

# Correlation Does Not Imply Compensation: Complexity and Irregularity in the Lexicon

Amanda Doucette<sup>1</sup> Ryan Cotterell<sup>2</sup> Morgan Sonderegger<sup>1</sup> Timothy J. O'Donnell<sup>1,3</sup>

<sup>1</sup>Dept. of Linguistics, McGill University <sup>2</sup>Dept. of Computer Science, ETH Zurich

<sup>3</sup>Canada CIFAR AI Chair, Mila

amanda.doucette@mail.mcgill.ca ryan.cotterell@inf.ethz.ch

morgan.sonderegger@mcgill.ca timothy.odonnell@mcgill.ca

## Abstract

It has been claimed that within a language, morphologically irregular words are more likely to be phonotactically simple and morphologically regular words are more likely to be phonotactically complex. This inverse correlation has been demonstrated in English for a small sample of words, but has yet to be shown for a larger sample of languages. Furthermore, frequency and word length are known to influence both phonotactic complexity and morphological irregularity, and they may be confounding factors in this relationship. Therefore, we examine the relationships between all pairs of these four variables both to assess the robustness of previous findings using improved methodology and as a step towards understanding the underlying causal relationship. Using information-theoretic measures of phonotactic complexity and morphological irregularity (Pimentel et al., 2020; Wu et al., 2019) on 25 languages from UniMorph, we find that there is evidence of a *positive* relationship between morphological irregularity and phonotactic complexity within languages on average, although the direction varies within individual languages. We also find weak evidence of a negative relationship between word length and morphological irregularity that had not been previously identified, and that some existing findings about the relationships between these four variables are not as robust as previously thought.<sup>1</sup>

## 1 Introduction

The *compensation hypothesis* (Martinet, 1955; Hockett, 1955) states that as a language increases in complexity in one area, another must decrease in complexity to compensate. A compensatory relationship could exist either *within* a language (i.e., words that are more complex in one way are less complex in another), or *across* languages (i.e., an entire lexicon that is more complex in one way

is less complex in another). One such compensatory relationship has been proposed between morphological irregularity and phonotactic complexity. Hay (2003), Hay and Baayen (2003), and Burzio (2002) argue that words within a language with irregular morphology tend to be phonotactically simple, while words with regular morphology tend to be phonotactically complex.

Although there is some evidence for this relationship in English (reviewed below), the existence of a correlation does not imply *compensation*, which we take to mean that an increase in one variable directly causes a decrease in the other (Pearl et al., 2016, §1.5). While we may observe a correlation between morphological irregularity and phonotactic complexity, it is possible that there is in fact no direct causal relationship between them. For example, they could share a common cause such as word frequency (Pearl et al., 2016, §2.2). The effect could also be mediated through a third mediator variable (Pearl et al., 2016, §3.7), as has been argued for the relationship between phonotactic complexity and frequency (Mahowald et al., 2018).

Therefore, to assess the relationship between morphological irregularity and phonotactic complexity, we need to examine any other variables they may be related to. Previous work suggests that both morphological irregularity and phonotactic complexity are correlated with word frequency. Wu et al. (2019) showed that morphological irregularity positively correlates with frequency, and Mahowald et al. (2018) showed that phonotactic complexity inversely correlates with frequency after controlling for word length. Phonotactic complexity is also known to be correlated with word length, with longer words conveying less information per phoneme (Pimentel et al., 2020), and more frequent words tend to be shorter (Zipf, 1935; Piantadosi et al., 2011; Piantadosi, 2014; Pimentel et al., 2023).

While there is evidence supporting a relation-

<sup>1</sup>Code is available at <https://osf.io/ax78p/>.

ship between some pairs of these four variables—morphological irregularity, phonotactic complexity, word length, and frequency—there is reason to be uncertain about the existence and direction of a correlation between others, whether the correlation holds within or across languages, and what other variables need to be controlled for to accurately assess the effect. The pairwise relationships between these variables have not yet been examined on a single data set of many languages, and some relationships have only been examined using orthographic rather than phonetic transcriptions. Therefore, in addition to examining the relationship between phonotactic complexity and morphological irregularity, we will also examine the relationships between all other pairs of variables in this set.

We find that within languages, there is a positive effect of phonotactic complexity on morphological irregularity after controlling for word length and frequency. Across languages, we find no consistent effect. We replicate previous findings of a negative effect of word length on frequency, and of a positive effect of frequency on morphological irregularity. We also find a negative effect of frequency on phonotactic complexity, although not as robust as previously suggested. Our results for the relationship between phonotactic complexity and word length complicate previous results: We find a positive effect for one data set and a negative effect for another. Finally, we present a novel analysis of the effect of word length on morphological irregularity, and find a negative effect in most languages.

## 2 Background

For each pair of variables, we summarize previous work demonstrating a correlation, any theoretical arguments supporting a positive, negative, or no relationship between the two, and what other variables must be controlled for to examine a potential causal relationship.

### 2.1 Phonotactic Complexity vs. Morphological Irregularity

It has been hypothesized that within a language, phonotactic complexity is negatively correlated with morphological irregularity. Hay and Baayen (2003), Hay (2003), and Burzio (2002) argued that for English, words that are phonotactically complex are more likely to be morphologically regular. For example, *dreamed* is morphologically regular, but contains the unusual consonant cluster [md], while

*went* is morphologically irregular, but has regular phonotactics. While this relationship has not been examined in other languages, there are several reasons to suspect a negative correlation as a universal tendency. First, low-probability phonotactic junctures can facilitate morphological decomposition, as argued by Hay (2003), who found that for a set of 12 English affixes the proportion of words creating an illegal phonotactic juncture was predictive of morphological productivity. Second, as argued by Burzio (2002), irregular forms are more likely to be memorized, while regular forms are constructed from individual morphemes. If phonotactically simple words are easier to store in memory, phonotactic complexity should be inversely correlated with morphological irregularity.

