Published in Corpus Linguistics and Linguistic Theory 12.2 (2016)

# Why frequency and morphological irregularity are not independent variables in Spanish: a response to Fratini et al. (2014)

Borja Herce
borja.herce@ehu.eus

**Abstract:** Fratini *et al*. (2014) concluded that frequency and morphological irregularity are in Spanish, unlike in English, independent variables. In this paper I take issue with that claim. On the one hand, it is argued that the borders between regularity and irregularity are diffuse. Many of the verbs classified by Fratini *et al.* (2014) as irregular might, therefore, not be so. In addition, the choice of lexemes they analyzed was far from adequate. Their set of irregular verbs contained many verbs formed by adding some prefix to a more frequent irregular verb (e.g. *a-venir, a-tener, con-decir, con-mover...*) and many highly infrequent lexemes in general, barely in use in the speech community (e.g. *abnegar, ablandecer, amoblar, amodorrecer...*). In an alternative corpus analysis it has been found that, when these and other shortcomings in Fratini *et al.* (2014) are dealt with, morphological irregularity and frequency are indeed strongly correlated variables also in Spanish.

**Keywords:** Spanish, irregular verbs, frequency, corpus, Fratini *et al*. (2014)

## 1 Introduction

That there is a general correlation between irregularity and frequency (Bybee 1991, Haspelmath & Sims 2010: 274-277) as well as between frequency and length (Zipf 1935) has long been *communis opinio*. Challenging received wisdom is a laudable enterprise but of course has to be well motivated. In this paper I argue that, *contra* Fratini *et al.* (2014), received wisdom was accurate in this case about the general correlation between irregularity and frequency[1]. In Section 2 I argue that there were several shortcomings in Fratini *et al*. (2014)'s research that may have had pernicious effects in both the data they obtained, their usage of that data and their eventual conclusions. In Section 3, in addition, an alternative corpus research is conducted to investigate, more adequately than in crude mass-data analyses it is argued, the connection between frequency and irregularity in Spanish verbs. In Section 4 I reach some final conclusions and suggest some topics for future research.

## 2 Evaluation of Fratini *et al*. (2014)

### 2.1 The distinction between regular and irregular in grammar

It is difficult to provide a discrete cut-off point between regularity and irregularity in a way which is not either arbitrary or exclusively theory-driven. In English we are "fortunate" to have some of the clearest distinctions cross-linguistically between regular and irregular verbs. On the one hand we have an enormous class (e.g. *start started started*) which includes the vast majority of verb types.

[1] Fratini et al. (2014) argued that the existence of differences in the frequency of regular and irregular forms constitutes evidence in favour of dual-route models of morphology whereas the opposite argues in favour of single-route models. It is not clear to me why this should be necessarily so. Many authors (e.g. Bybee) are well aware of the differences in frequency of regular and irregular items but still argue for single-route models. Conversely, the latest versions of dual-route models acknowledge that frequent regular forms can be listed in the lexicon, thus making the relation between regularity and rule-generation not univocal. Unlike Fratini et al. (2014), the present paper is not intended to present evidence for or against single- or dual-route models of morphology.

On the other hand we have the rest of the verbs, an utter minority, which show idiosyncratic vowel and/or consonant alternations and which can only be further grouped at most in very small internally homogeneous groups (e.g. *bear bore born, tear tore torn, wear wore worn*).

Other languages, by contrast, are very far from being so "tidy" in this respect. Spanish, for example, presents more problems when trying to classify many of its numerous verbal inflectional classes as either regular or irregular so the criteria for (ir)regularity become essential. Traditional grammars of Spanish do not state them openly but their practices in this respect appear to involve both convention (i.e. a continuation of the traditional verbal conjugations of Latin) and a determination to consider irregular everything that looks like a modification of the root as it appears in the infinitive. If we want the term "regular" to be a scientific and cross-linguistically applicable one and not an *ad hoc* label used for different things in different languages, we surely cannot let those be the criteria. Stolz et al (2007:9) also noticed that "[i]ndividual scholars tend to leave their ideas of irregularity unexplained." This is not a healthy scientific habit either and will be abandoned here.

Outside language, the term "regular" usually means "built according to some established rule" or "happening over and over again". When applied to grammar, it is defined as "conforming to the normal or usual manner of inflection" (Webster). The notion of "regular" as something common or frequent is, therefore, at the core of the definition. The more usual (i.e. frequent) a given way is to form the 1SG past tense, for example, the more regular that particular type of inflection will be, regardless of the changes it involves or what the traditional grammatical conventions have been. Similarly, a verb which is conjugated in the same way as another 1000 types will be more regular than another one whose inflection class contains 100 verbs. The latter, in turn, will be more regular than a verb whose inflection class contains only a handful of verbs. This will be the idea of regularity underlying this paper.

