Phonotactic Complexity and Its Trade-offs

Tiago Pimentel
University of Cambridge
tp472@cam.ac.uk

Brian Roark
Google
roark@google.com

Ryan Cotterell
University of Cambridge
rdc42@cam.ac.uk

Abstract

We present methods for calculating a measure of phonotactic complexity—bits per phoneme—that permits a straightforward cross-linguistic comparison. When given a word, represented as a sequence of phonemic segments such as symbols in the international phonetic alphabet, and a statistical model trained on a sample of word types from the language, we can approximately measure bits per phoneme using the negative log-probability of that word under the model. This simple measure allows us to compare the entropy across languages, giving insight into how complex a language's phonotactics is. Using a collection of 1016 basic concept words across 106 languages, we demonstrate a very strong negative correlation of -0.74 between bits per phoneme and the average length of words.

1 Introduction

One prevailing view on system wide phonological complexity is that as one aspect increases in complexity (e.g., size of phonemic inventory), another reduces in complexity (e.g., degree of phonotactic interactions). Underlying this claim—the so-called compensation hypothesis (Martinet, 1955; Moran and Blasi, 2014)—is the conjecture that languages are, generally speaking, of roughly equivalent complexity, that is, no language is overall inherently more complex than another. This conjecture is widely accepted in the literature and dates back at least to the work of Hockett (1958). Because along any one axis, a language may be more complex than another, this conjecture has a corollary that compensatory relationships between different types of complexity must exist. Such compensation has been hypothesized to be the result of natural processes of historical change, and is sometimes attributed to a potential linguistic universal of equal communicative capacity (Pellegrino et al., 2011; Coupé et al., 2019).

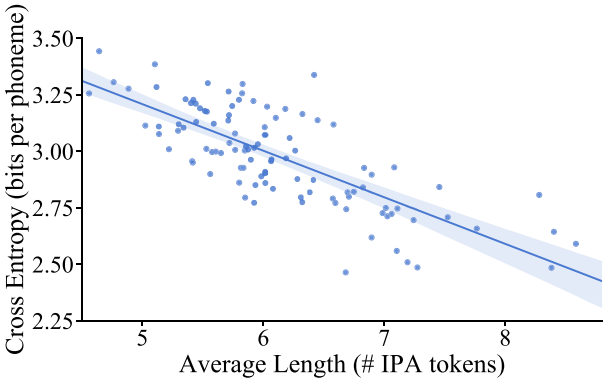

Figure 1: Bits per phoneme vs average word length using an LSTM language model.

Methods for making hypotheses about linguistic complexity objectively measurable and testable have long been of interest, though existing measures are typically relatively coarse—see, for example, Moran and Blasi (2014) and §2 below for reviews. Briefly, counting-based measures such as inventory sizes (e.g., numbers of vowels, consonants, syllables) typically play a key role in assessing phonological complexity. Yet, in addition to their categorical nature, such measures generally do not capture longer-distance (e.g., cross-syllabic) phonological dependencies such as vowel harmony. In this paper, we take an information-theoretic view of phonotactic complexity, and advocate for a measure that permits straightforward cross-linguistic comparison: bits per phoneme. For each language, a statistical language model over words (represented as phonemic sequences) is trained on a sample of types from the language, and then used to calculate the bits per phoneme for new samples, thus providing an upper bound of the actual entropy (Brown et al., 1992).

Transactions of the Association for Computational Linguistics, vol. 8, pp. 1–18, 2020. https://doi.org/10.1162/tacl_a_00296
Action Editor: Eric Fosler-Lussier. Submission batch: 5/2019; Revision batch: 9/2019; Published 1/2020.

Characterizing phonemes via information theoretic measures goes back to Cherry et al. (1953), who discussed the information content of phonemes in isolation, based on the presence or absence of distinctive features, as well as in groups, (e.g., trigrams or possibly syllables). Here we leverage modern recurrent neural language modeling methods to build models over full word forms represented as phoneme strings, thus capturing any dependencies over longer distances (e.g., harmony) in assigning probabilities to phonemes in sequence. By training and evaluating on comparable corpora in each language, consisting of concept-aligned words, we can characterize and compare their phonotactics. Probabilistic characterizations of phonotactics have been used extensively in psycholinguistics (see §2.4), but such methods have generally been used to assess single words within a lexicon (e.g., classifying high versus low probability words during stimulus construction), rather than information-theoretic properties of the lexicon as a whole, which our work explores.

The empirical portion of our paper exploits this information-theoretic take on complexity to examine multiple aspects of phonotactic complexity:

(i) **Bits per Phoneme and Word Length:** In §5.1, we show a very high negative correlation of -0.74 between bits per phoneme and average word length for the same 1016 basic concepts across 106 languages. This correlation is plotted in Figure 1. In contrast, conventional phonotactic complexity measures (e.g., number of consonants in an inventory) demonstrate poor correlation with word length. Our results are consistent with Pellegrino et al. (2011), who show a similar correlation in speech.[1] We additionally establish, in §5.2, that the correlation persists when controlling for characteristics of long words (e.g., early versus late positions in the word).

(ii) **Constraining Language:** Despite often being thought of as adding complexity, processes like vowel harmony and final-obstruent devoicing *improve* the predictability of subsequent segments by constraining the number of well-formed forms. Thus, they reduce complexity measured in bits per phoneme. We validate our models by systematically removing certain constraints in our corpora in §5.3.

(iii) **Intra- versus Inter-Family Correlation:** Additionally, we present results in §5.4 showing that our complexity measure not only correlates with word length in a diverse set of languages, but also intra language families. Standard measures of phonotactic complexity do not show such correlations.