There are also reasons to suspect no relationship between morphological irregularity and phonotactic complexity, related to the limitation that all previous work considers only a small set of words. Morphological and phonotactic processes could apply independently, and previously observed significant correlations could be statistical accidents due to small sample size. Indeed, responding to Hay (2003), Plag (2002) found no correlation between morphological irregularity and phonotactic complexity in a different sample of 12 English affixes. Alternatively, morphological irregularity and phonotactic complexity could be independent conditional on a third common cause or mediator variable, with which they are both correlated. This could result in a statistically significant correlation, while there is no causal relationship in reality.

One such common cause that could result in a positive observed correlation between phonotactic complexity and morphological regularity is word age. Hay and Baayen (2003) note that highly productive affixes are regularly used in creating new words. Thus, new words in a language will tend to have regular morphology, and it seems plausible they will also tend to have regular phonotactics. As the language changes over time, what is considered regular will also change, resulting in a positive correlation: Older words will have irregular morphology and high phonotactic complexity, while newer words will have regular morphology and low phonotactic complexity. Other possible common causes include word length and frequency: Previous work demonstrates correlations between frequency and phonotactic complexity, frequency and morphological irregularity, and word length and phonotactic complexity. A negative effect of word

length on morphological irregularity is also plausible, and will be established in our data. Both word frequency and word length are therefore common causes that should be controlled for in assessing the relationship between phonotactic complexity and morphological irregularity.

## 2.2 Phonotactic Complexity vs. Length

Pimentel et al. (2020) demonstrated a strong negative correlation between phonotactic complexity and average word length both across and within 106 languages. Pellegrino et al. (2011) suggest that this compensation is the result of a linguistic universal: The rate of information in every language is very similar, with the amount of information per word roughly constant. Thus, longer words should have less information per phoneme (Coupé et al., 2019; Meister et al., 2021). While previous work (discussed below) suggests that frequency is a common cause of both phonotactic complexity and word length, and should be controlled for in this analysis, we will not control for it in line with our goal of replicating previous studies using a single dataset.<sup>2</sup>

## 2.3 Morphological Irregularity vs. Frequency

A positive correlation between morphological irregularity and word frequency has been observed in English (Marcus et al., 1992; Bybee, 1985), but this correlation was questioned by Fratini et al. (2014) and Yang (2016). In a larger set of 21 languages, Wu et al. (2019) found a positive correlation between morphological irregularity and frequency. These correlations were found to be more robust when irregularity was considered as a property of lemmas rather than individual words. A potential mechanism is described by Hay and Baayen (2003): More frequent words are more likely to be accessed as whole words, and less frequent words are more likely to be parsed into their component morphemes. Because irregulars are more likely to be accessed as whole words in memory, there will be a positive correlation between frequency and morphological irregularity. In contrast, if we assume that lexicons are optimized for efficient communication, i.e., more frequent words should be less morphologically complex (Zipf, 1935), we would expect frequent words to have regular morphology, i.e., a negative correlation.

<sup>2</sup>However, preliminary models show that controlling for frequency only has a minimal impact on results.

## 2.4 Phonotactic Complexity vs. Frequency

A consequence of Zipf's (1935) hypothesis that the most frequent words in a language should require the least effort is that even within words of the same length, the most frequent ones should be easiest to produce and understand. This suggests that more frequent words should have lower phonotactic complexity. After controlling for word length, this is exactly what Mahowald et al. (2018) found in a study of 96 languages, using orthographic probabilities from Wikipedia as a proxy for phonotactic complexity. However, orthography can differ significantly from pronunciation—this correlation has not been confirmed with phonotactic probabilities from phonetic transcriptions. Following Mahowald et al. (2018), we will control for word length as a potential mediator in the relationship between phonotactic complexity and word frequency.

## 2.5 Morphological Irregularity vs. Length

We are not aware of previous work on the relationship between morphological irregularity and word length, although it is intuitively plausible that one influences the other. For example, a negative correlation within a language could arise because regular inflectional morphology involves combining multiple morphemes, causing words with regular morphology to be longer. Previous work also implies that frequency has an effect on both morphological irregularity and word length. Therefore, we control for frequency as a common cause in assessing the potential relationship between morphological irregularity and word length.

## 2.6 Length vs. Frequency

Zipf (1935) observed that the most frequent words in a language tend to be short. Since then, the inverse relationship between word length and frequency has been studied in depth and found to follow Zipf's law extremely systematically (see Piantadosi, 2014 for a review), although it is unclear whether word length correlates more strongly with surprisal (Piantadosi et al., 2011) or frequency (Meylan and Griffiths, 2021).

# 3 Methods

## 3.1 Data

Our morphological data comes from the UniMorph project, a database of morphologically annotated corpora for 182 languages (Batsuren et al., 2022). Each inflected form is annotated with its lemma

(the lexical meaning) and a set of morphological features; *walked* would be annotated as [VERB; SINGULAR; PAST], for example. While UniMorph provides data for a large set of languages in a universal schema, it does not provide phonetic transcriptions. Because we are interested in how morphology interacts with phonotactics, we use grapheme to phoneme models from Epitran (Mortensen et al., 2018) to convert orthographic transcriptions to IPA. Languages with no available Epitran model were excluded from our analyses.

For training models of phonotactic complexity, we use NorthEuraLex (Dellert et al., 2020), a database of phonetic transcriptions of 1,016 basic concepts for 107 Northern Eurasian languages also studied by Pimentel et al. (2020). For languages that are not included in NorthEuraLex, we use WikiPron (Lee et al., 2020), a database of pronunciation dictionaries from Wiktionary. Languages not in either NorthEuraLex or WikiPron are excluded.