## 2.2 Regularity and irregularity in Fratini et al. (2014).

In agreement with the more traditional grammatical descriptions, Fratini et al. (2014) classified as regular only three Spanish verb classes. These are identified by their thematic vowels, with infinitives in *-ar* (conjugation 1) *-er* (conjugation 2) and *-ir* (conjugation 3) and do not present any vowel or consonant alternations. It has to be noted at this point that the first of these classes, with 9615 members according to the count of Carreras Riudavets *et al*. (2010), is the one with by far the biggest number of verb types and the one where new coinages or borrowings are included (e.g. *twittear, escanear, chequear, tunear...*). Under a very strict definition of inflectional regularity it could well be argued that this is the only regular conjugation in Spanish. There is, indeed, a huge gap with the other classes, since regular conjugations 2 and 3 have "only" 151 and 354 members respectively according again to Carreras Riudavets *et al*. (2010)[2].

Apart from these three conjugations, however, there are a few other big verbal classes in Spanish. There are, for example, 349 verbs of the *-ar* type which are inflected with *o>ue* or *e>ie* stem vowel alternations like e.g. *contar > cuento, acertar > acierto* and 77 verbs or the -er type which do the same (Carreras Riudavets *et al*. 2010). These diphthongizations occur, of course, in predictable slots in the paradigm so the verbs in this big inflectional classes cannot be classified as irregular simply because traditional grammars have usually attached them that label. Some research papers concerned with the regular-irregular distinction in Spanish have in fact been aware of this fact.

[2] For additional arguments concerning the fundamental difference between the first class (-ar) and the other two (-er -ir) see Verissimo & Clahsen (2009) for Portuguese or Say & Clahsen (2002) for Italian.

Balaguer *et al*. (2005) for example, called these diphthongizing verbs "semi-regular" and distinguished them from both the traditional 3 classes and from other "more irregular" verbs.

Many verbs ending in *-ocer -acer* or *-ucir* and above all those with the very productive suffix *-ecer* (e.g. *palidecer, anochecer, acontecer, reverdecer, enorgullecer, reducir, parecer, enrojecer, esclarecer, nacer, conocer...*) regularly insert a /k/ phoneme before certain person affixes (i.e. *palidezco, palidezca...*) also in predictable slots in the paradigm. There are as many as 361 verbs in this class according to Carreras Riudavets *et al*. (2010). This class appears to be so productive, that it is those few verbs in *-(e)cer* which do not belong to this class, like *convencer* or *mecer,* traditionally considered regular, that are out of line in the language. This is supported by the continued occurrence of analogized forms like *convenzco, convenzca* and *mezco, mezca* despite the efforts of prescriptive grammarians.

Other traditionally irregular paradigms or alternations cannot be considered irregular either without further thought. All verbs ending in *-uir,* for example, 65 according to Carreras Riudavets *et al.* (2010) (e.g. *contribuir, argüir, construir, huir...*) add a /j/ before personal endings not starting with /i/ (e.g. *contribuyo, contribuyen...)* in a regular cuasi-morpho-phonological process to repair hiatus. Similarly, the 49 verbs in *-ir* ending in a palatal consonant (e.g. *bullir, gruñir*) can hardly be considered irregular just because they drop /j/ in endings beginning with /je/ /jo/ since native speakers of Spanish know that the phonological structure of their language does not allow them to produce forms like *bullió* or *bruñiera* because a palatal consonant cannot be followed by yod.

Also quite easy to identify are verbs in *-ir* with "e" as their stem vowel. They always present vowel alternation, so considering non-alternation to be their regular way of inflection when this is unattested may not be the best analysis. These verbs most usually either change "e" to "i" in predictable slots in their paradigm (e.g. *seguir, medir, freír, pedir, concebir, repetir, reír, regir, elegir, corregir, vestir...*) or display a predictable mix of "i" and "ie" (e.g. *advertir, erguir, herir, ingerir, invertir, mentir, preferir, sentir...*). These classes contain dozens of members as well (49 and 62 verb types respectively according to Carreras Riudavets *et al.* (2010)).

The verbs of all the above mentioned inflectional classes, some of which have hundreds of members and are bigger than the traditional "regular" classes 2 and 3, were included by Fratini *et al*. (2014) into their list of irregular verbs without any previous argumentation:

| -ar | -cer | -ir | -ar+diph. | -er | -er+diph. | -uir | e-ir(i) | e-ir(ii) |
|---|---|---|---|---|---|---|---|---|
| 9615 | 361 | 354 | 349 | 151 | 77 | 65 | 62 | 49 |

Table 1: Spanish verbal classes ordered by size. Those traditionally considered regular are shaded.

It is quite clear, therefore, concerning regularity, that between one extreme (the *-ar* class without alternations) and the other (probably the verb *ser* 'to be') there is a continuum with many different inflectional classes of various sizes. So many morphologists (e.g. Bybee 1991: 84-86 or Haspelmath & Sims 2010: 159) have noticed this lack of a sharp boundary between regularity and irregularity that the continuum-like character of the opposition is even considered by some to be "largely uncontroversial" (Stolz *et al.* 2007:22). Failing to acknowledge the existence of this continuum or to give it at least some theoretical consideration can only make us more prone to non-optimal experimental settings or approaches. A sizeable proportion of the verbs which Fratini *et al*. (2014), following traditional practice, classified as irregular are, therefore, highly disputable, since they belong to very big inflectional classes like the ones identified above. The high number of irregular verbs they used (more than 650) is by itself quite revealing of this onset problem in their study.

