Given these general rules of accentuation, the location of the stress can always be ascertained in Catalan from a word's spelling.

Concerning consonants, the most important synchronic process in Catalan when it comes to morphological predictability is word-final devoicing (Hualde & Zhang 2022). As in other languages like German or Russian, all word-final obstruents (i.e. stops, affricates, and fricatives) are voiceless. As a result, it is not possible to predict, given a word with a final obstruent, whether this will be voiced or voiceless in paradigmatically related word forms in a different (e.g. intervocalic) phonological environment (e.g. *poc* /ˈpɔk/ 'M.SG.little' - *poca* /ˈpɔkə/ 'F.SG.little' vs *groc* /ˈgrɔk/ 'M.SG.yellow' - *groga* /ˈgrɔgə/ 'F.SG.yellow')

Catalan consonantal orthography is comparatively phonemic, with mostly one-to-one correspondences between sounds and graphemes. When it comes to word-final devoicing, however, and unlike the aforementioned example might suggest, orthography varies, often indicating the underlying rather than surface voicing of a consonant (e.g. *pedagog* 'pedagogue', which is pronounced as /pədəgˈɔk/ and not */pədəgˈɔg/). Another process that Catalan orthography does not indicate consistently is the simplification, in Central Catalan, of some homorganic word-final consonant clusters (see Herrick 1999), and the deletion of word-final /r/. Thus, *camp* 'field/pitch' is pronounced /ˈkam/[6], *punt* 'point' is pronounced /ˈpun/, *perd* '3SG.lose' is pronounced /ˈpɛr/, etc. and *anar* 'go' is pronounced /əˈna/, *tenir* 'have' /təˈni/, *fer* 'do' /ˈfe/, etc.

Concerning the Catalan verbal system, agreement is found with person and number of the subject. Verbs also inflect for eight different TAMs and six nonfinite forms, for a total of 50 cells (see Figure 2). Morphologically, verbs are traditionally (Fabra 1937) divided into three main classes, defined on the basis of the thematic vowel. Conjugation I (continuing Latin First Conjugation) is defined by the thematic vowel -a-, Conjugation II (which continues Latin Second and especially Third Conjugation) by -e- and Conjugation III (which continues Latin Fourth Conjugation) by -i-. In modern Catalan, this system has become more opaque as the thematic vowels can have different allomorphs, or be absent, depending on the tense and agreement suffixes (see Oltra Massuet 1999 for an in-depth description). Of the roughly 4500 verbs listed in the Diccionari General de la Llengua Catalana (Fabra 1932) more than 3500 fall in the Conjugation I. All verbs in this class have *-ar* as the infinitive suffix and are regular with the exception of *anar* 'go' and *estar* 'be'. Conjugation II is the smallest one and, notably, is characterized by a high prevalence of irregular participial forms. Verbs in this class take *-re, -er* (unstressed), *-r,* and *-er* (stressed) as the infinitive suffixes. Finally, Conjugation III is the second most numerous class with around 700 verbs and is further divided in IIIa and IIIb. Both subdivisions use -ir as the infinitive suffix. Verbs of Conjugation IIIa, also known as the inchoative, have a stem extension in the form of *-eix-* in the standard variety in SG.PRS and 3PL.PRS. Verbs of Conjugation IIIb are fewer in number, lack the inchoative root extension, and are mostly irregular (Wheeler et al. 2002: 288-290).

3. Creating the inflected lexicon VeLeCa

While Catalan is a thriving language, public language resources and corpora are smaller than in most national standard Romance languages (see Bou 2020 for an

[6] Evidence for this comes from F.C. Barcelona's anthem's first verse 'tot el camp es un clam'. *Camp* rhymes with *clam* 'uproar' because the 'p' in the former is no longer pronounced.

overview of different available Catalan resources). To construct the inflected lexicon we have made use of a machine-readable structured JavaScript Object Notation (JSON) data file containing Catalan verb inflectional paradigms extracted from the open access linguistic resource Wiktionary (Ylonen 2022). Morphosyntactic values of words were standardized to Unimorph conventions (Sylak-Glassman 2016, Batsuren et al. 2022) and the inflected orthographical forms were subsequently phonemized using TransDic (Garrido et al. 2018). TransDic was chosen as a starting point for phonemization because of its dialect specific feature, especially relevant since we aimed at producing an accurate transcription of the Central dialect. Furthermore, unlike other public grapheme-to-phoneme systems, its performance was evaluated and claimed to produce satisfactory results when compared to human transcribers. Nevertheless, several issues had to be addressed and corrected manually.

An important goal in the development of the lexicon was to capture processes that occur invariably in the Central variety, i.e. to use a broad phonetic approach. For this reason, consonant devoicing or unstressed vowel neutralizations were reflected in the transcription, while consonant gemination, known to vary even within Central Catalan (Wheeler 2005: 265; Lloret & Prieto 2022: 14) was not. The spirantization of voiced obstruents /b/, /d/ and /g/ into [β], [ð] and [ɣ] was also not reflected in the transcription due to the allophonic nature of this process. Other adjustments to the phonemization included correcting the transcription of uix as [uʃ] rather than [ujʃ] (e.g. dibuixar ‘draw’, /diβuʃˈa/ and not /diβujʃˈa/) , the treatment of x as either [ʃ] or [ks] as appropriate (e.g. ixo ‘1SG.go.out’, /ˈiʃu/ and not /ˈiksu/), the incorrect transcription of the tj sequence sometimes as [tʒ] rather than [dʒ] (e.g. assetjar ‘besiege’, /əsədʒˈa/ and not /əsətʒˈa/), the treatment of vowels with diaeresis (ï and ü), which were erroneously assigned primary stress invariably, as well as correcting cases where the TransDic transcription did not reflect word-final cluster simplification or place of articulation assimilation in nasals (see Section 2).