Frequency data is retrieved from Wikipedia,<sup>3</sup> and calculated as log count per million words. Following Wu et al. (2019), we exclude all forms with zero frequency, which can differ by orders of magnitude in their true frequency (Baayen, 2001), and we exclude 15 languages where the average probability of the morphological irregularity model predicting the correct surface form is below 0.75. The UniMorph dataset for each language varies in size from 77 to 50,284,287 forms and 37 to 824,074 lemmas. Many of the excluded languages are those with smaller datasets, where the model does not have enough information to accurately predict surface forms.

The languages included in our analysis are: Albanian, Amharic, Azerbaijani, Catalan, Chewa, Czech, Dutch, English, French, German, Hungarian, Italian, Kazakh, Khalka Mongolian, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Spanish, Swedish, Turkish, Ukrainian, Uzbek, and Zulu. Further details about datasets used can be found in App. A.

### 3.2 Morphological Irregularity Models

While a binary distinction between regular and irregular morphology is useful in many theories of grammar, a more fine-grained quantitative measure of irregularity is needed to examine the potential relationships with other variables we are interested in. For example, an English speaker might judge

a verb like *walked* to be more regular than *sang*, which is, in turn, more regular than *went*. Wu et al. (2019) define an information theoretic measure of irregularity that captures such intuitions, and is applicable across languages.

This measure, called *degree of morphological irregularity*, abbreviated as MI, is defined using a probabilistic model. Let  $\Sigma$  be an alphabet of symbols<sup>4</sup> and let  $\mathcal{S}$  be a finite set of morphological features, e.g., those provided by the UniMorph dataset. Furthermore, we define the inflector function  $\iota: \Sigma^* \times \mathcal{S} \rightarrow \Sigma^*$  that maps a pair of a lemma and a set of morphological features to an inflected surface form, i.e.,  $(\ell, \sigma) \mapsto w$ . The inflector  $\iota$  is assumed to only operate on *known* lemmas. Thus, to get at a notion of morphological irregularity, we also require a probabilistic inflection model  $p(w | \ell, \sigma, \mathcal{L}_{-\ell})$ , a probability distribution over  $\Sigma^*$  conditioned on a lemma  $\ell \in \Sigma^*$ , a slot  $\sigma \in \mathcal{S}$ , and a lexicon with the target lemma removed  $\mathcal{L}_{-\ell}$ , that tells us which forms in  $\Sigma^*$  are probable inflected surface forms for the lemma  $\ell$  with morphological features  $\sigma$ . The distribution  $p(w | \ell, \sigma, \mathcal{L}_{-\ell})$  essentially corresponds to a wug-test probability (Berko, 1958), i.e., it tells us the likelihood of the model predicting the correct inflected form of a word it has never seen. To make the probabilities given by  $p(w | \ell, \sigma, \mathcal{L}_{-\ell})$  more interpretable, Wu et al. (2019) use the negative log odds of probability of the correct surface form, i.e.,

$$\text{MI}(w, \ell, \sigma) = -\log \frac{p(w | \ell, \sigma, \mathcal{L}_{-\ell})}{1 - p(w | \ell, \sigma, \mathcal{L}_{-\ell})} \quad (1)$$

We can interpret  $\text{MI}(w, \ell, \sigma)$  as follows. We achieve a  $\text{MI}(w, \ell, \sigma)$  of 0 if the probability of the correct surface form is exactly 0.5, a negative  $\text{MI}(w, \ell, \sigma)$  when a surface form is more predictable, and a positive  $\text{MI}(w, \ell, \sigma)$  when the form is less predictable.

Morphological irregularity can be considered either a property of an individual word, as in Eq. (1), or as a property of an entire lemma. We calculate the MI of a lemma as the mean MI score of all words in the lemma, i.e.,

$$\text{MI}(\ell) = \frac{1}{|\mathcal{S}|} \sum_{\sigma \in \mathcal{S}} \text{MI}(\iota(\ell, \sigma), \ell, \sigma) \quad (2)$$

**Estimating  $p(w | \ell, \sigma, \mathcal{L}_{-\ell})$  from data.** Because the probability distribution  $p(w | \ell, \sigma, \mathcal{L}_{-\ell})$  is con-

<sup>3</sup>Retrieved 4/16/23 from <https://dumps.wikimedia.org>.

<sup>4</sup>The symbols could be graphemes or phonemes, depending on the nature of the data annotation.

ditioned on a lexicon with the target lemma removed, the most accurate estimate of MI would require training a separate model for each target lemma in each language. In practice, MI is estimated by training models on a language with a set of lemmas removed, rather than just one. We train neural network models on the UniMorph data using code from Wu et al. (2019), which implements a monotonic hard attention string-to-string induction model described by Wu and Cotterell (2019). In our experiments, we use the same model architecture and training parameters as Wu et al. (2019), and split the lemmas for each language into thirty sets.

### 3.3 Phonotactic Complexity Models

Similar to our estimation of morphological irregularity, to estimate phonotactic complexity, we take a probabilistic approach. Following Pimentel et al. (2020), we consider a probability distribution  $p(w | \mathcal{L})$  over  $\Sigma^*$ . Then, given a word  $w \in \Sigma^*$ , we define the degree of phonotactic complexity, abbreviated as PC, as follows

$$\text{PC}(w) = -\frac{\log p(w | \mathcal{L})}{|w|} \quad (3)$$

where  $|w|$  is the length of the word  $w$ . Like the degree of morphological irregularity, PC is a surprisal-based metric that lends itself to easy interpretation. Specifically, if  $\text{PC}(w)$  is lower, it means that  $w$  is less surprising and therefore more regular.

**Estimating  $p(w | \mathcal{L})$  from data.** The distribution  $p(w | \mathcal{L})$  is a hypothetical construct that tells us the probability of an *unknown* word. When we estimate  $p(w | \mathcal{L})$  from data, we cannot use the estimated distribution to judge the complexity of those words in the training data. We split training data for each language into ten sets and train ten models, each with one set held out. PC is evaluated on the held-out set for each model. We use the model architecture and training procedure used in code provided by Pimentel et al. (2020), which implements a character-level LSTM (Hochreiter and Schmidhuber, 1997) language model with each phoneme represented by a set of phonetic features from Phoible (Moran et al., 2014).