(iv) **Explicit feature representations:** We also find (in §5.5) that methods for including features explicitly in the representation, using methods described in §4.1, yield little benefit except in an extremely low-resource condition.

Our methods[2] permit a straightforward cross-linguistic comparison of phonotactic complexity, which we use to demonstrate an intriguing trade-off with word length. Before motivating and presenting our methods, we next review related work on measuring complexity and phonotactic modeling.

2 Background: Phonological Complexity

2.1 Linguistic Complexity

Linguistic complexity is a nuanced topic. For example, one can judge a particular sentence to be syntactically complex relative to other sentences in the language. However, one can also describe a language as a whole as being complex in one aspect or another (e.g., polysynthetic languages are often deemed morphologically complex). In this paper, we look to characterize phonotactics at the language level. However, we use methods more typically applied to specific sentences in a language, for example in the service of psycholinguistic experiments.

In cross-linguistic studies, the term *complexity* is generally used chiefly in two manners, which Moran and Blasi (2014) follow Miestamo (2006) in calling *relative* and *absolute*. Relative

[1] See also Coupé et al. (2019), where syllable-based bigram models are used to establish a comparable information rate in speech across 17 typologically diverse languages.

[2] Code to train these models and reproduce results is available at `https://github.com/tpimentelms/phonotactic-complexity`.

complexity metrics are those that capture the difficulty of learning or processing language, which Miestamo (2006) points out may vary depending on the individual (hence, is relative to the individual being considered). For example, vowel harmony, which we will touch upon later in the paper, may make vowels more predictable for a native speaker, hence less difficult to process; for a second language learner, however, vowel harmony may increase difficulty of learning and speaking. Absolute complexity measures, in contrast, assess the number of parts of a linguistic (sub-)system (e.g., number of phonemes or licit syllables).

In the sentence processing literature, *surprisal* (Hale, 2001; Levy, 2008) is a widely used measure of processing difficulty, defined as the negative log probability of a word given the preceding words. Words that are highly predictable from the preceding context have low surprisal, and those that are not predictable have high surprisal. The phonotactic measure we advocate for in §3 is related to surprisal, though at the phoneme level rather than the word level, and over words rather than sentences. Measures related to phonotactic probability have been used in a range of psycholinguistic studies—see §2.4—though generally to characterize single words within a language (e.g., high versus low probability words) rather than for cross-linguistic comparison as we are here. Returning to the distinction made by Miestamo (2006), we will remain agnostic in this paper as to which class (relative or absolute) such probabilistic complexity measures fall within, as well as whether the trade-offs that we document are bona fide instances of complexity compensation or are due to something else, for example, related to the communicative capacity as hypothesized by Pellegrino et al. (2011). We bring up this terminological distinction primarily to situate our use of *complexity* within the diverse usage in the literature.

Additionally, however, we will point out that an important motivation for those advocating for the use of absolute over relative measures to characterize linguistic complexity in cross-linguistic studies is a practical one. Miestamo (2006, 2008) claims that relative complexity measures are infeasible for broadly cross-linguistic studies because they rely on psycholinguistic data, which is neither common enough nor sufficiently easily comparable across languages to support reliable comparison. In this study, we demonstrate that surprisal and related measures are not subject to the practical obstacles raised by Miestamo, independently of whichever class of complexity they fall into.

2.2 Measures of Phonological Complexity

The complexity of phonemes has long been studied in linguistics, including early work on the topic by Zipf (1935), who argued that a phoneme's articulatory effort was related to its frequency. Trubetzkoy (1938) introduced the notion of markedness of phonological features, which bears some indirect relation to both frequency and articulatory complexity. Phonological complexity can be formulated in terms of language production (e.g., complexity of planning or articulation) or in terms of language processing (e.g., acoustic confusability or predictability), a distinction often framed around the ideas of articulatory complexity and perceptual salience (see, e.g., Maddieson, 2009). One recent instantiation of this was the inclusion of both focalization and dispersion to model vowel system typology (Cotterell and Eisner, 2017).

It is also natural to ask questions about the phonological complexity of an entire language in addition to that of individual phonemes—whether articulatory or perceptual, phonemic or phonotactic. Measures of such complexity that allow for cross-linguistic comparison are non-trivial to define. We review several previously proposed metrics here.

Size of Phoneme Inventory. The most basic metric proposed for measuring phonological complexity is the number of distinct phonemes in the language's phonemic inventory (Nettle, 1995). There has been considerable historical interest in counting both the number of vowels and the number of consonants (see, e.g., Hockett, 1955; Greenberg et al., 1978; Maddieson and Disner, 1984). Phoneme inventory size has its limitations—it ignores the phonotactics of the language. It does, however, have the advantage that it is relatively easy to compute without further linguistic analysis. Correlations between the size of vowel and consonant inventories (measured in number of phonemes) have been extensively studied, with contradictory results presented in the literature—see, for example, Moran and Blasi

(2014) for a review. Increases in phonemic inventory size are also hypothesized to negatively correlate with word length measured in phonemes (Moran and Blasi, 2014). In Nettle (1995), an inverse relationship was demonstrated between the size of the segmental inventory and the mean word length for 10 languages, and similar results (with some qualifications) were found for a much larger collection of languages in Moran and Blasi (2014).[3] We will explore phoneme inventory size as a baseline in our studies in §5.

