Morphological Irregularity Correlates with Frequency

Shijie Wu
Department of Computer Science
Johns Hopkins University
Baltimore, US
shijie.wu@jhu.edu

Ryan Cotterell
The Computer Laboratory
University of Cambridge
Cambridge, UK
rdc42@cam.ac.uk

Timothy J. O'Donnell
Department of Linguistics
McGill University
Montréal, Canada
timothy.odonnell@mcgill.ca

Abstract

We present a study of morphological irregularity. Following recent work, we define an information-theoretic measure of irregularity based on the predictability of forms in a language. Using a neural transduction model, we estimate this quantity for the forms in 28 languages. We first present several validatory and exploratory analyses of irregularity. We then show that our analyses provide evidence for a correlation between irregularity and frequency: higher frequency items are more likely to be irregular and irregular items are more likely be highly frequent. To our knowledge, this result is the first of its breadth and confirms longstanding proposals from the linguistics literature. The correlation is more robust when aggregated at the level of whole paradigms—providing support for models of linguistic structure in which inflected forms are unified by abstract underlying stems or lexemes. Code is available at https://github.com/shijie-wu/neural-transducer.

1 Introduction

Irregularity is a pervasive phenomenon in the inflectional morphology of the world's languages and raises a number of questions about language design, learnability, and change. Nevertheless, irregularity remains an understudied phenomenon and many basic questions remain unanswered (Kiefer, 2000; Stolz et al., 2012). Do all languages exhibit irregularity? What is the relationship between irregularity and frequency? Is irregularity best thought of as a property of individual forms, or a property of more abstract objects like morphological paradigms? In this paper, we examine these questions, focusing in particular on the relationship between irregularity and frequency.

One of the fundamental challenges in studying irregularity is defining the phenomenon in a way that is applicable across languages. We begin the paper by addressing this question. First, we formalize the problem of inflectional morphology and present a novel, information-theoretic measure of the degree of irregularity of an inflected form. This definition builds on recent work that defines (ir)regularity in terms of the probabilistic predictability of a form given the rest of the language (Cotterell et al., 2018a; Ackerman and Malouf, 2013). Making use of a state-of-the-art model of morphological inflection, we estimate our measure of irregularity across a large number of word forms from 28 languages drawn from the UniMorph database (Kirov et al., 2018). Based on these estimates we perform three studies. First, we validate our estimates by examining the predictions on English past tense forms—showing that the model's predicts accord with human judgements of irregularity. We also examine the overall rate of accuracy of our model. Second, we examine the degree of irregularity across languages, showing that the model predicts wide variance in the average amount of irregularity between the languages in our sample. Finally, we provide empirical evidence for a correlation between irregularity and frequency across languages. While this relationship has been observed for individual languages (e.g., English: Marcus et al., 1992; Bybee, 1985), this is the first confirmation of the effect across this many languages. This result is especially relevant given recent discussions calling the relationship into question (e.g., Fratini et al., 2014; Yang, 2016). We find, furthermore, that the correlation between irregularity and frequency is much more robust when irregularity is considered as a property of whole *lexemes* (or *stems*/*paradigms*) rather than as a property of individual word forms. We discuss the implications of these findings.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5117–5126
Florence, Italy, July 28 - August 2, 2019.

2 Formalizing Inflectional Morphology

In this work, each word type is represented as a triple consisting of the following components:

- A **lexeme**[1] ℓ: An arbitrary integer or string that indexes an abstract word (e.g., GO, which provides an index to forms of the verb *go* such as *goes* and *went*).
- A **slot** σ: An arbitrary integer, string, or more structured object that indicates how the word is inflected (e.g., $[pos{=}\text{v}, tns{=}\text{past}, person{=}\text{3rd}, num{=}\text{sg}]$ for the form *went*).
- A **surface form** w: A string over a fixed phonological or orthographic alphabet Σ (e.g., *went*).

A **paradigm** $\boldsymbol{\ell}$ (boldface ℓ) is a lexeme-specific map from slots to surface forms for lexeme ℓ.[2] Typically, slots are indexed by structured entities—known as *morpho-syntactic feature vectors* or *morpho-syntactic tags*—represented by a set of key-value pairs: $\sigma = [k_1{=}v_1, \ldots, k_n{=}v_n]$. For example, the English verb form *runs*, which has the feature vector $[tns{=}\text{pres}, per{=}\text{3rd}, num{=}\text{sing}]$. In what follows, the keys k_i and the corresponding values v_i are taken from the universal inventory, defined by the UniMorph annotation scheme and denoted $\mathcal{M}$ (Kirov et al., 2018). We use dot notation to refer to specific forms or sets of forms in a paradigm indexed by some slot $\text{GO}.past = went$.