### 3.4 Analysis

Following previous studies on the interactions between phonotactic complexity, morphological irregularity, word length, and frequency, we report regression coefficients for each language and each

Y	X	Controls	Rand. Effs.
MI	PC	<u>FR</u> + mean(PC) + <u>WL</u> + mean(WL)	PC + FR + WL
PC	WL	mean(WL)	WL
MI	FR	–	FR
PC	FR	<u>WL</u> + mean(WL)	FR + WL
MI	WL	<u>FR</u> + mean(WL)	WL + FR
WL	FR	–	FR

Table 1: Controls and random effects included in regression models of properties  $X$  and  $Y$  of words (one row per word): MI = morphological irregularity; PC = phonotactic complexity; WL = word length; FR = frequency. Mean( $X$ ) is a language’s average value of  $X$ , across all its words. In lme4 syntax, models using words from all languages are:  $Y \sim X + \text{Controls} + (1 + \text{Random Effects} | \text{language})$ , and individual language models are:  $Y \sim X + \text{UnderlinedControls}$ .

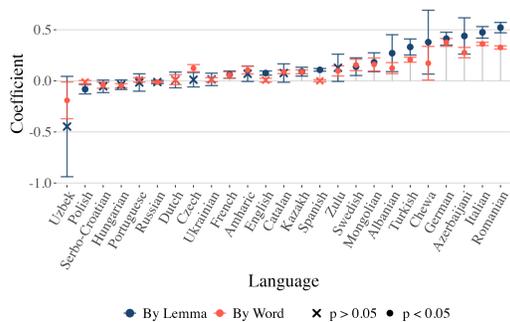
pair of variables, controlling for any necessary variables as described in Tab. 1—these are analogous to partial correlations between the variables of interest. The  $p$ -values have been adjusted for multiple comparisons using the Benjamini-Hochberg method. We also perform a linear mixed effects regression analysis across languages for each pair of variables, including random intercepts and slopes for the effect of language. The means of the dependent variable and controls within each language are also included as an additional predictor, to separate across-language effects from within-language effects (see Sonderegger, 2023, §8.10.2.3 and Antonakis et al., 2021). Because languages should not differ in mean word frequency, this predictor is excluded. All predictors were standardized, and log counts-per-million were used for frequencies.

For morphological irregularity analyses (where  $Y = \text{MI}$  in Tab. 1), we also report a regression model with data grouped by lemma, where phonotactic complexity and word length are taken as the average within each lemma, frequency is the sum of frequencies within each lemma, and morphological irregularity is calculated according to Eq. (2). Results for the effects of interest in these regression analyses are reported below. Plots of regression predictions and raw data for select languages are shown in App. B.

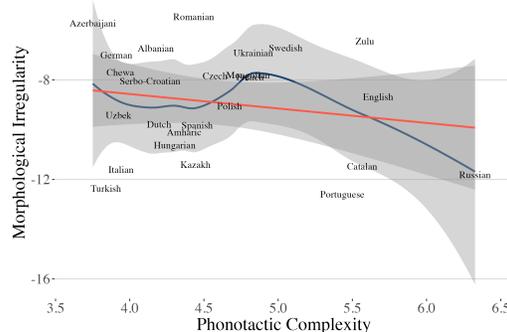
## 4 Results

### 4.1 Phonotactic Complexity and Morphological Irregularity

Within languages, we find a positive effect of phonotactic complexity on morphological irreg-

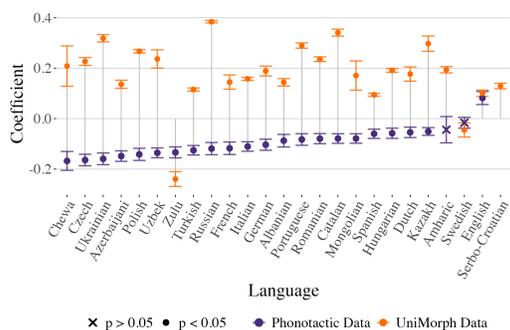


(a) Regression coefficients by language, grouped by lemma and by word, with 95% CIs.

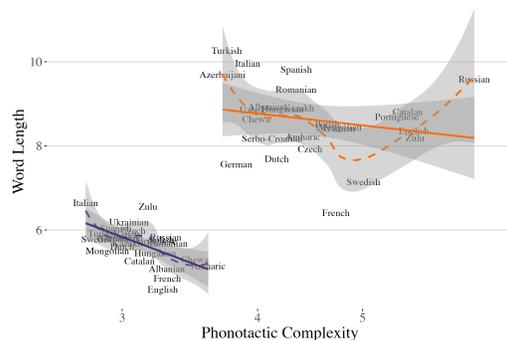


(b) By-language means, with linear (red/light) and LOESS (blue/dark) smoothers.

Figure 1: Phonotactic complexity and morphological irregularity.



(a) Regression coefficients by language, for phonotactic and UniMorph data, with 95% CIs.



(b) By-language means for phonotactic data (purple/dark) and UniMorph data (orange/light), with linear (solid) and LOESS (dotted) smoothers.

Figure 2: Phonotactic complexity and word length.

ularity after controlling for word length and frequency of 0.10 (95% CI [0.05, 0.15],  $p < 0.001$ ,  $\sigma = 0.188$ )<sup>5</sup> and a non-significant ( $p > 0.05$ ) effect of mean phonotactic complexity of -0.19 (95% CI [-0.50, 0.12],  $p = 0.224$ ). When grouped by lemma, there is an estimated effect of 0.14 (95% CI [0.07, 0.21],  $p < 0.001$ ,  $\sigma = 0.261$ ) and a non-significant effect of mean phonotactic complexity of 0.00 (95% CI [-0.36, 0.37],  $p = 0.993$ ). Although a majority of languages have a positive effect, as shown in Fig. 1a, some are negative or non-significant. Across languages, there is no evidence of a relationship between mean morphological irregularity and mean phonotactic complexity—we find a non-significant Spearman’s correlation of  $-0.045$  ( $p = 0.832$ ), shown in Fig. 1b.