Markedness in Phoneme Inventory. A refinement of phoneme inventory size takes into account markedness of the individual phonemes. McWhorter (2001) argues that one should judge the complexity of an inventory by counting the cross-linguistic frequency of the phonemes in the inventory, channeling the spirit of Greenberg (1966). Thus, a language that has fewer phonemes, but contains cross-linguistically marked ones such as clicks, could be more complex.[4] McWhorter justifies this definition with the observation that no attested language has a phonemic inventory that consists only of marked segments. Beyond frequency, Lindblom and Maddieson (1988) propose a tripartite markedness rating scheme for various consonants. In this paper, we are principally looking at phonotactic complexity, though we did examine the joint training of models across languages, which can be seen as modeling some degree of typicality and markedness.

Word Length. As stated earlier, word length, measured in the number of phonemes in a word, has been shown to negatively correlate with other complexity measures, such as phoneme inventory size (Nettle, 1995; Moran and Blasi, 2014). To the extent that this is interpreted as being a compensatory relation, this would indicate that word length is being taken as an implicit measure of complexity. Alternatively, word length has a natural interpretation in terms of information rate, so trade-offs could be attributed to communicative capacity (Pellegrino et al., 2011; Coupé et al., 2019).

Number of Licit Syllables. Phonological constraints extend beyond individual units to the structure of entire words themselves, as we discussed above; so why stop at counting phonemes? One step in that direction is to investigate the syllabic structure of language, and count the number of possible licit syllables in the language (Maddieson and Disner, 1984; Shosted, 2006). Syllabic complexity brings us closer to a more holistic measure of phonological complexity. Take, for instance, the case of Mandarin Chinese. At first blush, one may assume that Mandarin has a complex phonology due to an above-average-sized phonemic inventory (including tones); closer inspection, however, reveals a more constrained system. Mandarin only admits two codas: /n/ and /ŋ/.

Although syllable inventories and syllable-based measures of phonotactic complexity—for example, highest complexity syllable type in Maddieson (2006)—do incorporate more of the constraints at play in a language versus segment-based measures, (a) they remain relatively simple counting measures; and (b) phonological constraints do not end at the syllable boundary. Phenomena such as vowel harmony operate at the word level. Further, the combinatorial possibilities captured by a syllabic inventory, as discussed by Maddieson (2009), can be seen as a sort of categorical version of a distribution over forms. Stochastic models of word-level phonotactics permit us to go beyond simple enumeration of a set, and characterize the distribution in more robust information-theoretic terms.

2.3 Phonotactics

Beyond characterizing the complexity of phonemes in isolation or the number of syllables, one can also look at the system determining how phonemes combine to form longer sequences in order to create words. The study of which sequences of phonemes constitute natural-sounding words is called phonotactics. For example, as Chomsky and Halle (1965) point out in their oft-cited example, *brick* is an actual word in English;[5]

[3] Note that by examining negative correlations between word length and inventory size within the context of complexity compensation, word length is also being taken implicitly as a complexity measure, as we shortly make explicit.

[4] McWhorter (2001) was one of the first to offer a quantitative treatment of linguistic complexity at all levels. Note, however, he rejects the equal complexity hypothesis, arguing that creoles are simpler than other languages. As our data contain no creole languages, we cannot address this hypothesis; rather, we only compare non-creole languages.

[5] For convenience, we just use standard orthography to represent actual and possible words, rather than phoneme strings.

blick is not an actual word in English, but is judged to be a possible word by English speakers; and *bnick* is neither an actual nor a possible word in English, due to constraints on its phonotactics.

Psycholinguistic studies often use phonotactic probability to characterize stimuli within a language—see §2.4 for details. For example, Goldrick and Larson (2008) demonstrate that both articulatory complexity and phonotactic probability influence the speed and accuracy of speech production. Measures of the overall complexity of a phonological system must thus also account for phonotactics.

Cherry et al. (1953) took an explicitly information-theoretic view of phonemic structure, including discussions of both encoding phonemes as feature bundles and the redundancy within groups of phonemes in sequence. This perspective of phonemic coding has led to work on characterizing the explicit rules or constraints that lead to redundancy in phoneme sequences, including morpheme structure rules (Halle, 1959) or conditions (Stanley, 1967). Recently, Futrell et al. (2017) took such approaches as inspiration for a generative model over feature dependency graphs. We, too, examine decomposition of phonemes into features for representation in our model (see §4.1), though in general this only provided modeling improvements over atomic phoneme symbols in a low-resource scenario.

Much of the work in phonotactic modeling is intended to explain the sorts of grammaticality judgments exemplified by the examples of Chomsky and Halle (1965) discussed earlier. Recent work is typically founded on the commonly held perspective that such judgements are gradient,[6] hence amenable to stochastic modeling (e.g., Hayes and Wilson, 2008; Futrell et al., 2017—though cf. Gorman, 2013). In this paper, however, we are looking at phonotactic modeling as the means for assessing phonotactic complexity and discovering potential evidence of trade-offs cross-linguistically, and are not strictly speaking evaluating the model on its ability to capture such judgments, gradient or otherwise.

2.4 Phonotactic Probability and Surprisal

A word's phonotactic probability has been shown to influence both processing and learning of language. Words with high phonotactic probabilities (see brief notes on the operationalization of this below) have been shown to speed speech processing, both recognition (e.g., Vitevitch and Luce, 1999) and production (e.g., Goldrick and Larson, 2008). Phonotactically probable words in a language have also been shown to be easier to learn (Storkel, 2001, 2003; Coady and Aslin, 2004, *inter alia*); although such an effect is also influenced by *neighborhood density* (Coady and Aslin, 2003), as are the speech processing effects (Vitevitch and Luce, 1999). Informally, phonological neighborhood density is the number of similar sounding words in the lexicon, which, to the extent that high phonotactic probability implies phonotactic patterns frequent in the lexicon, typically correlates to some degree with phonotactic probability—that is, dense neighborhoods will typically consist of phonotactically probable words. Some effort has been made to disentangle the effect of these two characteristics (Vitevitch and Luce, 1999; Storkel et al., 2006; Storkel and Lee, 2011, inter alia).