Given the pieces just sketched, a complete model of inflectional morphology will specify a joint distribution over surface forms, lexemes, and slots, that is $\mathbb{P}(w, \ell, \sigma)$, or one of its associated conditional distributions, such as $\mathbb{P}(\ell, \sigma \mid w)$—the distribution over lexemes and features, given a surface form; or $\mathbb{P}(w \mid \ell, \sigma)$—the conditional probability of a surface form given a lexeme and inflectional features. In this paper, we will focus on the latter, defining a probabilistic model to approximate this distribution and using that to estimate degrees of irregularity.

[1] This terminology is characteristic of *word-and-paradigm* approaches to morphology. In *item-and-arrangement* approaches, this might be called the *stem* (Hockett, 1954).

[2] See (Baerman et al., 2015, Part II) for a tour of alternative views of inflectional paradigms.

3 Operationalizing Irregularity

The informal distinction between regular and irregular forms is an important one for many theories of grammar (e.g., Siegel, 1974), language processing (e.g., Hay, 2003), and language acquisition (e.g., Pinker, 1999; Marcus et al., 1992; McClelland and Patterson, 2002a,b; Pinker and Ullman, 2002b,a; Rumelhart and McClelland, 1986; Prasada and Pinker, 1993; Pinker and Prince, 1988). However, there have been few proposals for how the notion can be characterized precisely or measured quantitatively.

Clearly, the regularity of a form (or rule) can only be defined with respect to the language as a whole—what makes something irregular is that it does not behave in the way that would be expected given other forms in the language. But what is meant by *expected*? Here, we follow recent work by defining the notion of expectedness in terms of a probabilistic model of inflection which approximates $\mathbb{P}(w \mid \ell, \sigma)$ (Cotterell et al., 2018a; Ackerman and Malouf, 2013). However, there remains a wrinkle. A form like *went* is highly expected as the past tense of GO for an adult speaker of English, but is also irregular. How do we capture this?

We take the correct notion of expectedness to be the expectedness of the word form treated *as if* it were the first instance of that lexeme which had been observed. Thus, we base our measures of regularity on the conditional probability of a word type w given the rest of the forms in the language with the target lexeme removed.

$$\mathbb{P}(w \mid \ell, \sigma, \mathcal{L}_{-\ell}) \tag{1}$$

Of course, since the target language $\mathcal{L}$ is generally infinite, we will need to make use of some model-based estimate of this probability $p_{\boldsymbol{\theta}}(w \mid \ell, \sigma, \mathcal{L}_{-\ell})$. In essence, our definition of irregularity is based on *wug-testing* (Berko, 1958) such a probabilistic model to see how robustly it generalizes to the target form w. In practice, we will estimate this quantity by performing a holdout evaluation of the target form under our model.

More irregular forms will tend to have a lower *wug*-test probability $\mathbb{P}(w \mid \ell, \sigma, \mathcal{L}_{-\ell})$ than most regular forms. However, the absolute value of such a probability is not directly interpretable. To turn these probabilities into interpretable values which directly measure irregularity, we take the negative log odds of the probability of the correct word

form.

$$\iota(w) = -\log\left[\frac{\mathbb{P}(w \mid \ell, \sigma, \mathcal{L}_{-\ell})}{1 - \mathbb{P}(w \mid \ell, \sigma, \mathcal{L}_{-\ell})}\right] \quad (2)$$

We refer to this quantity as the *degree of irregularity* of a form. If probability of the correct form w is exactly 0.5, then eq. (2) will be 0. However, if $\mathbb{P}(w \mid \ell, \sigma, \mathcal{L}_{-\ell}) > \sum_{w' \neq w} \mathbb{P}(w' \mid \ell, \sigma, \mathcal{L}_{-\ell})$, then eq. (2) will be negative. Otherwise, the quantity is positive. In other words, the metric is more strongly positive when a form is less predictable given other forms in the language and more strongly negative when a form is more strongly predictable. The midpoint at 0 occurs when there is an equal amount of probability mass on the correct form and all other forms.

Note that this definition of ι neatly addresses several challenges in studying the notion of (ir)regularity. First, it doesn't require us to define a binary notion of regular versus irregular or even to explicitly define any such notion at all—a model may treat regularity as an implicit rather than explicit feature of a form or paradigm. Second, and relatedly, we do not require data annotated with the regularity of forms to train or test our model. Third, this definition inherently captures the idea of degree of regularity, for instance, capturing the distinction between wholly suppletive forms such as *went* and semi-productive inflectional classes such as *ring*/*rang*, *sing*/*sang*, etc. Fourth and finally, regularity is known to be correlated with other features of morphological structure, such as productivity. Our definition sidesteps the tricky issue of disentangling these different properties of inflection.