#### 4.2 Phonotactic Complexity and Word Length

As shown in Fig. 2a, we find a positive relationship between phonotactic complexity and word length within the majority of languages in the UniMorph

data set. We find a similar prediction from the linear model: the estimated effect is 0.18 (95% CI [0.13, 0.23],  $p < 0.001$ ,  $\sigma = 0.123$ ). Across languages, we find no evidence of a correlation between mean phonotactic complexity and mean word length ( $\rho = -0.333$ ,  $p = 0.104$ ), as shown in Fig. 2b. Similarly, the linear model estimates no significant effect of mean word length ( $\beta = -0.40$ , 95% CI [-0.95, 0.14],  $p = 0.145$ ). The positive effects found in most languages are opposite to the direction predicted by Pimentel et al. (2020) and Pellegrino et al. (2011). However, it is important to note that nearly all words in UniMorph are morphologically complex, while NorthEuraLex (the dataset used by Pimentel et al. (2020)), contains mostly morphologically simple words. As previously noted, morpheme boundaries can create low-probability phonotactic junctures, resulting in higher phonotactic complexity. This suggests that the relationship between word length and phonotactic complexity may be dependent on morphological complexity. We also note that the UniMorph

<sup>5</sup>We use  $\sigma$  to refer to random effect standard deviation.

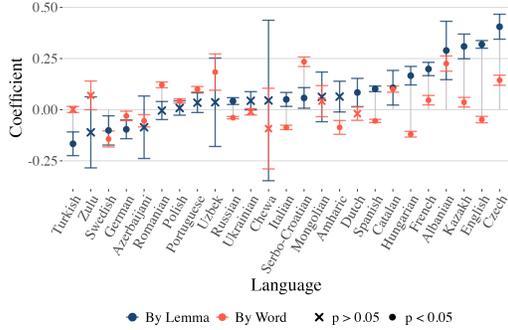


Figure 3: Morphological irregularity and frequency coefficients by language, grouped by lemma and by word, with 95% CIs.

data set contains significantly longer words than NorthEuraLex, suggesting a potential non-linear effect.

To test this, we evaluate the phonotactic complexity of the NorthEuraLex and WikiPron data used to train the phonotactic models and fit models for each language. These results are shown in Fig. 2a, where we can see that most languages show a negative effect in morphologically simple data, replicating the results of Pimentel et al. (2020). We find a Spearman’s correlation coefficient of  $-0.578$  ( $p = 0.003$ ) between average PC and average word length across languages, as shown in Fig. 2b. These results suggest that the finding that phonotactic complexity is negatively correlated with word length only holds for either morphologically simple or relatively short words.

### 4.3 Morphological Irregularity and Frequency

On the one hand, when morphological irregularity is considered a property of individual words, we find that 10/25 languages have a significant positive effect of frequency on morphological irregularity, and 9/25 have a significant negative effect (Fig. 3). In the linear mixed-effects model, we find no significant effect ( $\beta = 0.02$ , 95% CI  $[-0.02, 0.06]$ ,  $p = 0.224$ ,  $\sigma = 0.100$ ). On the other hand, when morphological irregularity is considered a property of lemmas rather than individual words, we find that 12/25 languages have a positive effect and only 3/25 have a negative effect. These correlations are shown in Fig. 3. The linear mixed-effects model predicts a positive effect of 0.08 (95% CI  $[0.02, 0.13]$ ,  $p = 0.005$ ,  $\sigma = 0.135$ ), consistent with those of Wu et al. (2019) who examined the same relationship, although using the original or-

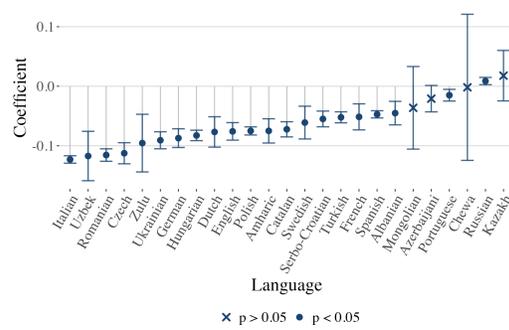


Figure 4: Phonotactic complexity and frequency regression coefficients by language, with 95% CIs.

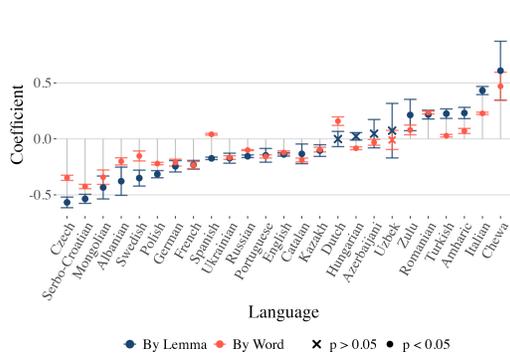
thographic transcriptions from UniMorph, rather than G2P transcriptions. Although the direction of correlation varies across individual languages, there is a tendency towards a positive correlation.