Within the psycholinguistics literature referenced above, phonotactic probability was typically operationalized by summing or averaging the frequency with which single phonemes and phoneme bigrams occur, either overall or in certain word positions (initial, medial, final); and neighborhood density of a word is typically the number of words in a lexicon that have Levenshtein distance 1 from the word (see, e.g., Storkel and Hoover, 2010). Note that these measures are used to characterize specific words, that is, given a lexicon, these measures allow for the designation of high versus low phonotactic probability words and high versus low neighborhood density words, which is useful for designing experimental stimuli. Our bits per phoneme measure, in contrast, is used to characterize the distribution over a sample of a language rather than specific individual words in that language.

Other work has made use of phonotactic probability to examine how such processing and learning considerations may impact the lexicon. Dautriche et al. (2017) take phonotactic probability as one component of ease of processing and learning—the other being perceptual confusability—that might influence how lexicons become organized over time. They operationalize phonotactic probability via generative phonotactic

[6]Gradient judgments would account for the fact that *bwick* is typically judged to be a possible English word like *blick* but not as good. In other words, *bwick* is better than *bnick* but not as good as *blick*.

models (phoneme n-gram models and probabilistic context-free grammars with syllable structure), hence closer to the approaches described in this paper than the work cited earlier in this section. Generating artificial lexicons from such models, they find that real lexicons demonstrate higher network density (as indicated by Levenshtein distances, frequency of minimal pairs, and other measures) than the randomly generated lexicons, suggesting that the pressure towards highly clustered lexicons is driven by more than just phonotactic probability.

Evidence of pressure towards communication efficiency in the lexicon has focused on both phonotactic probability and word length. The information content, as measured by the probability of a word in context, is shown to correlate with orthographic length (taken as a proxy for phonological word length) (Piantadosi et al., 2009, 2011). Piantadosi et al. (2012) show that words with lower bits per phoneme have higher rates of homophony and polysemy, in support of their hypothesis that words that are easier to process will have higher levels of ambiguity. Relatedly, Mahowald et al. (2018) demonstrate, in nearly all of the 96 languages investigated, a high correlation between orthographic probability (as proxy for phonotactic probability) and frequency, that is, frequently used forms tend to be phonotactically highly probable, at least within the word lengths examined (3–7 symbols). A similar perspective on the role of predictability in phonology holds that words that are high probability in context (i.e., low surprisal) tend to be reduced, and those that are low probabilty in context are prone to change (Hume and Mailhot, 2013) or to some kind of enhancement (Hall et al., 2018). As Priva and Jaeger (2018) point out, frequency, predictabilty and information content (what they call *informativity* and operationalize as expected predictability) are related and easily confounded, hence the perspectives presented by these papers are closely related. Again, for these studies and those cited earlier, such measures are used to characterize individual words within a language rather than the lexicon as a whole.

3 The Probabilistic Lexicon

In this work, we are interested in a hypothetical phonotactic distribution $p_{lex} : \Sigma^* \rightarrow \mathbb{R}_+$ over the lexicon. In the context of phonology, we interpret Σ^* as all ‘‘universally possible phonological surface forms,’’ following Hayes and Wilson (2008).[7] The distribution p_{lex}, then, assigns a probability to every possible surface form $\mathbf{x} \in \Sigma^*$. In the special case that p_{lex} is a log-linear model, then we arrive at what is known as a maximum entropy grammar (Goldwater and Johnson, 2003; Jäger, 2007). A good distribution p_{lex} should assign high probability to phonotactically valid words, including non-existent ones, but little probability to phonotactic impossibilities. For instance, the possible English word *blick* should receive much higher probability than $^*bnick$, which is not a possible English word. The lexicon of a language, then, is considered to be generated as samples without replacement from p_{lex}.

If we accept the existence of the distribution p_{lex}, then a natural manner by which we should measure the phonological complexity of language is through Shannon's entropy (Cover and Thomas, 2012)

$$H(p_{lex}) = -\sum_{\mathbf{x}\in\Sigma^*} p_{lex}(\mathbf{x}) \log p_{lex}(\mathbf{x}) \quad (1)$$

The units of $H(p_{lex})$ are bits as we take $\log$ to be base 2. Specifically, we will be interested in **bits per phoneme**, that is, how much information each phoneme in a word conveys, on average.

3.1 Linguistic Rationale

Here we seek to make a linguistic argument for the adoption of bits per phoneme as a metric for complexity in the phonological literature. Bits are fundamentally units of predictability: If the entropy of your distribution is higher (i.e., more bits), then it is *less* predictable, and if the entropy is lower, (i.e., fewer bits), then it is *more* predictable with an entropy of 0 indicating determinism.

Holistic Treatment. When we just count the number of distinctions in individual parts of the phonology, for example, number of vowels or number of consonants, we do not get a holistic picture of how these pieces interact. A simple probabilistic treatment will *inherently* capture nuanced interactions. Indeed, it is not clear how to balance the number of consonants, the number of vowels and the number of tones to

[7] Hayes and Wilson (2008) label Σ^* as Ω.