Note that our definition of ι conditions on $\mathcal{L}_{-\ell}$—the language without the target *lexeme*—rather than on $\mathcal{L}_{-w}$—the language without the target *word*. Thus, we are measuring the probability that the model will generalize to the correct form without any evidence of a lexeme at all. Thus, we rule out predictability that comes from similar forms within a paradigm $\boldsymbol{\ell}$. For example, in our approach a model cannot make use of the irregularity of the past tense form *ring* to guess that the past participle form was more likely to be *rung*. We discuss the implications of this assumption in more detail below §5.4.

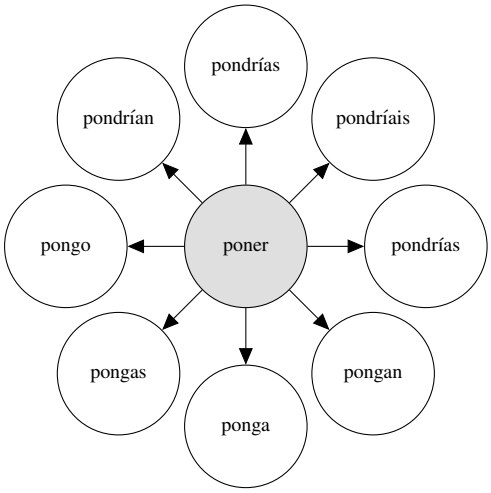

Figure 1: Lemma paradigm tree

4 Modeling Morphological Inflection

Our goal is to estimate $\mathbb{P}(w \mid \ell, \sigma, \mathcal{L}_{-\ell})$ from data. We do this by using a structured probabilistic model of string transduction which we call $p_{\boldsymbol{\theta}}$. In the following sections, we describe this model, how we handle syncretism in the model, our training (holdout and test) scheme, and our estimates of the degree of irregularity ι.

4.1 A Lemma-Based Model

In linguistic morphology, a major division is between *item-and-arrangement* or *morpheme-based* models and *word-and-paradigm* or *word-based* models (Hockett, 1954). Following (Cotterell et al., 2017b), we adopt a word-based approach. To do this, we designate a unique surface form for each paradigm $\boldsymbol{\ell}$ known as the *lemma*. The lemma is associated with a slot which we notate $\check{\sigma}$: $\boldsymbol{\ell}.\check{\sigma} \in \Sigma^*$. The lemma can be thought of as a dictionary or citation form of a word and is traditionally chosen by lexicographers of a language. For example, in many Western European languages the lemma of verb forms is the infinitive. Figure 1 shows several of the forms of the Spanish verb *poner* ("to put") organized around the lemma form. In what follows, we use the lemma to identify lexemes, and wherever a probability distribution would condition on the abstract lexeme ℓ we instead condition on the lemma $\boldsymbol{\ell}.\check{\sigma}$.

Our probabilistic model of string transduction $p_{\boldsymbol{\theta}}$ is a monotonic model with hard attention described in Wu and Cotterell (2019) and can be viewed as a graphical model over strings like the one shown in

	SG	PL	SG	PL
NOM	***Wort***	***Wörter***	*Herr*	***Herren***
GEN	*Wortes*	***Wörter***	***Herrn***	***Herren***
ACC	***Wort***	***Wörter***	***Herrn***	***Herren***
DAT	*Worte*	*Wörtern*	***Herrn***	***Herren***

Table 1: Full paradigms for the German nouns *Wort* ("word") and *Herr* ("mister") with abbreviated and tabularized UniMorph annotation. The syncretic forms are bolded and colored by ambiguity class. Note that, while in the plural, the nominative and accusative are always syncretic across all paradigms, the same is not true in the singular.

Figure 1. It is expressed as follows.

$$p_{\boldsymbol{\theta}}(w \mid \boldsymbol{\ell}.\breve{\sigma}, \sigma, \mathcal{L}_{-\ell}) = \sum_{\mathbf{a} \in \mathcal{A}(w, \boldsymbol{\ell}.\breve{\sigma})} p_{\boldsymbol{\theta}}(w, \mathbf{a} \mid \boldsymbol{\ell}.\breve{\sigma}, \sigma, \mathcal{L}_{-\ell}). \tag{3}$$

The definition of the model includes a sum over all monotonic (non-crossing) alignments $\mathcal{A}(w, \boldsymbol{\ell}.\breve{\sigma})$ between the lemma $\boldsymbol{\ell}.\breve{\sigma}$ and the output surface form w. The inner term of this sum is estimated using a sequence to sequence model. The sum itself is computable in polynomial time using a variant of the forward algorithm (Rabiner, 1989). The model achieves state-of-the-art performance on the SIGMORPHON 2017 shared task on morphological reinflection (Cotterell et al., 2017a). We follow the hyperparameter used by Wu and Cotterell (2019).