As explained in Section 2, stressed <e> and <o> can stand for either the stressed mid close vowels /e/ and /o/, or for the mid open /ɛ/ and /ɔ/. While recurring suffixes (e.g. -eix- as /ɛʃ/, or -es- as /es/) were transcribed mostly correctly by TransDic, infrequent roots were not, and this orthographic ambiguity resulted in many transcription mistakes. These were solved in the following way. First we identified these ambiguous cases in our lexicon (i.e. verbs where the tonic root vowel is either an accented <e> or <o>). Then we extracted the root vowel identity information provided in Wiktionary for the Central variety (phonemic information, and etymological information is generally provided alongside orthographic forms). Once this was done we adjusted the transcription of the stressed vowel.

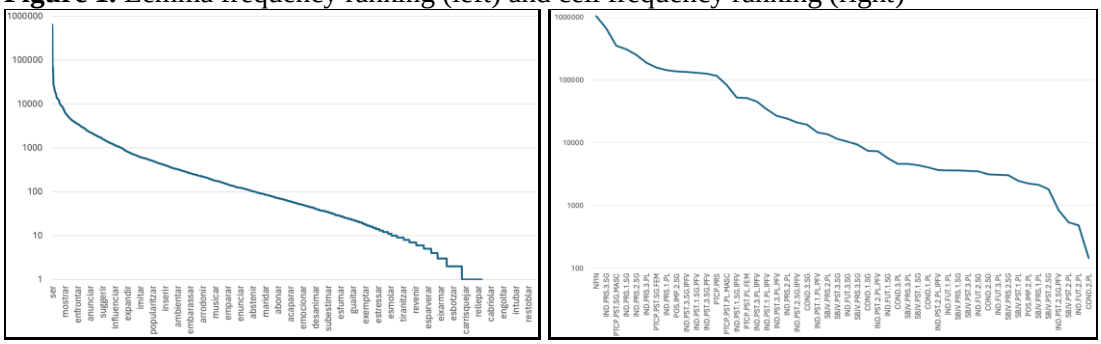
Alongside phonological transcription, we also needed to amend various aspects of the Wiktionary forms we relied on. Synonymous verbal forms in use in different geographical locations (e.g. *vindre* vs *venir* 'come'), as well as reflexive markers of intransitive verbs were removed from the lexicon. Sometimes, Wiktionary forms were inadequate in other respects. For example, in verbs where two different inflections are possible (e.g. conjugation IIIa vs IIIb as in *consumo* ‘1SG.consume’, *consums* ‘2SG.consume’, *consum* ‘3SG.consume’ vs *consumeixo, consumeixes, consumeix*) we needed to make sure that one alternative was chosen consistently within a given verb (instead of, for example, *consumeixo*, *consums*, *consum*). These inconsistencies and other isolated errors in Wikipedia were detected semi-automatically, by identifying

morphologically unique word forms and verbs, and suspicious alternations through Qumin (Beniamine 2018) and checking their correctness.

Non-second person imperatives were also removed, since these forms can be argued not to be true imperatives (see e.g. Jary & Kissine 2016, Holvoet 2023) and are morphologically identical with the corresponding forms of the PRS.SBJV tense.

Finally, the corpus was supplemented with etymological information from Wiktionary where available, and with frequency estimates extracted from the CUCWeb corpus (Boleda et al. 2006). The left panel of Figure 1 shows the token frequencies (ordered from more to less) of all the lemmas in our resource (note that the least frequent 395 lemmas are completely absent from CUCWeb). The right panel shows the ordered adjusted cell frequencies.

Figure 1. Lemma frequency ranking (left) and cell frequency ranking (right)



We decided to adjust cell frequencies to avoid two main limitations of the CUCWeb corpus or its tagging: 1) the tagging of forms that are syncretic between 1SG and 3SG (in most of these cases, tokens were attributed to 1SG, although this does not match the relative frequencies of 1 and 3 in the corpus when they are morphologically distinguished), and 2) due to the written nature of the corpus, 3rd person forms outnumber 1st and 2nd persons by around 10 to 1, and 30 to 1, respectively. More balanced corpora (e.g. SUBTLEX-CAT, Boada et al. 2020) suggest that 3rd person forms outnumber 1st persons only slightly. We correct for these two deficiencies in the estimation of per-cell frequencies in the following way: with non-syncretic forms we subtract 45% of the tokens of 3 and distribute them between the corresponding 1 and 2 forms in proportion to their observed frequencies[7]. With syncretic forms we attribute 65% of the tokens of 1SG to 3SG and 10% to 2SG.