If the frequency of regular and irregular verbs wants to be studied, the safest approach would probably be to select unmistakable cases of each by disregarding at least most of the fuzzy continuum between the two poles. This would ideally have meant comparing class 1 verbs with verbs not conjugated like any other (in Spanish these verbs would be *dar, poner, haber, tener, andar, ser, hacer, estar, querer, ir, venir, caber, caer, traer, poder, ver, decir, saber* and *valer*) or at least with verbs whose inflectional class numbers in the dozens and not in the hundreds. If the frequency of regular and irregular verb forms (not verb types) is the object of study, a suitable approach would be, in a verbal system as complex as the Spanish, to focus on a single slot of the paradigm at a time. This is the approach of later sections 3.3 and 3.4.

### 2.3 Prefixed irregular verbs[3]

Another aspect of the research by Fratini *et al*. (2014) which might be problematic was that, in the list of irregular verbs they used, there is a considerable proportion of verbs which are formed by the adding of a certain prefix to another irregular verb. Examples include *ab-negar, abs-tener, abs-traer, ad-venir, ante-poner, ante-ver*... The vast majority of these involve the formation of a much more infrequent verb out of a much more frequent irregular verb. For every underived irregular verb (e.g. *venir*) a great number of derived variants has been introduced: *antevenir, advenir, avenir, contravenir, convenir, desavenir, devenir, entrevenir, intervenir, reconvernir, prevenir, provenir, sobrevenir, revenir, subvenir*.

These derived variants of more basic irregular verbs are parasitic on them in that their conjugation is based on that of their basic verb source. However, unlike the base forms, derived variants might well be distributed randomly across the frequency range, which would introduce a lot of “noise” in the data, especially in the most common frequency ranges. It is not difficult to imagine how, if in English the researchers' list of irregular verbs would have included hundreds of verbs of this kind (e.g. *a-bear, a-bite, ac-know, be-fly*...) they would also have had much more difficulty in finding a correlation between irregularity and frequency.

Since these derived verbs are dependant on their base verbs and ultimately on their frequency for existence, this constitutes a big dilemma concerning how these derived verbs should be treated. A possibility to deal with this problem would be to assign to them the frequency of their bases or to add the frequencies of the derived verbs and their base verb. Probably the safest option, however, given the complexity of the challenges presented by these verb types, would be not to include them in any study attempting to elucidate the possible correlation between frequency and irregularity since these verbs are obviously “playing the game” not only with their own frequency but also, to some extent, with that of their bases.

### 2.4 Defunct irregular verbs

Another potentially weak aspect of the research by Fratini *et al*. (2014) is the vast number of extremely infrequent irregular verbs they included in their list. In this respect they simply refer back to Villar (2001) as the source for these verb types. Most grammarians, however, like to assemble as many irregular verbs as possible for the sake of exhaustiveness without this implying that all their verb types are actually in use in the speech community. One has to be, therefore, especially scrupulous in this respect. Fratini et al (2014) tackled this problem by doing a normative study in which only those irregular verbs were selected which were known by at least 4 out of 11

[3] For a criticism similar to the one presented in this section see Bybee (2007: 176-177)

speakers. I consider this number insufficient, specially since their regular verbs, by contrast, were known at least by 8 out of 11 speakers and most usually (94.5%) by all of them. These double standards are not easy to justify and they are liable to causing an important bias in the respective frequencies of one kind of verbs and the other.

Any native speaker of Spanish going through the two lists will notice the big difference between them in this respect as they will fail to use or have heard many of the verbs in the irregular verbs' list but not on the one containing regular verbs. Within the first 100 irregular verbs in Fratini et al. (2014)'s list we find for example *ablandecer, abnegar, aforar, aforarse, amarillecer, amodorrecer, anteferir, apedrar, asonar, aspaventar, carcomecer, circunvolar*... Their verb forms are all exceedingly infrequent (q<0.01 per million words in Ngram viewer)[4]. Note, as a way of comparison, that even the archaic form *truje* for *traje* 'I brought' is still 0.05 pmw in Ngram viewer or CREA despite its defunct status in the modern language.

A suitable solution to this bias would have to involve, first, having the same threshold for the inclusion of regular and irregular verbs. It might not be necessary to resort to preliminary normative studies for this; the threshold below which a given form is excluded because it is considered to be no longer in use can be established with respect to usage data alone from a given corpus. The choice of the concrete frequency threshold will be arbitrary to a big extent but should crucially remain the same for regular and irregular verbs. Secondly but also crucially, the criterion used to select the verb types in the list of regular verbs would have to be made explicit since, without randomization in this respect, we cannot know whether the frequency properties of the selected regular verbs will approximate those of the class of regular verbs as a whole. None of these two requirements were met in the research by Fratini *et al*. (2014).