4.2 Handling Syncretism

Many inflectional systems display *syncretism*—the morphological phenomenon whereby two slots with distinct morpho-syntactic tags may have an identical surface form. In contrast to many models of inflectional morphology, we collapse syncretic forms of a word into a single paradigm slot, thereby assuming that every every surface form w in a paradigm is distinct. An example of such a collapsed paradigm in German is given in Table 1. Our formalization includes a slot that merges the genitive, accusative and dative singular into a single slot due to the word *Herr*.

To accomplish this we assume that each lexeme is associated with a set of *syncretism classes* denoted by $\mathcal{C}^{\ell}$. $\mathcal{C}^{\ell} : \mathcal{M} \rightarrow \mathcal{M}$ is a map from a slot σ to a *citation form slot* σ' which indexed the canonical suface citation form for that combination of features. $\mathcal{C}^{\ell}$ is used to collapse paradigm cells with identical surface forms. For instance, all forms of the lexeme GO are realized as *went* in the English past tense, regardless of person and number features; thus, for example, $\mathcal{C}^{\text{GO}}([\textit{tns} = \text{past}, \textit{per} = \text{3rd}, \textit{num} = \text{sing}]) = \mathcal{C}^{\text{GO}}([\textit{tns} = \text{past}, \textit{per} = \text{2nd}, \textit{num} = \text{plural}])$. We say that two lexemes ℓ and ℓ' are *syncretically equivalent* if $\mathcal{C}^{\ell}(\sigma) = \mathcal{C}^{\ell'}(\sigma)$ for all σ. We assume the mappings $\mathcal{C}^{\ell}$ are known and given in advance in what follows.

We will use this syncretism-collapsed representation for all simulations below. In particular, this assumption will allow us to simply count the surface forms of each word in Wikipedia without dealing with the tricky issue of assigning individual words to the correct combination of morphosyntactic features (see, Cotterell et al., 2018b, for detailed discussion).

4.3 Handling Derived Forms

As discussed above, we hold out whole lexemes, including all of their inflected forms during training. However, derivational morphology presents a potential challenge for this approach. Consider the irregular verb *do/did/done*. This verb appears in a number of derived prefixed forms such as *redo* and *undo*. These forms all inflect identically to the base form *do*—for example, *redo/redid/redone*.[3] If we train our probability model on such derived forms, it is likely to estimate too high a *wug*-test probability for all forms which are built from the shared stem.

To obviate this problem, we remove all derived forms from the data we consider. To do so we develop a heuristic approach to isolate all words that may have been derived from another. Note that a key desideratum of heuristic is that it should be high precision with respect to finding derivational transformation—we would rather overexclude forms as potentially derivative of another, rather than leave a derived form in the data.

We consider a lexeme ℓ' to be derived from a lexeme ℓ if and only if there is a string $s \in \Sigma^{+}$ such that $(\forall \sigma)[\boldsymbol{\ell}'.\sigma = \boldsymbol{\ell}.\sigma \cdot s]$ or $(\forall \sigma)[\boldsymbol{\ell}'.\sigma = s \cdot \boldsymbol{\ell}.\sigma]$ where $s \cdot t$ denotes string concatenation of strings s and t. For example, DO and REDO satisfy this condition, while SING and RING do not. We

[3] An anonymous reviewer points out that in some languages, such as Dutch, forms derived from irregular verbs become regular (e.g., *zeggen/zei* but *toezeggen/toezegde*). In those languages, it should be unnecessary to apply our heuristic approach.

perform a search for candidate s for all pairs of lexemes in each language and remove all ℓ' that meet this criterion.

4.4 Measuring Irregularity

With the above definitions in place, we can define an approximation to our degree of irregularity ι.

$$\iota(w) = -\log \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{\ell}.\check{\sigma}, \sigma, \mathcal{L}_{-\ell})}{1 - p_{\boldsymbol{\theta}}(w \mid \boldsymbol{\ell}.\check{\sigma}, \sigma, \mathcal{L}_{-\ell})} \tag{4}$$

In our analyses below, we will also wish to measure the irregularity of lexemes as a whole. To do this, we take the average irregularity score over the entire paradigm $\boldsymbol{\ell}$.

$$\iota(\ell) = \frac{\sum_{\{(w,\sigma,\ell) \in \boldsymbol{\ell} \mid w \neq \boldsymbol{\ell}.\check{\sigma}\}} -\log \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{\ell}.\check{\sigma}, \sigma, \mathcal{L}_{-\ell})}{1 - p_{\boldsymbol{\theta}}(w \mid \boldsymbol{\ell}.\check{\sigma}, \sigma, \mathcal{L}_{-\ell})}}{|\boldsymbol{\ell}| - 1} \tag{5}$$

5 Studies of Irregularity

The empirical portion of our work consists of three studies. We first validate and examine the accuracy of the model (§5.2.1). Second, we examine the distribution of irregularity across the languages in our sample (§5.3). Finally, we examine the correlation between irregularity and frequency (§5.4). Before presenting these studies we first give an overview of the data and simulations common to all of them.