Two peculiarities of Catalan frequencies that deserve to be mentioned with respect to those of other Romance languages are the extraordinary frequency of *anar* 'go' among lemmas, and of the infinitive among cells. The former is almost as frequent as the verb *ser* 'be', which is the most frequent one in the language and very notably so in other Romance (and non-Romance) languages. The latter is the most frequent cell in the paradigm, above the 3SG.PRS.IND which is the most frequent one in other languages. Both aspects could be due to the idiosyncratically Catalan past periphrases (e.g. *vaig parlar*, lit. go.1SG speak.INF) which, unlike in other Romance languages,

[7] We do not do this in the PRS.SBJV because 2PL is highly inflated due to polite imperative uses, so in this case we redistribute tokens at a 50-50 proportion between 1PL and 2PL. We also cannot adjust the underrepresentation of 2 imperatives this way, so instead we have multiplied their observed frequency by 20.

involve an auxiliary verb 'go' and the infinitive, rather than the verb 'have' or 'be' and the participle (see e.g. Juge 2006).

4. Quantitative morphological analysis of the PCFP in Catalan verbs

The PCFP (see Ackerman et al. 2009) is the name given to the challenge that speakers of languages with allomorphy face when producing some forms in the paradigm on the basis of other forms. Research has shown (see Blevins et al. 2017) that no matter how big a corpus/input is, most inflected forms of most lexemes are not attested even once, which means they need to be predicted from other forms[8]. The uncertainty that speakers face when filling out their verbal paradigms has come to be generally measured through entropy (Shannon 1948). Entropy (see the formula below) is a measure of information or uncertainty. It is expressed in bits, 1 bit being equivalent to the uncertainty of a (fair) coin toss.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Because in language a lexeme can only ever be known to exist, in practice, if at least one form is known, conditional entropy (see the formula below) tends to be used more often as a measure of the PCFP. It captures the uncertainty remaining in one variable, given knowledge of another. For example, how difficult is it to guess the form of the infinitive given knowledge of the 3SG.PRS.IND.

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x)$$

Nowadays, ready-made software solutions exist that allow us to calculate conditional entropies and other measures of morphological complexity directly from inflected forms. Here we will use Beniamine's (2018) Qumín. This set of Python scripts (which has also been used with comparable lexicons in other Romance languages) allows us to calculate these from whole unsegmented word forms, hence avoiding the segmentation problem that other software requires us to solve in advance. An automated approach to segmentation, hence, improves replicability and saves time. The way Qumín works is locally, through sequence-to-sequence alignment and the extraction of by-pair morphological alternations.

[8] Complete inflected lexicons, like the one we develop here, are based on competence, which typically allows to produce even highly infrequent forms. It should be kept in mind, however, that the token frequency of forms is a crucial property when it comes to the realistic exploration of the PCFP, which is the reason why it has been supplied to the resource, as explained around Figure 1.

Table 2. Some illustrative forms and alternations in Catalan verbs

	INF	1PL.COND	1SG.COND	INF 1PL.COND	INF 1SG.COND	1PL.COND 1SG.COND
Trobar 'find'	trubá	trubəríəm	trubəríə	á ⇌ əríəm	á ⇌ əríə	_ ⇌ _m
Donar 'give'	duná	dunəríəm	dunəríə	á ⇌ əríəm	á ⇌ əríə	_ ⇌ _m
Anar 'go'	əná	əniríəm	əniríə	á ⇌ iríəm	á ⇌ iríə	_ ⇌ _m
Fer 'do'	fé	fəríəm	fəríə	é ⇌ əríəm	é ⇌ əríə	_ ⇌ _m
Dir 'say'	dí	diríəm	diríə	í ⇌ iríəm	í ⇌ iríə	_ ⇌ _m

Some of these are illustrated in Table 2. The extracted alternations (e.g. _ ⇌ _m) represent what the shortest and most general way is to transform one form into another. In the case of the morphological relationship between the 1PL.COND and the 1SG.COND, we find that this is predictable (one must always add /m/ at the end to derive the former from the latter). This means that there is no uncertainty (i.e. conditional entropy = 0) involved in predicting between these forms, which hence belong to the same predictability domain in the paradigm (also known as 'distillation'[9], see Stump & Finkel 2013). The morphological relation between the INF and COND forms, by contrast, is not fully predictable. While in most verbs one must replace COND /əríə/ by /á/, in the verb *fer* 'do' this is replaced by /é/. In the opposite direction, INF /á/ is usually replaced by /əríə/, but in *anar* 'go' this is replaced instead by /iríə/. The result is that conditional entropy is not zero between the INF and the COND in Catalan. These forms belong, hence, to different predictability domains.

Figure 2. Morphological interpredictability domains in Catalan verbs

	1SG	2SG	3SG	1PL	2PL	3PL
IMP	-	Z1	-	-	Z2	-
IND.PRS	Z3	Z4	Z5	Z6	Z6	Z7
SBJV.PRS	Z8	Z8	Z8	Z9	Z9	Z8
IND.IPFV	Z10	Z10	Z10	Z10	Z10	Z10
IND.PST	Z11	Z12	Z12	Z12	Z12	Z12
SBJV.PST	Z13	Z13	Z13	Z13	Z13	Z13
COND	Z14	Z14	Z14	Z14	Z14	Z14
FUT	Z14	Z14	Z14	Z14	Z14	Z14

INF	Z15	
GER	Z16	
PTCP.M	Z17	Z17
PTCP.F	Z18	Z18

Figure 2 displays the morphological predictability domains that exist in Catalan verbs. This and subsequent results are based on the 1457 verbs which have 100 tokens in the CUCWeb corpus or more. The number of predictability domains is 18, which is somewhat higher than in the Romance verbal inflectional systems analyzed so far with the same methodology (these number up to 15 in Italian, 14 in Spanish, French and Romanian, and 12 in Portuguese, see Pellegrini & Cignarella 2020, Herce 2023, Bonami et al. 2014, Herce & Pricop 2024 and Beniamine et al. 2021 respectively). They also could be contrasted with the 11 domains that Guerrero (2014) found regarding stem allomorphy with a different methodology (consider the notion of 'stem

[9] We will avoid using this technical term throughout the paper simply because it is not a widely known one.

space' in Montermini & Bonami 2013). Discrepancies between the domains of Guerrero and ours (e.g. Z11=Z12=Z13 in Guerrero 2014:161) are due to suffixal allomorphy (see Table 3), which is not considered in this literature.