English	Turkish	English	Turkish
ear	kulak	throat	boğaz
rain	yağmur	foam	köpük
summit	zirve	claw	pençe
nail	tırnak	herd	sürü
horse	beygir	dog	köpek

Table 1: Turkish evinces two types of vowel harmony, front-back and round-unround. Here we focus on just front-back harmony. The examples in the table are such that all vowels in a word are either back (ı, u, a, o) or front (i, ü, e, ö), which is generally the case.

get a single number of phonological complexity. Probabilistically modeling phonological strings, however, does capture this. We judge the complexity of a phonological system as its entropy.

Longer-Distance Dependencies. To the best of our knowledge, the largest phonological unit that has been considered in the context of cross-linguistic phonological complexity is the syllable, as discussed in §2.2. However, the syllable clearly has limitations. It cannot capture, tautologically, cross-syllabic phonological processes, which abound in the languages of the world. For instance, vowel and consonant harmony are quite common cross-linguistically. Naturally, a desideratum for any measure of phonological complexity is to consider all levels of phonological processes. Examples of vowel harmony in Turkish are presented in Table 1.

Frequency Information. None of the previously proposed phonological complexity measures deals with the fact that certain patterns are more frequent than others; probability models inherently handle this as well. Indeed, consider the role of the unvoiced velar fricative /x/ in English; while not part of the canonical consonant inventory, /x/ nevertheless appears in a variety of loanwords. For instance, many native English speakers do pronounce the last name of composer Johann Sebastian Bach as /bax/. Moreover, English phonology acts upon /x/ as one would expect: Consider Morris Halle's (1978) example *Sandra out-Bached Bach*, where the second word is pronounced /out-baxt/ with a final /t/ rather than a /d/. We conclude that /x/ is in the consonant inventory of at least some native English speakers. However, counting it on equal status with the far more common /k/ when determining complexity seems incorrect. Our probabilistic metric covers this corner case elegantly.

Relatively Modest Annotation Requirements. Many of these metrics require a linguist's analysis of the language. This is a tall order for many languages. Our probabilistic approach only requires relatively simple annotations, namely, a Swadesh (1955)-style list in the international phonetic alphabet (IPA) to estimate a distribution. When discussing why he limits himself to counting complexities, Maddieson (2009) writes:

> [t]he factors considered in these studies only involved the inventories of consonant and vowel contrasts, the tonal system, if any, and the elaboration of the syllable canon. It is relatively easy to find answers for a good many languages to such questions as 'how many consonants does this language distinguish?' or 'how many types of syllable structures does this language allow?'

The moment one searches for data on more elaborate notions of complexity, for example, the existence of vowel harmony, one is faced with the paucity of data—a linguist must have analyzed the data.

3.2 Constraints Reduce Entropy

Many phonologies in the world use hard constraints (e.g., a syllable final obstruent must be devoiced or the vowels in a word must be harmonic). Using our definition of phonological complexity as entropy, we can prove a general result that *any* hard-constraining process will reduce entropy, thus making the phonology *less* complex. The fact that this holds for any hard contraint, be it vowel harmony or final-obstruent devoicing, is a fact that conditioning reduces entropy.

3.3 A Variational Upper Bound

If we want to compute Equation (1), we are immediately faced with two problems. First, we do not know p_{lex}: we simply assume the existence of such a distribution from which the words of the lexicon were drawn. Second, even if we did know p_{lex}, computation of the $H(p_{lex})$ would

be woefully intractable, as it involves an infinite sum. Following Brown et al. (1992), we tackle both of these issues together. Note that this line of reasoning follows Cotterell et al. (2018) and Mielke et al. (2019), who use a similar technique for measuring language complexity at the sentence level.

We start with a basic inequality from information theory. For any distribution q_{lex} with the same support as p_{lex}, the cross-entropy provides an upper bound on the entropy, that is

$$H(p_{lex}) \leq H(p_{lex}, q_{lex}) \tag{2}$$

where cross-entropy is defined as

$$H(p_{lex}, q_{lex}) = -\sum_{\mathbf{x} \in \Sigma^*} p_{lex}(\mathbf{x}) \log q_{lex}(\mathbf{x}) \tag{3}$$

Note that Equation (2) is tight if and only if $p_{lex} = q_{lex}$. We still are not done, as Equation (3) still requires knowledge of p_{lex} and involves an infinite sum. However, we are now in a position to exploit samples from p_{lex}. Specifically, given $\tilde{\mathbf{x}}^{(i)} \sim p_{lex}$, we approximate

$$H(p_{lex}, q_{lex}) \approx -\frac{1}{N} \sum_{i=1}^{N} \log q_{lex}(\tilde{\mathbf{x}}^{(i)}) \tag{4}$$

with equality if we let $N \to \infty$. In information theory, this equality in the limit is called the asymptotic equipartition property and follows easily from the weak law of large numbers. Now, we have an empirical procedure for estimating an upper bound on $H(p_{lex})$. For the rest of the paper, we will use the right-hand side of Equation (4) as a surrogate for the phonotactic complexity of a language.

How to Choose q_{lex}? Choosing a good q_{lex} is a two-step process. First, we choose a variational family $\mathcal{Q}$. Then, we choose a specific $q_{lex} \in \mathcal{Q}$ by minimizing the right-hand side of Equation (4)

$$q_{lex} = \text{argsup}_{q \in \mathcal{Q}} \frac{1}{N} \sum_{i=1}^{N} \log q(\tilde{\mathbf{x}}^{(i)}) \tag{5}$$

This procedure corresponds to maximum likelihood estimation. In this work, we consider two variational families: (i) a phoneme n-gram model, and (ii) a phoneme-level RNN language model. We describe each in §4.1.