5.1 Simulations

Data Provenance. All word forms, paradigms, and morphosyntactic features are taken from the UniMorph project (Kirov et al., 2018). Specifically, we examine the following 28 languages: Albanian, Arabic, Armenian, Basque, Bulgarian, Czech, Danish, Dutch, English, Estonian, French, German, Hebrew, Hindi, Irish, Italian, Latvian, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish, Ukrainian, Urdu, and Welsh. The languages come from 4 stocks (Indo-European, Afro-Asiastic, Finno-Urgic and Turkic) with Basque, a language isolate, included as well. Although this sample represents a reasonable degree of typological diversity, the Indo-European family is overrepresented in the UniMorph dataset, as is the case for most current multilingual corpora. However, within the Indo-European family, we consider a diverse set of subfamilies: Albanian, Armenian, Slavic, Germanic, Romance, Indo-Aryan, Baltic, and Celtic. For each subfamily, we sample subset of languages randomly.

All of our form-level frequencies were computed from Wikipedia.[4] Lexeme counts are achieved by summing over all entries in the paradigm associated with a lexeme. In all simulations, we predict the orthographic form of the target word w from the orthographic form of the lemma $\boldsymbol{\ell}.\check{\sigma}$ as a proxy for phonological transcriptions which do not exist for our all languages in UniMorph.

Lexeme-Based Cross Validation. In the studies that follow, we train a separate instance of our model on the forms in each language using the following procedure. We first remove morphologically-complex forms that are derived from other lemmas in the corpus using the heuristic technique described in §4.3. We then randomly assign the remaining lexemes of each language to one of ten splits. Note that each split will contain all of the forms associated with each lexeme and a lexeme will never be divided across splits. We then perform 10-fold cross-validation, training the model $p_{\boldsymbol{\theta}}$ on 8 splits, tuning on one of the remaining two splits, and testing on the final remaining split. Note that this approach to cross-validation allows us to approximate $\mathcal{L}_{-\ell}$ without the costly procedure of retraining for every held-out lexeme. However, also note that this approach has a potential confound. Lexemes can often be grouped into *inflectional classes* in which all lexemes mark different slots in the same way. For example, verbs such as *sing*/*sang*/*sung* and *ring*/*rang*/*rung* form an inflectional class in English. Inflectional classes vary in their size and regularity (Stump, 2001). If all or most lexemes in the same irregular inflectional class end up together in the test split under our approach, we may systematically overestimate their irregularity.

5.2 Validation and Accuracy

5.2.1 Validation on English Verbs

The first question we wish to ask is whether the irregularity predictions made by our model are consistent with human intuitions. To answer this question, we examine the predictions of our model on the English past tense—a morphological system which has been intensely studied for decades (see Pinker, 1999, for overview) and for which there is general agreement about which forms are regular

[4] Wikipedia data retrieved on Feb 1st, 2019.

Albright and Hayes (2003)	O'Donnell (2015)
0.670	0.559

Table 2: Validation of our irregularity metric. Spearman's ρ between gold-standard irregularity annotations from Albright and Hayes (2003) and O'Donnell (2015) and our irregularity metric.

or irregular. We make use of the databases of Albright and Hayes (2003) which consists of 4039 English verb forms and the dataset of O'Donnell (2015) which consists of 15202 verb forms, both hand-annotated for irregularity by experts.

We present our results in Table 2. We find that our measure of irregularity strongly correlates with human intuitions on English verbs. We take this as tentative validation of our metric. Future work will investigate the linguistic plausibility of our metric on a greater diversity of languages.

5.2.2 *Wug*-Test Accuracy

Language	Family	Avg. Accuracy	Lexemes	Forms	Avg. Forms/Lexeme
Albanian	Indo-European	0.83	537	26993	50.4
Arabic	Semitic	0.63	3559	89879	25.5
Armenian	Indo-European	0.95	4614	144841	31.4
Basque	**Isolate**	**0.01**	**26**	**10382**	**441.9**
Bulgarian	Slavic	0.94	2042	36007	17.7
Czech	Slavic	0.92	4470	61251	13.8
Danish	Germanic	0.65	2580	19968	7.8
Dutch	Germanic	0.94	3932	20680	5.3
English	Germanic	0.95	9915	40210	4.1
Estonian	Uralic	0.79	817	31711	38.9
French	Romance	0.86	5378	195638	37.4
German	Germanic	0.92	14739	69190	4.7
Hebrew	Semitic	0.78	492	11240	23.3
Hindi	Indo-Aryan	0.74	254	26404	104.0
Irish	Celtic	0.85	6527	69551	10.7
Italian	Romance	0.99	6495	269908	41.9
Latvian	Baltic	0.97	5347	60146	11.9
Persian	Iranian	0.70	271	26336	98.3
Polish	Slavic	0.93	8317	106914	13.0
Portuguese	Romance	0.98	2621	138372	52.9
Romanian	Romance	0.78	3409	51670	15.3
Russian	Slavic	0.95	19991	243748	12.2
Spanish	Romance	0.97	3904	232676	59.9
Swedish	Germanic	0.89	6451	43118	6.7
Turkish	Turkic	0.85	2697	150477	55.9
Ukrainian	Slavic	0.86	1426	13844	9.8
Urdu	**Indo-Aryan**	**0.38**	**180**	**5581**	**31.0**
Welsh	**Celtic**	**0.41**	**179**	**9083**	**50.8**