Commonalities with other Romance languages, probably inherited from from Proto-Romance, are many: the fact that future and conditional form a single domain, the contrast between SG+3PL PRS.SBJV and 1PL+2PL PRS.SBJV, the fact that the 1SG.PRS.IND forms an area all by itself, the greater morphological complexity of the PRS.IND compared to other tenses, etc. These commonalities are analyzed in much of the literature on Romance morphomes (consider the ones called Fuèc, N-morphome, L-morphome, and their interactions, see e.g. Maiden 2018, Herce 2019).

Differences from other Romance languages include the fact that 2PL.IMP and 2PL.PRS.IND are not interpredictable (i.e. the contrast between Z2 and Z6), and the absence of perfect predictability between the IND.PST and SBJV.PST (i.e. Z12 and Z13). The former results from a small number of verbs having different forms for these two values (e.g. *digueu* 'say.2PL.IMP' /digˈɛw/ vs *dieu* 'say.2PL.PRS.IND2' /diˈɛw/) while most have full syncretism (e.g. *trobeu* 'find.2PL.IMP' /tɾubˈɛw/ vs *trobeu* 'find.2PL.PRS.IND' /tɾubˈɛw/). The latter constitutes a split between former perfective tenses that typically behave in other Romance languages as a single unit (so-called *Pretérito Y Tiempos Afines*, PYTA, Maiden 2001) in processes of morphological change. As Table 3 illustrates, these unpredictabilities result from different conjugation neutralizations in different cells. While the 1SG.PST suffix -í neutralizes all conjugation distinctions (including the one between the first conjugation like *donar* 'give' and others like *poder* 'be able to', *morir* 'die'), this is not the case of the other endings (e.g. -ésim, -ísim; -árəs, -érəs, -íras). These non-isomorphic neutralizations translate into predictive uncertainties between these domains. The absence of allomorphy in the 1SG.PST cell, in particular, makes this the least informative form in the Catalan verbal paradigm (see Figure 3). This form can hence be easily predicted from other cells (e.g. 2SG.PST or 1PL.SBJV.PST) but is not very useful to predict other inflected forms.

Table 3. Some former-perfective forms and alternations in Catalan

	IND.PST. 1.SG.PFV	IND.PST. 2.SG.PFV	SBJV.PST. 1.PL	IND.PST. 1.SG.PFV ⇌ IND.PST. 2.SG.PFV	IND.PST. 1.SG.PFV ⇌ SBJV.PST. 1.PL	IND.PST. 2.SG.PFV ⇌ SBJV.PST. 1.PL
Trobar 'find'	tɾubí	tɾubárəs	tɾubésim	í ⇌ árəs	í ⇌ ésim	árə ⇌ é_im
Donar 'give'	duní	dunárəs	dunésim	í ⇌ árəs	í ⇌ ésim	árə ⇌ é_im
Anar 'go'	əní	ənárəs	ənésim	í ⇌ árəs	í ⇌ ésim	árə ⇌ é_im
Morir 'die'	muɾí	muɾírəs	muɾísim	⇌ ɾəs	⇌ sim	ɾə ⇌ _im
Poder 'can'	pugí	pugérəs	pugésim	í ⇌ érəs	í ⇌ ésim	ɾə ⇌ _im

Across the paradigm, the average conditional entropy in Catalan verbs is 0.217 bits, which rises to 0.3 bits when only conditional entropies between different interpredictability domains are taken into account as in Table 3. This measure of PCFP complexity is higher in Catalan than in the other Romance languages that have been analyzed in the same way so far (0.18 for French, and 0.17 for Portuguese (Beniamine 2018), 0.15 for Romanian (Herce & Pricop 2024, and 0.07 for Spanish (Herce 2023).

Figure 3. Conditional entropies between domains (higher values in darker gray)