### 2.5 Handling of data

It has to be stressed that, even after these shortcomings in the setting of their research, the data obtained by Fratini *et al*. (2014) still showed a highly significant (p=0.000) correlation between frequency and irregularity. The mean frequency of irregular forms (despite the inclusion of grammar-book irregular verbs like *abnegar, amarillecer* or *aspaventar*) was still found to be more than twice that of the regular verb forms. It was, thus, only the researchers' subsequent handling and interpretation of that data that led them to conclude otherwise; that (Fratini *et al*. 2014: 297) “[their] results do not support the general correlation between irregularity and frequency”.

Some of that handling involved the discarding of outliers[5]. In practice this probably meant the discarding of most verb forms of underived irregular verbs (e.g. *venir*) but not of their numerous derivates (e.g. *antevenir, advenir, avenir, contravenir, convenir, desavenir, devenir, entrevenir, intervenir, reconvernir, prevenir, provenir, sobrevenir, revenir, subvenir*). Further handling involved the division of the remaining verb forms into three frequency groups. This resulted in a further narrowing of the range of values within which correlations were explored[6]. The overall effect is

[4] This does not mean that these verbs cannot be understood by most native speakers. Since these infrequent verbs are most usually based upon some more frequent word (adjectives like *blando, abnegado, aforado, amarillo* or nouns like *modorra, piedra, aspaviento* or *carcoma*) the semantics of the verbs can remain transparent, which is maybe the reason why they were accepted at least by 4 out of 11 speakers in Fratini *et al.* (2014)'s normative study.

[5] 579 verb forms were discarded in total because of their high frequency. The threshold appears to have been 1000 tokens in CREA, or equivalently a frequency of 6,48 per million words; a frequency similar to that of English words like *utter, resembled, learnt, threatens, paramount, advise* or *Denver.* It is likely that this proceeding effectively excluded most of the truly irregular unprefixed verb forms (see Graphic 3).

[6] Fratini et al. (2014) started their research with 13947 verb forms and end up working with 6393.

that, even under optimal circumstances and with a normal distribution, conditions which are probably far from the present, the possibilities to obtain a statistically significant correlation would be much reduced after these operations:

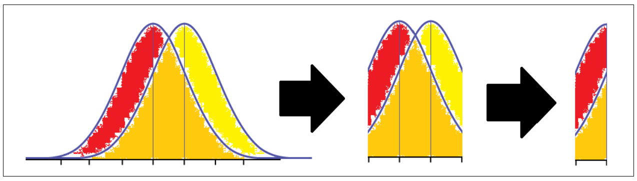


Figure 1: The narrowing of the data ranges

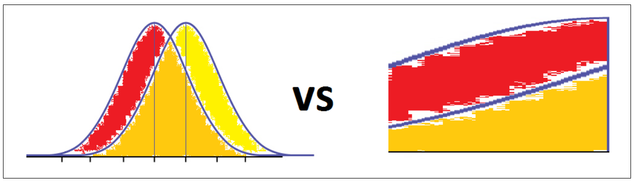


Figure 2: Achieved effect: weaker correlation

Figure 1 shows graphically the way data were processed in Fratini et al. (2014). A first narrowing of the data ranges involved the discarding of outliers. A second one involved splitting the remaining verb forms in 3 frequency groups. The result (Figure 2) is that even if there had been initially a statistically significant difference between the two groups of verb forms, this would have been considerably weakened after the handling of the data in Fratini et al. (2014). If to the narrowing of the data ranges we add the corresponding reduction of the sample size, the possibility to obtain a statistically significant correlation would have been considerably reduced. Note that a correlation was still found within each frequency range although this no longer reached statistical significance.

## 2.5 Conclusion

Because of the reasons commented in Sections 2.2, 2.3 and 2.4, most of the irregular verbs in Fratini *et al*. (2014)'s list should be discarded. Most of them either belong to some of the big or predictable inflectional classes introduced in section 2.2 (i.e. it could be argued that they are regular), or are formed by prefixation from some much more frequent irregular verb, or are defunct, mainly text or grammar book examples of irregular verbs.

Quantifying this claim, of the first 100 verbs in their list[7], at least 84 can be excluded by some of these criteria: As for the objections in Section 2.2, 23 of them are verbs in *-ecer* which as we saw constituted a very large and productive class. So productive, in fact, that even *carcomecer*, one of those verbs in Fratini *et al.* (2014)'s list which is completely absent from both Ngram viewer and CREA (frequency = 0) can still be conjugated by native speakers as *carcomezco, carcomeces... carcomezca, carcomezcas...* without much hesitation despite having never been read or heard before by most of them. This would have hardly been possible if the verb were indeed irregular. Another 20 verbs among their first 100 are diphthongizing verbs in *-ar*, another very numerous inflection class. Another 7 are diphthongizing verbs in -er, another 3 are verbs in -uir etc. As for the objections raised in Sections 2.3 and 2.4, 19 verbs are prefixed derivates of much more frequent irregular verbs and another 23 verbs are absent from the contemporary speech community altogether, applying as a threshold a very undemanding minimum frequency of 0.05 per million words to the verb type as a whole in Ngrams Viewer.