Table 3: Accuracy per language.

Our lexeme-based cross-validation setup differs substantially from the form-based setup typically used to evaluate models of inflectional morphology (see, e.g., Cotterell et al., 2017a). In the typical evaluation setup, individual surface word forms are heldout, rather than all of the forms associated with entire lexemes. This means, amongst other things, that words from irregular lexemes will often be split between test and train, giving models an opportunity to learn partially productive and

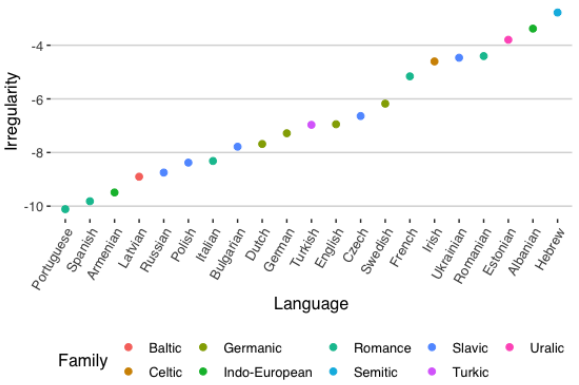

Figure 2: Average degree of irregularity ι across languages.

semi-regular patterns of inflection. Our approach however makes this impossible by strictly assigning all forms from each lexeme to either train or test.

It is important to ask, therefore, how well does our model predict the forms of heldout lexemes given this stricture? The results are displayed in Table 3. This table displays the average accuracy for each language in our sample as well as the number of lexemes for that language, the total number of forms, and the average number of forms per lexeme. The majority of languages show very high generalization accuracy to our lexeme-based *wug*-tests: 21 out of 28 have an average accuracy of 75% or higher. Three languages stand out in terms of their low accuracy and are highlighted in Table 3: Basque, Urdu, and Welsh. These languages, Basque especially, are characterized by smaller numbers of lexemes and larger numbers of forms per lexeme.

In the §5.4, we discuss the correlation between irregularity and frequency. The interpretation of these results relies on the ability of our model to accurately capture regular structure in the inflectional systems of the languages that we study. For this reason, we make the conservative choice to exclude all languages whose average accuracy was below 75% from all further analyses below.

5.3 Irregularity across Languages

It is often observed that there are differences in the prevalence of irregularity across languages (Stolz et al., 2012). On one end of the spectrum, some languages have widespread (often suppletive) allomorphy in their marking of inflectional features. For example, Arabic marks plurality on nouns in

one of more than a dozen different ways and these are idiosyncratic to the noun stem. Similarly, Georgian verbs often have different roots depending on their tense, aspect, or mood marking. On the other end of the spectrum, it is sometimes claimed that agglutinative languages like Turkish exhibit no irregularity whatsoever.

Figure 2 displays the average irregularity score per language for the 21 languages remaining after our 75% accuracy criterion. Recall from eq. (2) that the degree of irregularity ι is positive when the majority of predicted probability mass falls on forms that are not the correct target form (i.e., the form is irregular), and negative when the majority of probability mass falls on the predicted form (i.e., the form is regular). As can be seen from the figure, average irregularity is negative across languages. This is expected—most forms in these languages are predicted accurately by the model. However, there is wide variability in the average irregularity score between languages. In particular, in the most regular language, Portuguese, correct forms are about 25,000 times more likely on average than alternative forms. In the most irregular language, Hebrew, correct forms are only about 16 times more likely on average than alternative forms. We leave it to future work to validate and further study these cross-linguistic differences in irregularity predictions.

5.4 Irregularity and Frequency

In some morphological systems, such as the English past tense, there is a strong and well-known correlation between irregularity and frequency is well-known (Marcus et al., 1992; Pinker, 1999). In such systems, the most frequent past forms tend to be irregular and irregular forms tend to come from the most frequent verbs. Based on cases like this, it is widely believed in linguistics and psycholinguistics that there is an association between frequency and irregularity (Bybee, 1991; Haspelmath and Sims, 2010; Kiefer, 2000). However, to our knowledge, this relationship has never been explicitly tested quantitatively across many languages at once.