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	Z11	Z12	Z13	Z14	Z15	Z16	Z17	Z18
Z1		0.039	0.032	0	0.005	0.039	0.019	0.031	0.014	0.036	0.018	0.02	0.018	0.036	0.058	0.036	0.015	0.015
Z2	0.878		0.867	0.883	0.891	0.007	0.893	0.875	0.139	0.313	0.141	0.299	0.138	0.172	0.151	0.357	0.125	0.125
Z3	0.517	0.047		0.137	0.51	0.046	0.003	0.001	0.046	0.121	0.006	0.109	0.054	0.084	0.112	0.113	0.061	0.062
Z4	0.134	0.045	0.022		0.133	0.044	0.009	0.022	0.026	0.036	0.024	0.027	0.026	0.041	0.067	0.017	0.02	0.02
Z5	0.007	0.037	0.029	0		0.037	0.025	0.026	0.014	0.037	0.015	0.017	0.015	0.036	0.055	0.035	0.014	0.014
Z6	0.884	0.021	0.877	0.89	0.894		0.907	0.889	0.149	0.326	0.152	0.292	0.149	0.192	0.169	0.367	0.12	0.12
Z7	0.684	0.093	0.08	0.376	0.687	0.094		0.119	0.09	0.258	0.05	0.156	0.094	0.187	0.284	0.212	0.11	0.114
Z8	0.572	0.048	0.068	0.187	0.563	0.048	0.062		0.044	0.122	0.006	0.298	0.051	0.085	0.149	0.11	0.163	0.16
Z9	0.859	0.079	0.876	0.85	0.855	0.073	0.864	0.927		0.185	0.003	0.388	0	0.137	0.095	0.216	0.232	0.232
Z10	0.89	0.184	0.878	0.847	0.876	0.188	0.964	0.877	0.163		0.093	0.156	0.156	0.128	0.094	0.188	0.071	0.071
Z11	1.347	0.655	1.347	1.335	1.343	0.65	1.333	1.406	0.598	0.683		0.974	0.595	0.671	0.666	0.76	0.798	0.798
Z12	0.792	0.025	0.806	0.786	0.787	0.006	0.809	0.831	0.002	0.006	0		0	0.03	0.004	0.006	0	0
Z13	0.85	0.086	0.886	0.85	0.849	0.073	0.867	0.927	0.004	0.171	0.003	0.447		0.144	0.088	0.242	0.254	0.244
Z14	0.938	0.017	0.828	0.919	0.936	0.022	0.854	0.819	0.023	0.108	0.049	0.078	0.029		0.067	0.119	0.019	0.028
Z15	0.791	0.013	0.812	0.784	0.784	0.013	0.807	0.812	0.011	0.013	0.013	0.015	0.013	0.025		0.017	0.003	0.026
Z16	0.782	0.012	0.807	0.781	0.783	0	0.844	0.807	0.02	0	0.017	0.03	0.03	0.029	0.011		0.004	0.004
Z17	0.814	0.018	0.819	0.811	0.81	0.013	0.822	0.828	0.032	0.013	0.03	0.045	0.03	0.029	0.003	0.016		0.071
Z18	0.787	0.007	0.792	0.783	0.782	0.002	0.794	0.8	0.004	0.002	0.002	0.002	0.002	0.003	0.003	0.002	0	

Figure 3 allows us to zoom in to inspect finer-grained structural aspects. It displays the conditional entropies between all of the morphological-predictive subdomains in Figure 2. Alongside the unusually uninformative character of the 1SG.PST, we can visually appreciate structural principles typical of Romance verbal inflection, such as, most notably, the stark split between word forms from Z1, Z3, Z4, Z5, Z7 and Z8 vs. the other zones. This corresponds to a division between rhizotonic (i.e. root-stressed) word forms (e.g. /əkˈuzə/ 3SG.PRS.IND.accuse) and arhizotonic (i.e. suffix-stressed) word forms (e.g. /əkuzˈa/ INF.accuse). This is a morphological aspect (note that stress is not phonologically predictable in Catalan, as the provided minimal pair shows) that has been discussed extensively in the literature on Romance morphology, often under the term N-pattern (see Maiden 2018). The observation, when it comes to conditional entropies in Catalan verbs, is that while rhizotonic forms are good predictors for all or most forms in the paradigm (i.e. of rhizotonic and arhizotonic forms alike), arhizotonic forms are quite bad predictors of rhizotonic forms. This is due to a reduced number of vowel quality distinctions in unstressed syllables compared to stressed syllables. This is a feature inherited from Proto-Western-Romance to some extent, as a seven-vowel system in stressed environments was opposed already in the proto-language to a five-vowel system in unstressed syllables (this is the system still found in Italian, for example). While this has been inherited by all Romance languages to some extent, Catalan has taken this dichotomy a step further due to the further reduction of unstressed vowel qualities to just three in the standard Central Catalan dialect that we focus on here). As a result, it has become even more difficult to guess a rhizotonic form from an arhizotonic one. While in *acusar* ‘accuse’ before, unstressed /u/ corresponded also to /u/ when stressed, in other verbs this can correspond to stressed /o/ (e.g. in *donar* 'give'), or /ɔ/ (e.g. in *morir* 'die').

Another development typical of Catalan but also found in other Romance languages is the rapprochement of the former-perfective (i.e. PYTA) forms of the paradigm with the participles. Although perfect predictability is not found in the opposite direction, Z12 (e.g. a 2SG.IND.PST) can fully predict other PYTA forms (Z11 and Z13) and the participial forms (Z17 and Z18). This high predictability results from a tendency, both in Catalan and other Romance languages (see Wheeler 2011, Badal 2024), to unify the morphology of these two parts of the paradigm[10].

An unusual finding, found in Catalan but not in other Romance languages, is the split, when it comes to morphological interpredictability, of masculine and feminine participles. This results mostly from the devoicing of consonants at the end of the word. That is, because both /d/ and /t/ are realized as /t/ word-finally, we have participial forms like regular F.SG.go /ənˈadə/ vs M.SG.go /ənˈat/ and F.SG.give /dunˈadə/ vs M.SG.give /dunˈat/, opposed to irregular participles like F.SG.say /dˈitə/ vs M.SG.say /dˈit/, F.SG.do /fˈetə/ vs M.SG.do /fˈet/. Hence, the masculine participle cannot fully predict the feminine one. The reduction of some consonant clusters at the end of the word introduces further occasional uncertainty (e.g. M.SG.open /ubˈɛɾtə/ vs M.SG.open /ubˈɛɾ/).