### 4.4 Phonotactic Complexity and Frequency

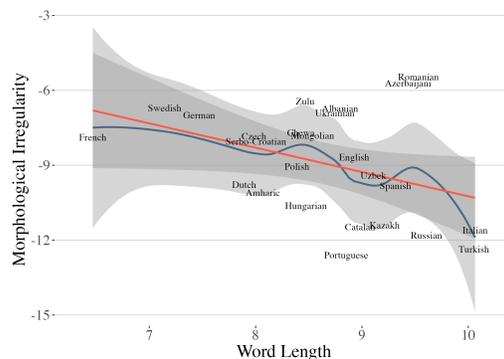
Within most languages, we find a significant negative effect of frequency on phonotactic complexity after controlling for word length, as shown in Fig. 4. The linear model predicts an effect of  $-0.06$  (95% CI  $[-0.08, -0.05]$ ,  $p < 0.001$ ,  $\sigma = 0.034$ ). These findings are similar to those of Mahowald et al. (2018), who found a negative or non-significant correlation in 100% of languages using orthographic data and a simpler measure of phonotactic complexity. Using phonetic transcriptions, we find significantly positive effects in only 1/25 languages, suggesting that this effect is fairly consistent across languages.

### 4.5 Morphological Irregularity and Length

Within most languages, we find a negative effect of word length on morphological irregularity after controlling for frequency, shown in Fig. 5a. The linear model also estimates a negative effect of  $-0.07$  (95% CI  $[-0.15, 0.00]$ ,  $p = 0.058$ ,  $\sigma = 0.193$ ), although it is non-significant. This effect is also somewhat consistent across languages, as shown in Fig. 5b. Across languages, we find a non-significant Spearman’s correlation of  $-0.38$  ( $p = 0.061$ ), while the linear model estimates the effect of mean word length to be  $-0.60$  (95% CI  $[-1.09, -0.12]$ ,  $p = 0.015$ ). When grouped by lemma, we find a non-significant effect of  $-0.08$  (95% CI  $[-0.19, 0.03]$ ,  $p = 0.131$ ,  $\sigma = 0.277$ ), and a non-significant effect of mean word length of  $-0.59$  (95% CI  $[-1.28, 0.11]$ ,  $p = 0.098$ ).



(a) Regression coefficients by language, grouped by lemma and by word, with 95% CIs.



(b) By-language means, with linear (red/light) and LOESS (blue/dark) smoothers.

Figure 5: Morphological irregularity and word length.

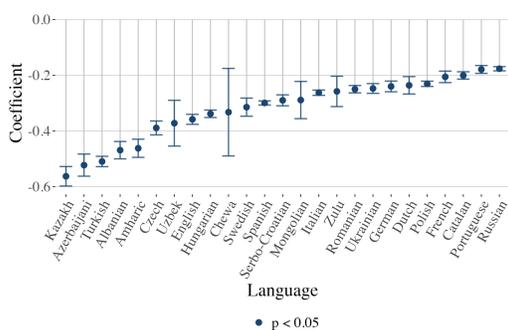


Figure 6: Word length and frequency regression coefficients by language, with 95% CIs.

#### 4.6 Word Length and Frequency

Finally, our results for word length and frequency are exactly as expected, following the prediction of Zipf (1935). We find consistently negative effects of frequency on word length, shown in Fig. 6, and an effect of  $-0.32$  (95% CI  $[-0.36, -0.27]$ ,  $p < 0.001$ ,  $\sigma = 0.109$ ) in the linear model.

### 5 Discussion

In examining the interactions between phonotactic complexity, morphological irregularity, word length, and frequency, we have found several unexpected and unintuitive results.

**Within-language Results.** Within most languages, we find significant effects for every pair of variables, although the direction of the effect varies. In our analysis of phonotactic complexity and frequency, we see a strong tendency towards negative effects. In our analysis of morphological irregularity and word length, we see a tendency towards a negative effect, but a positive effect in several languages. In analyses of morphological irregularity and frequency, and of phonotactic complexity

and morphological irregularity, we see a strong tendency towards positive effects. However, the only analysis that is without exception in all languages examined is word length and frequency (by far the best supported by previous work, i.e., Zipf’s Law), where there is a negative effect for all languages. For phonotactic complexity and word length, we found that the direction of the effect for each language changes with the data used. These results complicate the claims of previous work examining several of these pairs across languages, which generally conclude that there is strong support for the relationship. Our results for phonotactic complexity and morphological irregularity also contradict those of Hay and Baayen (2003), Hay (2003), and Burzio (2002). We find evidence of a positive effect of phonotactic complexity on morphological irregularity within language, rather than the negative effect that has previously been argued for. Our analysis controlled for word length and frequency, while previous work did not.

**Across-language Results.** Across languages, however, we find a negative effect of word length on morphological irregularity and a positive effect of word length on phonotactic complexity, although the direction of this effect changes with the data set used. We also find no evidence of a linear effect of phonotactic complexity on morphological irregularity across languages. However, as can be seen in Fig. 1b, there may be a nonlinear effect across languages. In exploratory data analysis, we also identified several possible nonlinear effects within languages. We leave fully describing any such relationships to future work.

**Incomplete Picture.** The results presented here suggest that, although these four variables do influence each other, we do not have enough infor-

mation to make claims about one compensating for another, either within a single language or universally across languages. The effects found are potentially consistent with several causal models. It is the causal effects captured by these models that we are interested in—how much does an intervention in one variable *directly* affect another, if at all? If a decrease in morphological irregularity causes a decrease in word length, which then causes an increase in phonotactic complexity, there is no compensatory relationship between morphological irregularity and phonotactic complexity, no matter how correlated they appear to be.

**Causal Modeling.** Our results assume that the underlying causal structure implied by previous work is correct—that morphological irregularity and phonotactic complexity have a common cause of frequency and word length, that the effect of frequency on phonotactic complexity is mediated by word length, and that morphological irregularity and word length have a common cause of frequency. However, it is possible that a different causal model underlies this data. We leave proposing such causal models and testing their implications for future work. While the models discussed in this work provide a starting point for understanding the structure of the lexicon, evidence supporting the underlying causal structure responsible for generating the data is necessary to evaluate any compensatory relationships in a set of highly correlated variables (Pearl et al., 2016).