Recently, several authors have questioned the received wisdom that irregularity and frequency are related (Yang, 2016; Fratini et al., 2014).[5] Thus, it has become important to test this relationship empirically. An example of such a challenge to

Figure 3: Correlations between irregularity and frequency at the form level.

the standard assumption comes from Yang (2016) who proposed an influential theory of morphological productivity known as the *tolerance principle*. The mathematical derivation of the tolerance principle relies on the assumption that irregular forms are uniformly distributed throughout the frequency range (Yang, 2016).[6]

Here we present the first study to probe the relationship between irregularity and frequency at scale. We first examine the relationship between the degree of irregularity ι and the frequency of individual word forms. To study this question, we examined the Pearson correlation between the log-transformed frequency of word forms in each language and their predicted irregularity scores $\iota(w)$. Because word occurrences fall into the class of *large number of rare event* distributions, finite samples will tend to underestimate the probability of infrequent words—word forms that appear 0 times in some sample often differ by orders of magnitude in their true probability (Chitashvili and Baayen, 1993; Baayen, 2001). For this reason, we chose to exclude all frequency 0 forms from our analyses.

The correlations for the 21 languages considered in this study are shown in Figure 3 with significant correlations ($p < 0.05$) marked in blue. Overall, a slight trend towards a positive correlation between irregularity and frequency is discernible in this set of word forms. Following Mahowald et al. (2018), we tested this by fitting a mixed-effect model with irregularity as the dependent variable, language as a random effect (slopes and intercepts) and log count as a fixed effect (Gelman and Hill, 2007). The

[5] But see Herce (2016).

[6] Yang tentatively proposes that the correlation between frequency and irregularity might be accidental in languages such as English. He argues, however, that his theory is not contingent on this being the case (Yang, 2016, pp. 65).

Figure 4: Correlations between irregularity and frequency at the lexeme level.

results give a positive coefficient of 0.064 for the log count factor. The AIC-corrected log-odds ratio in favor of the model with a fixed effect of count (compared to a model with just random effects) is 3.44. A nested-model likelihood-ratio χ-squared test shows that the log factor is significant with $p < 0.04$.

An important question about irregularity is whether it is a property of individual forms, or rather whether it inheres to whole paradigms (Baerman et al., 2010; Stolz et al., 2012; Herce, 2016). To examine this question more closely, we ran an alternative correlational analysis examining the correlation between the sum of the counts of all forms associated with a lexeme and the average irregularity score for all forms associated with the lexeme (as in eq. (5)). Figure 4 shows the results. Overall, a stronger trend towards a positive correlation between irregularity and frequency is discernible at the lexeme level than at the word-form level. We tested this by fitting a mixed-effect model with irregularity as the dependent variable, language as a random effect (slopes and intercepts) and log count as a fixed effect. The models gives a positive coefficient of 0.14 for the log count factor. The AIC-corrected log-odds ratio in favor of the model with a fixed effect of count (compared to a model with just random effects) is 11.8. A nested-model likelihood-ratio χ-squared test shows that the log count factor is significant with $p < 0.001$. Thus, the correlation between irregularity and frequency is considerably more robust when considered at the lexeme level.

6 Conclusion

In this paper, we have introduced a measure of irregularity based on *wug*-testing a model of morphological inflection. In §5.2.1, we showed that this measure produces results that are consistent with human judgements. Focusing on a subset of the languages for which the model was able to recover the correct inflected forms at a high rate (§5.2.2), we showed that average irregularity varies a good deal between languages. This result is consistent with the findings of Cotterell et al. (2018a) which gave large scale empirical evidence of a tradeoff between the size of morphological paradigms and the predictability of individual forms within each paradigm.

The main novel empirical result of our paper was presented in §5.4 which showed that irregularity is correlated with frequency both at the level of individual forms as well as at the level of lexemes. To our knowledge, this is the first large-scale empirical demonstration of this piece of linguistic folk wisdom and provides evidence relevant to recent proposals questioning this generalization (Fratini et al., 2014; Yang, 2016).

Perhaps of greater interest than this positive result is the difference in the strength of the correlation between the level of individual forms and the level of lexemes. This difference appears to be driven by the fact that, in many cases, lexemes that contain high-frequency forms will also contain a few low frequency forms as well. Adopting the terminology of Yang (2002), we can say that low frequency forms *free-ride* on the higher frequency members of the lexeme.

This finding lends credence to models of linguistic structure which group words together by their lexeme or stem. Such models seem necessary to account for paradigmatic structure cross linguistically and to deal with phenomena such as the existence of *defective paradigms*—the phenomenon whereby certain inflected forms of a word seem to be impossible for speakers (Baerman et al., 2010). A canonical example is the past participle of *stride* (e.g., **strode*/**stridden*/**strided*). In these cases, the problem seems to be that the irregularity of the overall lexeme is known, but the particular word form has never been observed. Our results provide further support for the view that inflected forms represent surface exponence of common underlying morphological objects.