Figure 4. Average entropy (in bits) associated with predicting from other cells (left, predictability) and to other cells (right, predictiveness)

Predictability		Predictiveness	
Z1	0.737	Z11	0.939
Z5	0.735	Z6	0.435
Z4	0.66	Z2	0.427
Z8	0.647	Z13	0.411
Z7	0.64	Z9	0.404
Z3	0.637	Z10	0.401
Z12	0.197	Z14	0.344
Z16	0.166	Z17	0.306
Z10	0.143	Z16	0.292
Z18	0.124	Z15	0.291
Z15	0.122	Z12	0.288
Z14	0.119	Z18	0.28
Z17	0.118	Z7	0.217
Z2	0.084	Z8	0.161
Z13	0.082	Z3	0.119
Z9	0.081	Z4	0.042
Z6	0.079	Z1	0.025
Z11	0.037	Z5	0.024

Figure 4 shows the average predictability and predictiveness of the various zones in Figure 2. Because a high degree of allomorphy makes a form more informative but also more difficult to guess, there is an inverse correlation between how difficult it is to guess other cells from a given form, and how difficult it is to guess that form from other forms. In Catalan, Z11 (i.e. the 1SG.IND.PST) is the form that is easiest to guess (0.037 bits) but most difficult to guess another form from (0.939).

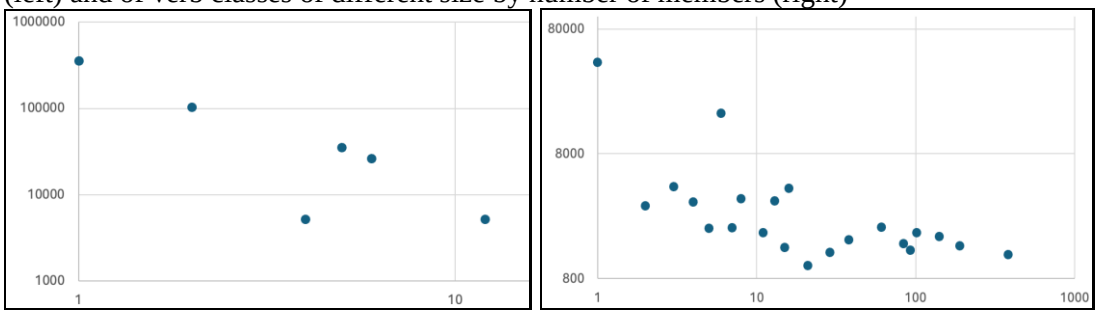
[10] The tendency to level the preterite and the participle morphology is found also in Germanic strong verbs (See e.g. *stand stood* see Dammel et al. 2010), and might be seen as semantically motivated.

Conversely, the most difficult forms to guess, and the easiest forms to guess from are Z1 and Z5, i.e. 2SG.IMP and 3SG.PRS.IND respectively.

A final finding, in this case shared probably with all languages (Wu et al. 2019), concerns the robust relationship, in Catalan verbs, between regularity and token frequency (see Figure 5). Figure 5's left panel shows the average frequency of cells from paradigmatic domains of different sizes. Thus, cells from one-cell domains like Z1, Z2, Z3 etc. are the most frequent while the average frequency of cells from larger domains (e.g. Z14 with 12 cells) tends to be lower. This is related to traditional observations (e.g. Manczak 1966, Milizia 2019) whereby less frequent meanings are prone to lack forms of their own (i.e. are syncretized or expressed through separative exponence).

The same can be said about lemma frequencies. Morphologically unique lemmas (i.e. singleton inflection classes) are the most frequent on average, while lemmas from larger classes tend to be less frequent. This inverse correlation between type frequency (i.e. class size, regularity) and token frequency is well known (Bybee 1995, Lieberman 2007, Herce 2016, Wu et al. 2019) and possibly universal.

Figure 5. Average frequency of morphological domains of different size by number of cells (left) and of verb classes of different size by number of members (right)



5. Conclusions

This paper has presented a new resource, VeLeCa, which presents all inflected forms (50 as identified here) from 3,484 verbal lexemes of Catalan, for 174,200 total forms. Orthographic forms from Wiktionary were checked and phonologically transcribed, and supplemented with lemma-level and cell-level frequencies. An analysis of the PCFP in the system was also conducted. The lexicon, freely available for non-commercial purposes, and all results can be found online at https://osf.io/xzjfy/?view_only=4cb45d355ade4da5b4b8b529ff548ee6.