[7] These are: *abastecer, abducir, ablandecer, abnegar, aborrecer, absolver, abstener, abstraer, acaecer, acertar, aclacecer, acontecer, acordar, acrecentar, acrecer, adherir, adolecer, adormecer, adquirir, aducir, advenir, advertit, aflorar, aforar, aforarse, agradecer, alentar, almorzar, amanecer, amarillecer, amoblar, amodorrecer, amolar, andar, anegar, anochecer, anteferir, anteponer, antevenir, apacentar, apacer, aparecer, apedrar, apetecer, apretar, aprobar, argüir, arrepentir, ascender, asentar, asentir, aserrar, asir, asonar, aspaventar, atañer, atardecer, atender, atener, atentar, autoabastecerse, avenir, aventar, avergonzar, balbucir, bendecir, bienquerer, blanquecer, bullir, caber, caer, calentar, carcomecer, carecer, cegar, ceñir, cerner, cernir, circunvolar, clarecer, cocer, coextenderse, colar, colegir, colgar, comedir, comenzar, compadecer, comparecer, competir, complacer, componer, comprobar, concebir, concernir, concertar, concluir, concordar, condecir* and *condescender.*

## 3. Investigating the relation between frequency and irregularity

### 3.1 Preliminary observations

When trying to investigate the correlation between irregularity and frequency it is necessary to take into account these preliminary theoretical issues which have been presented so far: First it is necessary to avoid verb types or verb forms difficult to classify clearly as either regular or irregular. Secondly, one has to avoid derivate verbs which are parasitic, so to speak, on the morphology and frequency of their base verbs[8]. Third, one has to avoid verb types and forms which do not have an existence beyond grammar books and academic discussions. Finally, these criteria and those applied in the choice of the verbs which will be used in the research have to be explicit and remain constant for regular and irregular verbs.

Trying to investigate the correlation between frequency and irregularity in paradigm-rich languages like Spanish may be very difficult if all the verb forms are studied simultaneously, on the one hand, because of the vast amount of data alone. On the other, because many verb forms are bound to be homographs with other words. Fratini *et al*. (2014), for example, stated that they were aware from the onset of this problem and claimed to have removed "most of" those forms, however, in the little information they provide about the actually inspected verb forms, we read about (Footnote 4) *importe, doble, lucha, cosa, firma, libres, cena, meta, salto, canto, calma, parto, gobierno, cuenta, pueblo*... all of them verb forms homograph with other very frequent words.

An alternative possibility is therefore proposed. If like Fratini *et al*. (2014) one wishes to analyze the frequency of verb forms and not verbs types[9] (although, of course, the frequency of a verb type and that of its corresponding verb forms will be very strongly correlated in any case), a good idea could be to focus on a single cell of the paradigm and then compare the frequencies of regular and irregular forms. The simple future forms may be good candidates in this respect. On the one hand, because of their stress in the last syllable, they are unlikely in Spanish to be homographs with other words. On the other hand, the morphology of these forms is relatively new from a diachronic perspective. That makes it easy to classify a given form as regular or irregular since there are no competing morphs and the endings *-é -ás -á -emos -éis -án* are in use with all verbs. 99.9% of the verbs just add these endings to the infinitive form (e.g. *morir>moriré, amar>amaré, ser>seré*). It is thus not too risky to classify as irregular the very few forms which do not follow this rule. The correlation of frequency and irregularity in the third person singular of the future tense will be analyzed as the most frequent form in the future paradigm.

[8] In this sense, it is also problematic how to treat verbs like *abducir* or *inquirir.* Even if there is no verb **ducir* or **quirir*, those bases appear in many other verbs (e.g. *reducir, conducir, inducir, inquirir, adquirir*) and thus these words may not be comparable to morphologically undecomposable verb types.

[9] The decision at which level to study the relation between (ir)regularity and frequency is very important. Irregularity in Spanish verbs is very rarely (e.g. *sé* 'I know', *eres* 'you are', *roto* 'broken') confined to an individual verb form. Instead, verb forms are most usually organized into bigger groups, internally homogeneous regarding (ir)regularity. Thus, if the 1SG.FUT is irregular, the same irregularity will we found in the rest of the future and the conditional forms (i.e. *saldré* implies *saldrá, saldría* etc.). If a verb has some root irregularity in the 1SG.PRES, it will also have it in the subjunctive (i.e. *tengo* implies *tenga, tengas* etc. and *quepo* implies *quepa, quepas* etc.). If the 2PL.PAS has a suppletive root, the rest of the past forms too (i.e. *cupisteis* implies *cupe, cupiste* etc. *tuvisteis* implies *tuve, tuviste* etc.). All of the forms conforming these groups contribute to a given irregularity. Thus, individual forms like *quepo* or *cupisteis* may be infrequent but the rest of the forms in their group need not be and they also provide, indirectly, evidence of the irregularity of other forms. Given the organization of the Spanish verb paradigms, therefore, one may wonder whether studying (ir)regularity at the level of the individual verb form is the most sensible choice.