3.4 A Note on Types and Tokens

To make the implicit explicit, in this work we will exclusively be modeling types, rather than tokens. We briefly justify this discussion from both theoretical and practical concerns. From a theoretical side, a token-based model is unlikely to correctly model an out of vocabulary distribution as very frequent tokens often display unusual phonotactics for historical reasons. A classic example comes from English: Consider the appearance of /ð/. Judging by token-frequency, /ð/ is quite common as it starts some of the most common words in the language: *the*, *they*, *that*, and so forth. However, novel words categorically avoid initial /ð/. From a statistical point of view, one manner to justify type-level modeling is through the Pitman–Yor process (Ishwaran and James, 2003). Goldwater et al. (2006) showed that type-level modeling is a special case of the stochastic process, writing that they ''justif[y] the appearance of type frequencies in formal analyses of natural language.''

Practically, using token-level frequencies, even in a dampened form, is not possible due to the large selection of languages we model. Most of the languages we consider do not have corpora large enough to get reasonable token estimates. Moreover, as many of the languages we consider have a small number of native speakers, and, in extreme cases, are endangered, the situation is unlikely to remedy itself, forcing the phonotactician to rely on types.

4 Methods

4.1 Phoneme-Level Language Models

Notation. Let Σ be a discrete alphabet of symbols from the IPA, including special beginning-of-string and end-of-string symbols. A character level language model (LM) models a probability distribution over Σ^*

$$p(\mathbf{x}) = \prod_{i=1}^{|\mathbf{x}|} p\left(x_i \mid \mathbf{x}_{<i}\right) \tag{6}$$

Trigram LM. n-grams assume the sequence follows a $(n-1)$-order Markov model, conditioning the probability of a phoneme on the $(n-1)$ previous ones

$$f_n(x_i \mid \mathbf{x}_{<i}) = \frac{count(x_i, x_{i-1}, \ldots, x_{i+1-n})}{count(x_{i-1}, \ldots, x_{i+1-n})} \tag{7}$$

where we assume the string $\mathbf{x}$ is properly padded with beginning and end-of-string symbols.

The trigram model used in this work is estimated as the deleted interpolation (Jelinek, 1980) of the trigram, bigram, and unigram relative frequency estimates

$$p_3(x_i \mid \mathbf{x}_{<i}) = \sum_{n=1}^{3} \alpha_n f_n(x_i \mid \mathbf{x}_{<i}) \qquad (8)$$

where the mixture parameters α_n were estimated via Bayesian optimization with a Gaussian prior maximizing the expected improvement on a validation set, as discussed by Snoek et al. (2012).

Recurrent Neural LM. Recurrent neural networks excel in language modeling, being able to capture complex distributions $p(x_i \mid \mathbf{x}_{<i})$ (Mikolov et al., 2010; Sundermeyer et al., 2012). Empirically, recent work has observed dependencies on up to around 200 tokens (Khandelwal et al., 2018). We use a character-level Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber, 1997) language model, which is the state of the art for character-level language modeling (Merity et al., 2018).

Our architecture receives a sequence of tokens $\mathbf{x} \in \Sigma^*$ and embeds each token $x_i \in \Sigma$ using a dictionary-lookup embedding table. This results in vectors $z_i \in \mathbb{R}^d$ which are fed into an LSTM. This LSTM produces a high-dimensional representation of the sequence, often termed hidden states

$$h_i = LSTM\left(z_{i-1}, h_{i-1}\right) \in \mathbb{R}^d \qquad (9)$$

These representations are then fed into a softmax to produce a distribution over the next character

$$p\left(x_i \mid \mathbf{x}_{<i}\right) = \mathit{softmax}\left(W h_i + b\right) \qquad (10)$$

where $W \in \mathbb{R}^{|\Sigma| \times d}$ is a final projection matrix and $b \in \mathbb{R}^{|\Sigma|}$ is a bias term. In our implementation, h_0 is a vector of all zeros and z_0 is the lookup embedding for the beginning-of-string token.

Phoneme Embedding LM. When developing a phoneme-level recurrent neural LM, one can use a base of phonemic features—for example, Phoible (Moran et al., 2014)—to implement a multi-hot embedding such that similar phonemes will have similar embedding representations. A phoneme i will have a set of binary attributes $a_i^{(k)}$ (e.g., stress, sonorant, nasal), each with its corresponding embedding representation $z^{(k)}$. A phoneme embedding will, then, be composed by the element-wise average of each of its features lookup embedding

$$z_i = \frac{\sum_k a_i^{(k)} z^{(k)}}{\sum_k a_i^{(k)}} \qquad (11)$$

where $a_i^{(j)}$ is 1 if phoneme i presents attribute j and $z^{(j)}$ is the lookup embedding of attribute j. This architecture forces similar phonemes, measured in terms of overlap in distinctive features, to have similar representations.

Concept	Language	Word	IPA
eye	portuguese	olho	/oʎu/
ear	finnish	korva	/kɔrʋɑ/
give	north karelian	antua	/ɑntʊɑ/
tooth	veps	hambaz	/hɑmbɑz/
black	northern sami	čáhppes	/t͡ʃaahppes/
immediately	hill mari	töpök	/tørøk/

Table 2: Sample of the lexicon in NorthEuraLex corpus.