The analysis of the PCFP in Catalan verbs that we conducted shows both cross-linguistically observed general tradeoffs, and continuities with other Romance languages. Among the first is the inverse correlation between frequency and regularity: verbs that belong to larger inflectional classes tend to have lower token frequency on average. Among the latter is the big divide we found in the predictability of root-stressed SG+3PL present forms and the rest of the paradigm. Largely due to vowel reduction (from seven to only three different qualities in Central Catalan), there is considerably higher uncertainty to predict root-stressed forms from suffix-stressed forms in Catalan than the other way around. A greater degree of morphological complexity has been found in Catalan verbs, overall, than in the related Romance

languages analyzed so far, both in terms of average conditional entropies, and in the number of paradigmatic domains. Catalan is also the smallest of these languages by number of speakers and the only one which is not the standard language of a large nation-state. Further research and data collection would be needed (in terms of further documentation and generation of large inflected lexicons from minoritized Romance languages) to assess the possible effect of sociolinguistic factors like population size, or number of L2 learners (Kusters 2003, Trudgill 2011).

Another fruitful avenue for future research would be the diachronization of quantitative PCFP research. As more and larger resources accumulate, particularly in a family like Romance with a highly documented direct ancestor, the opportunity arises to study morphological change through similar methods as are currently used for the quantitative analysis of paradigm synchrony. Computerized forward reconstruction (Sims-Williams 2018, Marr & Mortensen 2023, List forthcoming), for example, could be applied to recapitulate the effect of sound change over ancestral paradigms, and to separate this effect from that of analogical morphological change. Phylogenetic comparative methods (see e.g. Cathcart 2018) can also be employed to investigate the dynamics and trends of change from synchronic comparative data in Romance and beyond, and artificial language learning experiments (e.g. Saldana et al. 2022) can help us understand what are the cognitive biases that drive this change.

Acknowledgments

We would like to thank the feedback of three anonymous reviewers if Isogloss, as well as the funding of SNSF (Spark grant n. 220720).

References

Ackerman, Farrell, James P. Blevins, & Robert Malouf. 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*: 54-82. Oxford: Oxford University Press.

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

Badal, Manuel. 2024. El proceso de velarización de los participios de la segunda conjugación del catalán: Un ejemplo de analogía retrasada. *Verba: Anuario Galego de Filoloxía* 51: 1-21.

Batsuren, Khuyagbaatar, Goldman, Omer, Salam, Khalifa, Habash, Nizar, Kieraś, Witold, Bella, Gábor, Leonard, Brian et al. 2022. *UniMorph 4.0: universal morphology*. https://doi.org/10.48550/arXiv.2205.03608

Beniamine, Sacha. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. Ph.D. dissertation, University Paris Diderot.

Beniamine, Sacha, Bonami, Olivier, & Ana R. Luís. 2021. The fine implicative structure of European Portuguese conjugation. *Isogloss. Open Journal of Romance Linguistics* 7: 1–35.

Blevins, James P., Milin, Petar & Michael Ramscar. 2017. The Zipfian paradigm cell filling problem. In K. Ferenc, J. P. Blevins, & H. Bartos (eds.), *Perspectives on Morphological Organization*, 139-158. Leiden: Brill.

Boada, Roger, Guasch, Marc, Haro, Juan, Demestre, Josep, & Pilar Ferré. 2020. SUBTLEX-CAT: Subtitle word frequencies and contextual diversity for Catalan. *Behavior Research Methods* 52: 360-375.

Boleda, Gemma, Bott, Stefan, Meza, Rodrigo, Castillo, Carlos, Badia, Toni, & Vicente López. 2006. CUCWeb: a Catalan corpus built from the Web. In *Proceedings of the 2nd International Workshop on Web as Corpus*.

Bonami, Olivier, Caron, Gauthier, & Clément Plancq 2014. Construction d'un lexique flexionnel phonétisé libre du français. *SHS Web of Conferences* 8: 2583–2596. EDP Sciences.

Bou, Joan S. 2020. Language corpora. In J. A. Argenter, & J. Lüdtke (eds.), *Manual of Catalan Linguistics*, 421–440. Berlin: Walter de Gruyter.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5): 425–55.

Carbonell, Joan, & Joaquim Llisterri. 1999. Catalan. *Handbook of the International Phonetic Association*.

Cathcart, Chundra A. 2018. Modeling linguistic evolution: A look under the hood. *Linguistics Vanguard* 4(1): 20170043.

Dammel, Antje, Nowak, Jessica, & Mirjam Schmuck. 2010. Strong-verb paradigm leveling in four Germanic languages: A category frequency approach. *Journal of Germanic Linguistics* 22(4): 337-359.

Dols, Nicolau. 2020. Phonology, phonetics, intonation. In J. A. Argenter, & J. Lüdtke (eds.), *Manual of Catalan Linguistics*, 101-128. Berlin: De Gruyter.

Erdmann, Alexander, Elsner, Micha, Wu, Shijie, Cotterell, Ryan, & Nizar Habash. 2020. *The paradigm discovery problem*. https://doi.org/10.48550/arXiv.2005.01630

Fabra, Pompeu. 1932. *Diccionari general de la llengua catalana*. Barcelona: Llibreria Catalònia.

Fabra, Pompeu. 1937. *La conjugació dels verbs en català*. Barcelona: Barcino.

Garrido, Juan M., Codina, Marta, & Kimber Fodge. 2018. TransDic, a public domain tool for the generation of phonetic dictionaries in standard and dialectal Spanish and Catalan. In *IberSPEECH*, 291–295.

Guerrero, Aurélie. 2014. *Analyse thématique de la flexion en catalan central standard*. PhD dissertation, Université Toulouse le Mirail - Toulouse II.

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

Herce, Borja. 2019. Morphome interactions. *Morphology* 29(1): 109-132.