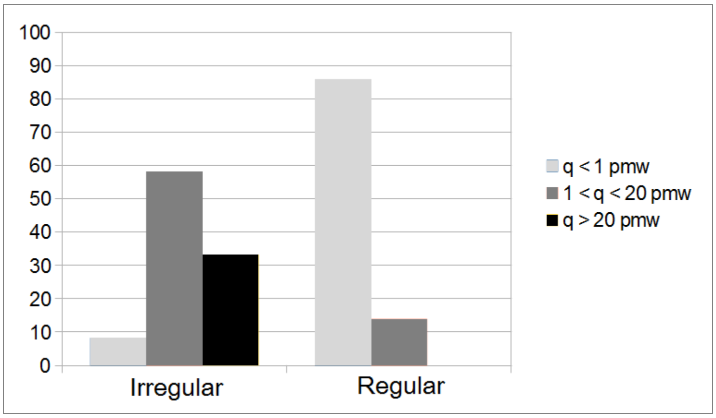
### 3.2 The setting

The only unprefixed verb forms which are irregular in Spanish in the third person singular of the future tense are *podrá, saldrá, vendrá, querrá, pondrá, habrá, cabrá, dirá, hará, vendrá, sabrá* and *tendrá (ante-pondrá, des-hará* etc. have been excluded, see Section 2.3). Those verbs also have the same irregularity in the rest of the persons in the paradigm. The fact that there are only 12 verbs which are irregular in this slot allows us, quite conveniently, to consider all of them in this analysis.

To contrast them to irregular verbs, 50 other verbs having a regular third person singular future have been selected by means of a random verb generator[10]. In agreement with the requirements in Section 2, these verbs were checked to be non-derived and above the frequency threshold of 0.05 per million words. Some of the generated verbs were discarded because of this: *desmontar* and *representar* for being derived and *amancebar, varear, gorronear, afrendar, molificar, guatear, zurear, marinar, chapar* and *repujar* for being below the required frequency threshold. These verbs were replaced by other randomly generated verbs until reaching the goal number of 50. The analyzed verbs were: *excusar, llevar, zurrar, trasvasar, encabezar, vagabundear, juntar, ir, castigar, zozobrar, silenciar, marchar, descoyuntar, quemar, ejemplarizar, arreglar, orientar, moderar, finalizar, zigzaguear, ensalzar, coronar, finiquitar, ultrajar, entender, vigilar, libertar, fustigar, nublar, marcar, legar, balbucear, herir, echar, hartar, esquivar, perforar, cargar, volver, quebrantar, picar, mezclar, irritar, nivelar, propinar, fastidiar, juzgar, gestionar, implicar* and *inquietar.* The search in CREA of their third person singular future yielded the following results.

### 3.3 Results

The mean frequency of the irregular verb forms was found to be 36,55 per million words (pmw) whereas for the regular verb forms it was 1,43 pmw[11], note the vast difference. The standard deviation was 43,07 for the irregular verbs and 4,35 for regular verbs. This is not surprising given that the lower frequency rages are the ones concentrating most of the regular verb forms. By performing a Mann-Whitney U-test, these differences have been found to be, of course, statistically significant (p=0,004).



Graphic 1: Frequency of irregular and regular FUT.3SG verb forms

An additional observation which supports the correlation between irregularity and frequency concerns the competition between regular and irregular variants for the same cell in the paradigm

[10] Available online at [http://www.danielpinero.com/random-words-generator-spanish]. Accessed 15/12/2015.

[11] Remember at this point that verb forms occurring with a frequency above 6.48 pmw had been discarded by Fratini *et al.* (2014) as “outliers”.

of a lexeme. It has been noted before (Bybee 1995: 236) that relatively infrequent irregular forms (like *wept* or *leapt* in English) tend to be regularized more frequently than very frequent ones (such as *kept* or *slept*). We can find the same tendency in the presently analyzed forms. If we compare in Google N-grams the frequency of use of the very frequent *habrá* (63 pmw), the intermediate *sabrá* (4,7 pmw) and the relatively infrequent *cabrá* (0,52 pmw) to the frequency of their regularized variants *haberá, saberá* and *caberá* we find that regularization is most frequent in the lesser used verb forms. Thus, we find that the least frequent irregular, *cabrá,* is “only” 54 times more frequent than its regularized version *caberá; sabrá,* in turn, is 770 times more frequent than *saberá* and the most frequent *habrá* is 3937 times more frequent than *haberá*. This pattern is not difficult to understand. In Pinker's (1999: 10) words “[a] simple explanation is that irregular forms (…) have to be memorized repeatedly, generation after generation, to survive in a language”. True irregulars can only be acquired successfully, therefore, if they are frequent enough.