### Acknowledgements

We thank the Montreal Computational & Quantitative Linguistics Lab for helpful feedback. The first author was supported by funding from the Fonds de recherche du Québec - Société et culture (FRQSC). The third author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (RGPIN-2023-04873). The senior author also gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada and the Canada CIFAR AI Chairs Program. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada.

### References

John Antonakis, Nicolas Bastardoz, and Mikko Rönkkö. 2021. *On ignoring the random effects assumption*

*in multilevel models: Review, critique, and recommendations. Organizational Research Methods, 24(2):443–483.*

R. Harald Baayen. 2001. *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer Dordrecht.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. *UniMorph 4.0: Universal Morphology*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855. European Language Resources Association.

Jean Berko. 1958. *The child's learning of English morphology. Word, 14(2-3):150–177.*

Luigi Burzio. 2002. *Missing players: Phonology and the past-tense debate. Lingua, 112(3):157–199.*

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

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. *Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. Science Advances, 5(9):eaaw2594.*

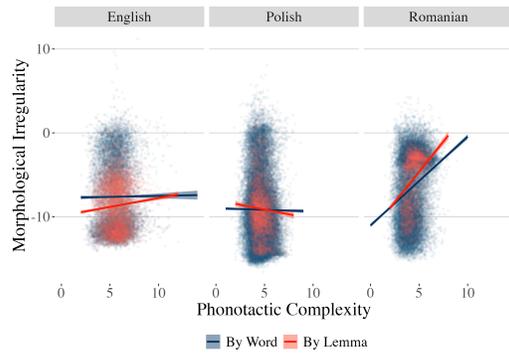
- Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. [NorthEuraLex: a wide-coverage lexical database of Northern Eurasia](#). *Language Resources and Evaluation*, 54:273–301.
- 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.
- Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge.
- Jennifer Hay and Harald Baayen. 2003. Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, 15:99–130.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- C.F. Hockett. 1955. *A Manual of Phonology*. Waverly Press.
- Jackson L. Lee, Lucas F. E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. [Massively multilingual pronunciation modeling with WikiPron](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228. European Language Resources Association.
- Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T. Piantadosi. 2018. [Word forms are structured for efficient use](#). *Cognitive Science*, 42(8):3116–3134.
- Gary F. Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harald Clahsen. 1992. [Overregularization in language acquisition](#). *Monographs of the Society for Research in Child Development*, 57(4):i–178.
- André Martinet. 1955. *Économie des changements phonétiques: Traité de phonologie diachronique*. A. Francke.
- Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. [Revisiting the uniform information density hypothesis](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980. Association for Computational Linguistics.
- Stephan C. Meylan and Thomas L. Griffiths. 2021. [The challenges of large-scale, web-based language datasets: Word length and predictability revisited](#). *Cognitive Science*, 45(6):e12983.
- Steven Moran, Daniel McCloy, and Richard Wright, editors. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology.
- David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. [Epitran: Precision G2P for many languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.
- François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. [A cross-language perspective on speech information rate](#). *Language*, 87(3):539–558.
- Steven T. Piantadosi. 2014. [Zipf’s word frequency law in natural language: A critical review and future directions](#). *Psychonomic Bulletin & Review*, 21(5):1112–1130.
- Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. [Word lengths are optimized for efficient communication](#). *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. [Revisiting the optimality of word lengths](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255. Association for Computational Linguistics.
- Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. [Phonotactic complexity and its trade-offs](#). *Transactions of the Association for Computational Linguistics*, 8:1–18.
- Ingo Plag. 2002. [The role of selectional restrictions, phonotactics and parsing in constraining suffix ordering in English](#). In Geert Booij and Jaap Van Marle, editors, *Yearbook of Morphology 2001*, pages 285–314. Springer.
- Morgan Sonderegger. 2023. *Regression Modeling for Linguistic Data*. MIT Press.
- Shijie Wu and Ryan Cotterell. 2019. [Exact hard monotonic attention for character-level transduction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537. Association for Computational Linguistics.
- Shijie Wu, Ryan Cotterell, and Timothy O’Donnell. 2019. [Morphological irregularity correlates with frequency](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126. Association for Computational Linguistics.
- Charles Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT press.
- George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