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

Herce, Borja. 2023. VeLeSpa: An inflected verbal lexicon of Peninsular Spanish and a quantitative analysis of paradigmatic predictability. *Research Square*.

Herce, Borja, & Bogdan Pricop. 2024. VeLeRo: an inflected verbal lexicon of standard Romanian and a quantitative analysis of morphological predictability. *Language Resources and Evaluation*: 1-17.

Herrick, Dylan. 1999. Catalan cluster simplification and nasal place assimilation. *Phonology at Santa Cruz* 6: 25-37.

Herrick, Dylan. 2003. *An acoustic analysis of phonological vowel reduction in six varieties of Catalan*. Ph.D. dissertation, University of California, Santa Cruz.

Holvoet, Axel. 2023. Towards an enhanced semantic map for imperatives. *STUF-Language Typology and Universals* 76(4): 635-657.

Hualde, Jose I., & Jennifer Zhang. 2022. Intervocalic lenition, contrastiveness, and neutralization in Catalan. Isogloss. *Open Journal of Romance Linguistics* 8(4): 1-20.

Jary, Mark, & Mikhail Kissine. 2016. When terminology matters: The imperative as a comparative concept. *Linguistics* 54(1): 119-148.

Juge, Matthew. 2006. Morphological factors in the grammaticalization of the Catalan "go" past. *Diachronica* 23(2): 313-339.

Kusters, Wouter. 2003. *Linguistic complexity*. Utrecht: Netherlands Graduate School of Linguistics.

Lamuela, Xavier. 2020. Spelling. In J. A. Argenter, & J. Lüdtke (eds.), *Manual of Catalan Linguistics*, 81-100. Berlin: De Gruyter.

Lieberman, Erez, Michel, Jean-Baptiste, Jackson, Joe, Tang, Tina, & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163): 713-716.

List, Johann-Mattis. (Forthcoming). *Modelling sound change with ordered layers of simultaneous sound laws*. https://doi.org/10.17613/4n5z-9y52

Lloret, María R., & Pilar Prieto. 2022. Catalan. In C. Gabriel, G. Randall, & T. Meisenburg (eds.), *Manual of Romance Phonetics and Phonology* 27, 743-778. Berlin: De Gruyter.

Maiden, Martin. 2001. A strange affinity: 'perfecto y tiempos afines'. *Bulletin of Hispanic Studies* 78(4): 441-464.

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

Mańczak, Witold. 1966. La nature du supplétivisme. *Linguistics* 4: 82–89.

Marr, Clayton, & David Mortensen. 2023. Large-scale computerized forward reconstruction yields new perspectives in French diachronic phonology. *Diachronica* 40(2): 238-285.

Mascaró, Joan. 1991. Iberian spirantization and continuant spreading. *Catalan Working Papers in Linguistics* 1: 167-179.

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

Milizia, Paolo. 2019. Diachrony and morphological equilibrium: The case of the southern New Indo-Aryan verb. In M. Cennamo, G. Giusti, B. Sevdali, & M. Taine-Cheikh (eds.), *Historical Linguistics 2015: Selected papers from the 22nd International Conference on Historical Linguistics*, 150-169. Amsterdam: John Benjamins.

Nogués-Graell, Jordina. 2019. *Vowel Reduction in Catalan Varieties: Catalan Typologies and Property Analysis*. Master's thesis, UiT Norges arktiske universitet.

Oltra-Massuet, Maria I. 1999. *On the notion of theme vowel: A new approach to Catalan verbal morphology*. Ph.D. dissertation, Massachusetts Institute of Technology.

Pellegrini, Matteo, & Alessandra T. Cignarella. 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.

Perea, María-Pilar, & Hiroto Ueda. 2010. Applying quantitative analysis techniques to La flexió verbal en els dialectes catalans. Dialectologia et Geolinguistica 18: 99–114.

Saldana, Carmen, Herce, Borja, & Balthasar Bickel. 2022. More or less unnatural: Semantic similarity shapes the learnability and cross-linguistic distribution of unnatural syncretism in morphological paradigms. *Open Mind* 6: 183-210.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3): 379-423.

Sims-Williams, Patrick. 2018. Mechanising historical phonology. *Transactions of the Philological Society* 116(3): 555-573.

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

Sylak-Glassman, John. 2016. *The composition and use of the universal morphological feature schema (unimorph schema)*. Johns Hopkins University.

Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press, USA.

Vila, F. Xavier. 2020. Language demography. In J. A. Argenter, & J. Lüdtke (eds.), *Manual of Catalan Linguistics*, 629-648. Berlin: De Gruyter.

Wheeler, Max, Yates, Alan, & Nicolau Dols. 2002. *Catalan: A comprehensive grammar*. London: Routledge.

Wheeler, Max. 2005. *The phonology of Catalan*. Oxford: Oxford University Press.

Wheeler, Max. 2011. The evolution of a morphome in Catalan verb inflection. In M. Maiden, J. C. Smith, M. Goldbach, & M-O. Hinzelin (eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*, 183-209. Oxford & New York: Oxford University Press.

Wu, Shijie, Cotterell, Ryan, & Timothy J. O'Donnell. 2019. *Morphological irregularity correlates with frequency*. https://doi.org/10.48550/arXiv.1906.11483

Ylonen, Tatu. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the International Conference on Language Resources and Evaluation*, 1317-1325. European Language Resources Association (ELRA).