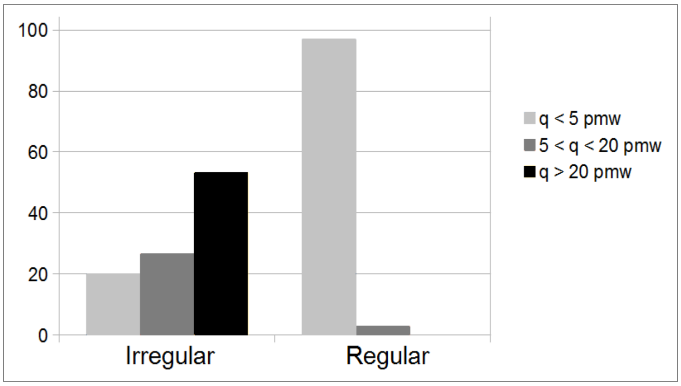
### 3.4 A re-test

If the pattern observed for the third person singular future is representative of the verbal system as a whole, a similar analysis of the frequency of regular and irregular forms in other cells of the paradigm should reveal similar numbers. Much like the future, the preterite forms are also convenient targets for analysis because a relatively clear distinction is found here as well between regular and irregular forms. The preterite forms of the verb are always stressed, throughout all persons, in 99.9% of the verbs. They receive the endings *-é -aste -ó -amos -asteis -aron* or *-í -iste -ió -imos -isteis -ieron* depending on whether they belong to the first or second/third conjugations respectively. It is therefore not too risky to consider exceptional those few verbs which do not follow this pattern.

The only morphologically simple verbs which do not follow the above rules are *andar, tener, estar, poner, saber, caber, traer, venir, querer, hacer, decir, ir, ser, haber* and *poder.*[12] These verbs have a different set of suffixes and, apart from *ir* and *ser,* unstressed first and third person singular endings. The third person plural has been chosen for analysis over the singular to prevent homograph forms like *cupo* 'fit.PAS.3SG'/'quota' or *vino* 'come.PAS.3SG'/'wine'. An additional 50 regular verbs has also been randomly obtained again for contrast: *fracasar, juntar, notar, yacer, juzgar, narrar, hundir, nivelar, mediar, nominar, llenar, jabonar, nublar, zigzaguear, simbolizar, nacer, entregar, nutrir, igualar, atacar, hastiar, glosar, lucir, parir, golpear, eximir, hallar, oprimir, zanjar, planear, votar, proyectar, incluir, sumergir, velar, irritar, sollozar, tergiversar, defender, intuir, pegar, quemar, buscar, vocear, liderar, generalizar, oscurecer, valer, untar, husmear.*

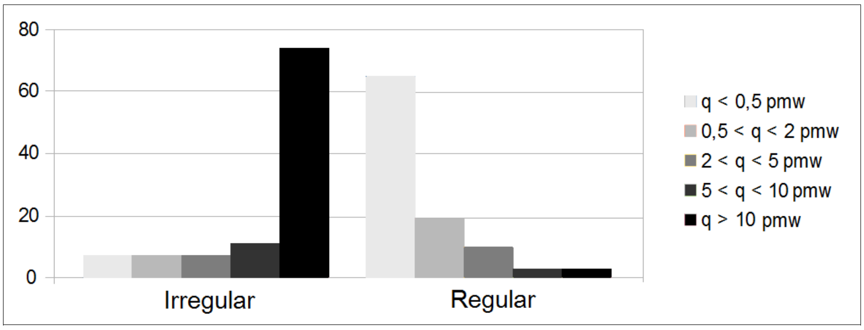
The obtained results once again evidence the clear correlation between frequency and irregularity. The mean frequency of the irregular verb forms in this cell was found to be 49,48 per million words (pmw) whereas for the regular verb forms it was 1,20 pmw. The standard deviation was 85,29 for the irregular verbs and 1,81 for regular verbs. A Mann-Whitney U-test shows these differences to be, again, statistically highly significant (p=0,00014).

[12] Verbs like *conducir, aducir, reducir, inducir, abducir*... have been excluded since they involve prefixation. The verb *dar* is also exceptional in that in the preterite it behaves as a second or third conjugation verb. This kind of irregularity (so-called heteroclisis), however, is only revealed at the paradigm level and not at the cell level analyzed here so the verb has not been considered irregular in this cell.



Graphic 2: Frequency of irregular and regular PAS.3PL verb forms

Putting together the results obtained for the two cells might be a little bit like adding apples and oranges but reveals, with more significant numbers, that irregular verb forms tend to cluster in relatively high frequencies whereas regular verb forms do so in the lowest frequency ranges:



Graphic 3: Frequency of the analyzed irregular and regular verb forms

These patterns and results are robust enough not to be the result of chance or some bias in the chosen forms. My contention is that the same pattern would also be obtained in any other cell like, for example, the imperative (where irregular forms would be those of the very frequent verbs *tener, poner, venir, salir, hacer* and *decir*) or the participle (with the verbs *abrir, cubrir, decir, escribir, hacer, morir, poner, romper, ver* and *volver* generating irregular forms). The fact that many of the verbs (e.g. *decir, tener, poner, venir* or *hacer*) are found to be irregular in most of the mentioned cells is a result of the fact that irregularity is found, overwhelmingly, in the most frequent verbs.