More generally, we observe that our *wug*-test

techniques provides a general way of studying regularity and predictability within languages and may prove useful for attacking other difficult problems in the literature, such as detecting inflectional classes. By measuring which words or lexemes are most predictable from one another, a general picture of morphological relatedness within a language can be built in a bottom-up way.

Acknowledgments

The third author gratefully acknowledges support from the Fonds de Recherche du Québec—Société et Culture and the Natural Sciences and Engineering Research Council of Canada.

References

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

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

R. Harald Baayen. 2001. *Word Frequency Distributions*. Springer, Berlin, Germany.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2015. *Understanding and measuring morphological complexity: An introduction.* Oxford University Press.

Matthew Baerman, Greville G. Corbett, and D. P. Brown. 2010. *Defective Paradigms: Missing forms and what they tell us*. Oxford University Press, Oxford, England.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14:150–177.

Joan L. Bybee. 1985. *Morphology: A Study of the Relation between Meaning and Form*. John Benjamins, Amsterdam.

Joan L. Bybee. 1991. Natural morphology: The organization of paradigms and language acquisition. In Thom Huebner and Charles A. Ferguson, editors, *Cross Currents in Second Language Acquisition and Linguistic Theory*. John Benjamins Publishing Company.

Revas J. Chitashvili and R. Harald Baayen. 1993. Word frequency distributions. *Quantitative Text Analysis*, pages 54–135.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2018a. On the complexity and typology of inflectional morphological systems. *Transaction of the Association for Computational Linguistics (TACL)*.

Ryan Cotterell, Christo Kirov, Sebastian J. Mielke, and Jason Eisner. 2018b. Unsupervised disambiguation of syncretism in inflected lexicons. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 548–553. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017a. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30. Association for Computational Linguistics.

Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017b. Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL2017)*.

Viviana Fratini, Joana Acha, and Itziar Laka. 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.

Andrew Gelman and Jennifer Hill. 2007. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*. Hodder Education.

Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge, New York, NY.

Borja Herce. 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).

Charles F. Hockett. 1954. Two models of grammatical description. *Word*, 10:210–231.

Ferenc Kiefer. 2000. Regularity. In *Morphologie: Ein internationales Handbuch zur Flexion und Wortbildung/Morphology: An international Handbook on Inflection and Word-Formation*. Walter de Gruyter, Berlin.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. *arXiv preprint arXiv:1810.11101*.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven Thomas Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Gary F. Marcus, Steven Pinker, Michael T. Ullman, Michelle Hollander, T. John Rosen, and Fei Xu. 1992. *Overregularization in Language Acquisition*. Monographs of the society for research in child development. University of Chicago Press, Chicago, IL.

James L. McClelland and Karalyn Patterson. 2002a. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6(11):465–472.

James L. McClelland and Karalyn Patterson. 2002b. 'Words or Rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences*, 6(11):464–465.

Timothy J. O'Donnell. 2015. *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. The MIT Press, Cambridge, Massachusetts.

Steven Pinker. 1999. *Words and Rules*. HarperCollins, New York, NY.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Steven Pinker and Michael T. Ullman. 2002a. Combination and structure, not gradedness, is the issue. *Trends in Cognitive Sciences*, 6(11):472–474.

Steven Pinker and Michael T. Ullman. 2002b. The past and future of the past tense debate. *Trends in Cognitive Sciences*, 6(11):456–463.

Sandeep Prasada and Steven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8(1):1–56.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.*, volume 2, pages 216–271. Bradford Books/MIT Press, Cambridge, MA.

Dorothy Siegel. 1974. *Topics in English Morphology*. Ph.D. thesis, Massachusetts Institute of Technology.

Thomas Stolz, Hitomi Otsuka, Aina Urdze, and Johan van der Auwera. 2012. Introduction: Irregularity — glimpses of a ubiquitous phenomenon. In Thomas Stolz, Hitomi Otsuka, Aina Urdze, and Johan van der Auwera, editors, *Irregularity in Morphology (and Beyond)*, pages 7–38. Akademie Verlag, Berlin, Germany.

Gregory T. Stump. 2001. Inflection. In *Handbook of Morphology*. Blackwell, Oxford, England.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. *arXiv preprint arXiv:1905.06319*.

Charles D. Yang. 2002. *Knowledge and Learning in Natural Language*. Oxford linguistics. Oxford University Press, New York.

Charles D. Yang. 2016. *The Price of Productivity: How Children Learn to Break the Rules of Language*. The MIT Press, Cambridge, Massachusetts.