## A Languages in Analysis

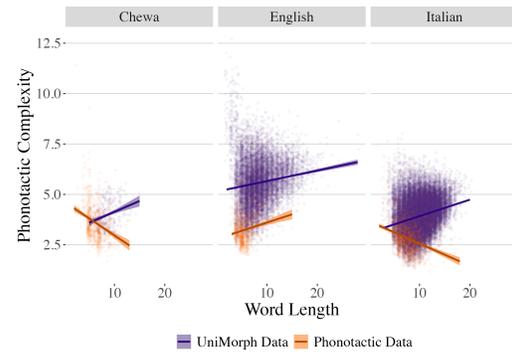
Language	Family	Type	Forms	Lemmas	Phon.	Acc.
Albanian	Indo-Eur.	Fus.	33483	589	NL	81.16
Amharic	Afro-Asiatic	Fus.	46224	2461	WP	92.82
Azerbaijani	Turkic	Agg.	8004	340	NL	76.15
Catalan	Indo-Eur.	Fus.	81576	1547	NL	95.88
Chewa	Atl. Congo	Agg.	4370	227	WP	85.91
Czech	Indo-Eur.	Fus.	50284287	824074	NL	89.89
Dutch	Indo-Eur.	Fus.	55467	4993	NL	92.33
English	Indo-Eur.	Fus.	115523	22765	NL	80.89
French	Indo-Eur.	Fus.	367732	7535	NL	84.96
German	Indo-Eur.	Fus.	179339	15060	NL	86.95
Hungarian	Uralic	Agg.	490394	13989	NL	93.27
Italian	Indo-Eur.	Fus.	509574	10009	NL	98.85
Kazakh	Turkic	Agg.	40283	1755	NL	98.00
Khalka Mongolian	Mongolic	Agg.	30143	2140	NL	82.35
Polish	Indo-Eur.	Fus.	13882543	274550	NL	90.93
Portuguese	Indo-Eur.	Fus.	303996	4001	NL	97.59
Romanian	Indo-Eur.	Fus.	80266	4405	NL	78.09
Russian	Indo-Eur.	Fus.	473481	28068	NL	95.53
Serbo-Croatian	Indo-Eur.	Fus.	840799	24419	NL	92.30
Spanish	Indo-Eur.	Fus.	382955	5460	NL	83.35
Swedish	Indo-Eur.	Fus.	78411	10553	NL	83.23
Turkish	Turkic	Agg.	275460	3579	NL	96.07
Ukrainian	Indo-Eur.	Fus.	20904	1493	NL	85.11
Uzbek	Turkic	Agg.	810	68	WP	87.86
Zulu	Atl. Congo	Agg.	49562	621	WP	75.45
<i>Bengali</i>	Indo-Eur.	Fus.	4443	136	NL	37.75
<i>Cebuano</i>	Austronesian	Agg.	618	97	WP	54.17
<i>Hindi</i>	Indo-Eur.	Fus.	54438	258	NL	68.79
<i>Indonesian</i>	Austronesian	Agg.	27714	3877	WP	49.26
<i>Kabardian</i>	Abkhaz-Adyge	Agg.	3092	250	WP	63.36
<i>Kashubian</i>	Indo-Eur.	Fus.	509	37	WP	1.43
<i>Kyrgyz</i>	Turkic	Agg.	5544	98	WP	29.81
<i>Maltese</i>	Afro-Asiatic	Fus.	3584	112	WP	29.82
<i>Swahili</i>	Atl. Congo	Fus.	14130	185	WP	35.17
<i>Tagalog</i>	Austronesian	Agg.	2912	344	WP	46.22
<i>Tajik</i>	Indo-Eur.	Fus.	77	75	WP	27.01
<i>Telugu</i>	Dravidian	Agg.	1548	127	NL	41.67
<i>Turkmen</i>	Turkic	Agg.	810	68	WP	30.42
<i>Urdu</i>	Indo-Eur.	Fus.	12572	182	WP	50.89
<i>Uyghur</i>	Turkic	Agg.	8178	90	WP	15.12

Table 2: All languages in analysis. *Italicized* languages are excluded. Abbreviations: Fus. - Fusional; Agg. - Agglutinative; Phon. - Phonotactic data source; NL - NorthEuraLex; WP - WikiPron; Acc. - Morphology Model Accuracy

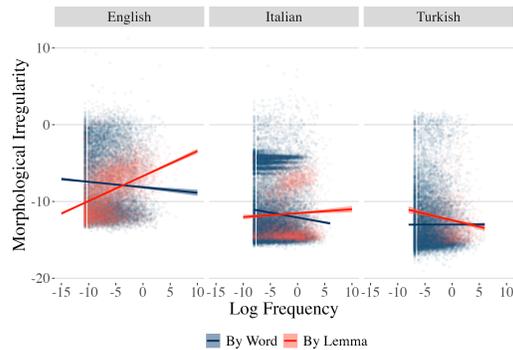
## B Plots of Individual Languages



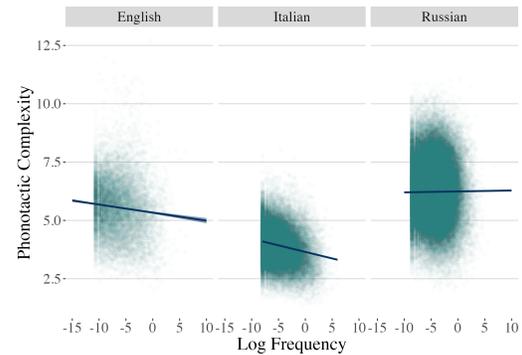
(a) Morphological irregularity and phonotactic complexity data for English, Polish, and Romanian, with linear model predictions from Tab. 1 and 95% CIs.



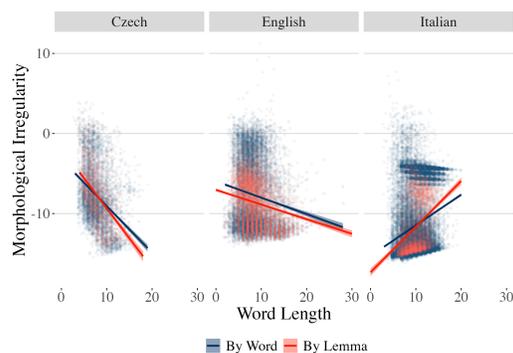
(b) Phonotactic complexity and word length data for Chewa, English, and Italian, with linear model predictions from Tab. 1 and 95% CIs.



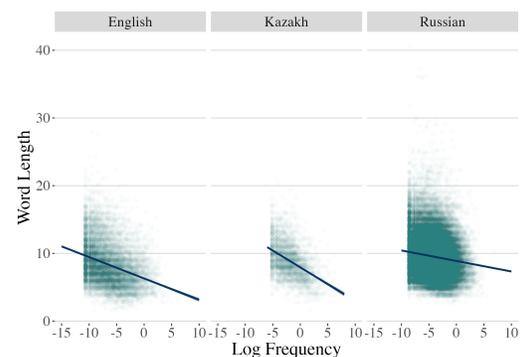
(c) Morphological irregularity and frequency data for English, Italian, and Turkish, with linear model predictions from Tab. 1 and 95% CIs.



(d) Phonotactic complexity and frequency data for English, Italian, and Russian, with linear model predictions from Tab. 1 and 95% CIs.



(e) Morphological irregularity and word length data for Czech, English, and Italian, with linear model predictions from Tab. 1 and 95% CIs.



(f) Word Length and frequency data for English, Kazakh, and Russian, with linear model predictions from Tab. 1 and 95% CIs.