## 4 Conclusion

As shown repeatedly in literature (e.g. Erker & Guy 2012 vs Bayley 2013) methodological choices are decisive when it comes to operationalizing even *a priori* simple notions like ‘regularity’ or ‘frequency’. The present paper has argued that the research conducted by Fratini et al. (2014) had relevant shortcomings in this respect which have adversely affected the validity of their conclusions. Any investigation dealing with the opposition of regularity and irregularity will have to come to terms with the fact that this is not a dichotomous dimension. Likewise, researchers studying synchronic frequency (i.e. contemporary use) need to make sure that the items they study are still alive in the speech community in general and do not merely appear sporadically in academic linguistic writings. In addition, if the object of study is morphological irregularity and its correlation with frequency, it has to be taken into account that the knowledge which speakers have of the morphological systems and paradigms of their languages is very complex and not completely understood. It seems, however, safe to say that verb forms like *venir* and *pre-venir, vino* and *pre-vino*, *vendrá* and *pre-vendrá* etc. are associated in the minds of speakers. Every use of *vendrá*, therefore, may well, to some extent reinforce not only *vendrá*, but also *pre-vendrá*, *contra-vendrá*, *con-vendrá* etc. The existence of deviations in this respect (e.g. *decir* > *pre-decir; dirá* > *pre-decirá*, **pre-dirá*) sure reminds us that the association is not perfect, however, until understanding perfectly the nature and strength of those associations, those verb types should be used with great caution in relation with frequency.

Contrary to the conclusions of Fratini *et al.* (2014), my corpus research has found a statistically significant difference between the frequency of regular and irregular verb forms in Spanish. In agreement with the vast majority of earlier insights into this topic (e.g. Ullman 1999), irregular forms have been found to be, on average, much more frequent than regular forms. The reason for this is not difficult to understand. Irregular forms that are indeed truly irregular have to be learnt independently. This can hardly happen unless that particular form is encountered frequently enough. Trying to determine the approximate value of that frequency threshold is a legitimate research objective. Future efforts could also be aimed at the measurement and quantification of the regularity/irregularity of a given inflectable word, probably by looking at the size of its inflectional class but also at its connection to other classes, the principal parts in its paradigm, observed analogical tendencies etc. The fact, however, that, in language after language, irregular verbs and nouns are predominantly found among the most frequent (e.g. *be, go, come, give, see; person, man, woman, child*...) should suffice to conclude that frequency and irregularity are indeed highly correlated variables, also in Spanish.

**Bibliography**

Bayley, Robert; Greer, Kristen & Holland, Cory. (2013). Lexical frequency and syntactic variation: A test of a linguistic hypothesis. University of Pennsylvania Working Papers in Linguistics 19.2: 4.

Balaguer, R. De Diego; Sebastian-Galles, Nuria; Díaz, Begoña & Rodriguez-Fornells, Antoni. (2005). Morphological processing in early bilinguals: An ERP study of regular and irregular verb processing. *Cognitive Brain Research 2*5,*1*. 312-327.

Bowden, Harriet Wood, Gelfand, Mathew P.; Sanz, Cristina & Ullman Michael T. (2010). Verbal inflectional morphology in L1 and L2 Spanish: A frequency effects study examining storage versus composition. *Language Learning 60, 1.* 44-87.

Bybee, Joan L. (1991). Natural morphology: The organization of paradigms and language acquisition. *Crosscurrents in second language acquisition and linguistic theories.* Amsterdam: Benjamins: 67-92.

Bybee, Joan L. (1995). Diachronic and typological properties of morphology and their implications for representation. *Morphological aspects of language processing*: 225-246. Hove: Psycology Press.

Bybee, Joan. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Carreras Riudavets, Francisco J., Hernández Figueroa, Zenón J. & Rodríguez Rodríguez, Gustavo. (2010). La conjugación de verbos en español y su morfología. Morrisville: Lulu

CREA, REAL ACADEMIA ESPAÑOLA: Corpus de referencia del español actual. [on line: http://corpus.rae.es/creanet.html]. Accessed 15 December 2015.

Erker, Daniel & Guy, Gregory R. (2012). The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language 88.3*: 526-557.

Fratini, Viviana; Acha, Joana & Laka, Itziar. (2014). Frequency and morphological irregularity are independent variables. Evidence from a corpus study of Spanish verbs. *Corpus Linguistics and Linguistic Theory* 10, 2: 289-314.

Google Books Ngram Viewer. http://books.google.com/ngrams. Accessed 20/12/2015.

Haspelmath, Martin & Sims, Andrea. (2010). Understanding Morphology. Oxford: Oxford University Press.

Pinker, Steven. (1999). *Words and Rules*. New York: Harper Perennial.

Say, Tessa & Clahsen, Harald. (2002). Words, rules and stems in the Italian mental lexicon. *Storage and Computation in the Language Faculty*: 93-129. New York: Springer.

Stolz, Thomas; Otsuka, Hitomi; Urdze, Aina & van der Auwera, Johan (Eds.). (2007). *Irregularity in Morphology (and beyond).* Berlin: Akademie Verlag.

Ullman, Michael T. (1999). Acceptability Ratings of Regular and Irregular Past-tense Forms: Evidence for a Dual-system Model of Language from Word Frequency and Phonological Neighbourhood Effects. Language and Cognitive processes, 14(1): 47-67.

Veríssimo, Joao & Clahsen, Harald. (2009). Morphological priming by itself: A study of Portuguese conjugations. *Cognition*, *112:* 187-194.

Villar, Celia. (2001). Guía de verbos españoles. Madrid: Espasa.

Zipf, George Kingsley. (1935). The psycho-biology of language. Boston: Houghton Mifflin Co.