4.2 NorthEuraLex Data

We make use of data from the NorthEuraLex corpus (Dellert and Jäger, 2017). The corpus is a concept-aligned multi-lingual lexicon with data from 107 languages. The lexicons contains 1016 ''basic'' concepts. Importantly, NorthEuraLex is appealing for our study as all the words are written in a unified IPA scheme. A sample of the lexicon is provided in Table 2. For the results reported in this paper, we omitted Mandarin, because no tone information was included in its annotations, causing its phonotactics to be greatly underspecified. No other tonal languages were included in the corpus, so all reported results are over 106 languages.

Why Is Base-Concept Aligned Important? Making use of data that are concept-aligned across the languages provides a certain amount of control (to the extent possible) of the influence of linguistic content on the forms that we are modeling. In other words, these forms should be largely comparable across the languages in terms of how common they are in the active vocabulary of adult speakers. Further, base concepts as defined for the collection are more likely to be lemmas without inflection, thus reducing the

influence of morphological processes on the results.[8]

To test this latter assertion, we made use of the UniMorph[9] morphological database (Kirov et al., 2018) to look up words and assess the percentage that correspond to lemmas or base forms. Of the 106 languages in our collection, 48 are also in the UniMorph database, and 46 annotate their lemmas in a way that allowed for simple string matching with our word forms. For these 46 languages, on average we found 313 words in UniMorph of the 1016 concepts (median 328). A mean of 87.2% (median 93.3%; minimum 58.6%) of these matched lemmas for that language in the UniMorph database. This rough string matching approach provides some indication that the items in the corpus are largely composed of such base forms.

Dataset Limitations. Unfortunately, there is less typological diversity in our dataset than we would ordinarily desire. NorthEuraLex draws its languages from 21 distinct language families that are spoken in Europe and Asia. This excludes languages indigenous to the Americas,[10] Australia, Africa, and South-East Asia. Although lamentable, we know of no other concept-aligned lexicon with broader typological diversity that is written in a unified phonetic alphabet, so we must save studies of more typologically diverse sets of languages for future work.

In addition, we note that the process of base concept selection and identification of corresponding forms from each language (detailed in Dellert, 2015, 2017) was non-trivial, and some of the corpus design decisions may have resulted in somewhat biased samples in some languages. For example, there was an attempt to minimize the frequency of loanwords in the dataset, which may make the lexicons in loanword heavy languages, such as English with its extensive Latinate vocabulary, somewhat less representative of everyday use than in other languages. Similarly, the creation of a common IPA representation over this number of languages required choices that could potentially result in corpus artifacts. As with the issue of linguistic diversity, we acknowledge that the resource has some limitations but claim that it is the best currently available dataset for this work.

Splitting the Data. We split the data at the concept level into 10 folds, used for cross validation. We create train-dev-test splits where the training portion has 8 folds ($\approx$ 812 concepts) and the dev and test portions have 1 fold each ($\approx$ 102 concepts). We then create language-specific sets with the language-specific words for the concept to be rendered. Cross-validation allows us to have all 1016 concepts in our test sets (although evaluated using different model instances), and we do our following studies using all of them.

4.3 Artificial Languages

In addition to naturally occurring languages, we are also interested in artificial ones. Why? We wish to validate our models in a controlled setting, quantifying the contribution of specific linguistic phenomena to our complexity measure. Thus, developing artificial languages, which only differ with respect to one phonological property, is useful.

The Role of Final-Obstruent Devoicing. Final-obstruent devoicing *reduces* phonological complexity under our information-theoretic metric. The reason is simple: There are fewer valid syllables as all those with voiced final obstruents are ruled out. Indeed, this point is also true of the syllable counting metric discussed in §2.2. One computational notion of complexity might say that the complexity of the phonology is equal to the number of states required to encode the transduction from an underlying form to a surface form in a minimal finite-state transduction. Note that all Sound Pattern of English (SPE)-style rules may be so encoded (Kaplan and Kay, 1994). Thus, the complexity of the phonotactics could be said to be related to the number of SPE-style rules that operate. In contrast, under our metric, any process that constrains the number of possibilities will, inherently, reduce complexity. The studies in §5.3 allow us to examine the *magnitude* of such a reduction, and validate our models with respect to this expected behavior.

We create two artificial datasets without final-obstruent devoicing based on the German and Dutch portions of NorthEuraLex. We reverse the

[8] Most of the concepts in the dataset do not contain function words and verbs are in the bare infinitive form – (e.g., *have*, instead of *to have*) although there are a few exceptions. For example, the German word *hundert* is represented as *a hundred* in English.

[9] `https://unimorph.github.io`.

[10] Inuit languages, which are genetically related to the languages of Siberia, are included in the lexicon.

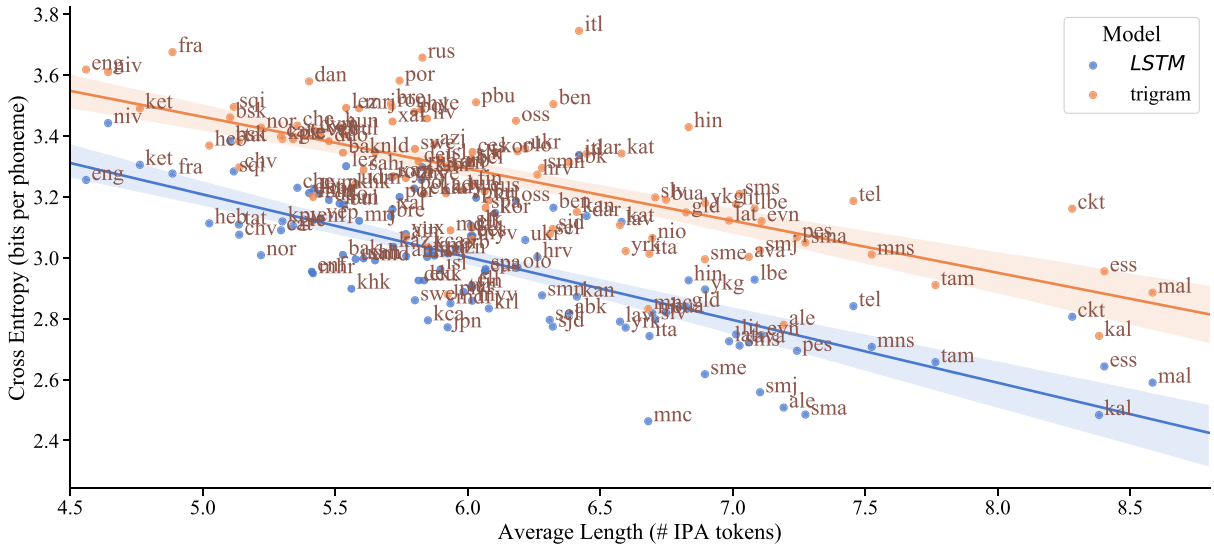

Figure 2: Per-phoneme complexity vs average word length under both a trigram and an LSTM language model.

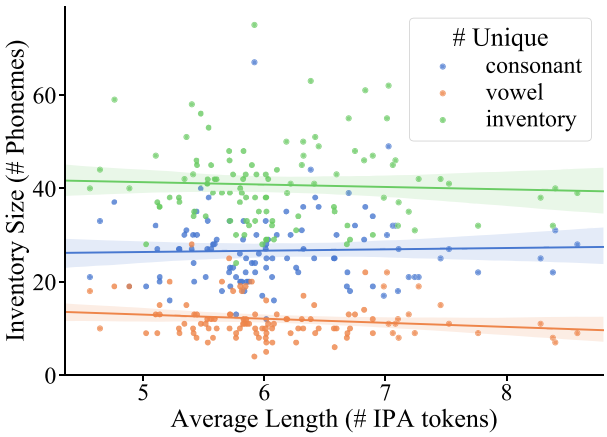

Figure 3: Conventional measures of phonological complexity vs average word length. These complexity measures are based in inventory size.

	Correlation	
Measure	Pearson r	Spearman ρ
Number of:		
phonemes	−0.047	−0.054
vowels	−0.164	−0.162
consonants	0.030	0.045
Bits/phoneme:		
unigram	−0.217	−0.222
trigram	−0.682	−0.672
LSTM	−0.762	−0.744

Table 3: Pearson and Spearman rank correlation coefficients between complexity measures and average word length in phoneme segments.

process, using the orthography as a guide. For example, the German /tsu:k/ is converted to /tsu:g/ based on the orthography *Zug*.

The Role of Vowel Harmony. Like final obstruent devoicing, vowel harmony plays a role in reducing the number of licit syllables. In contrast to final obstruent devoicing, however, vowel harmony acts cross-syllabically. Consider the Turkish lexicon, where most, but not all, basic lexical items obey vowel harmony. Processes like this reduce the entropy of p_{lex} and, thus, can be considered as creating a less complex phonotactics.

For vowel harmony, we create 10 artificial datasets by randomly replacing each vowel in a word with a new sampled (with replacement) vowel from that language's vowel inventory. This breaks all vowel harmony, but keeps the syllabic structure.

5 Results

5.1 Study 1: Bits Per Phoneme Negatively Correlates with Word Length

As stated earlier, Pellegrino et al. (2011) investigated a complexity trade-off with the information density of speech. From a 7-language study they found a strong correlation ($R = -0.94$) between the information density and the syllabic complexity of a language. One hypothesis adduced to explain these findings is that, for functional

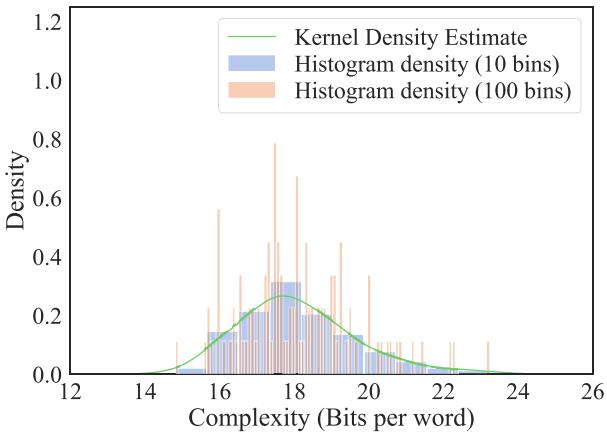

Figure 4: Kernel density estimate (KDE) of the average phonotactic complexity per word across 106 different languages. Different languages tend to present similar complexities (bits per word).

	Correlation	
Measure	Pearson r	Spearman ρ
Per Word:		
all languages	-0.269	-0.312
each language (avg)	-0.220	-0.257
each language (min)	-0.561	-0.607
Per Language:		
Fake (avg)	-0.270	-0.254
Fake (min)	-0.586	-0.568
Real	-0.762	-0.744

Table 4: Pearson and Spearman rank correlation coefficients between complexity measures and word length in phoneme segments. All correlations are statistically significant with $p < 10^{-8}$.

reasons, the rate of linguistic information is very similar cross-linguistically. Inspired by their study, we conduct a similar study with our phonotactic setup. We hypothesize that the bits per phoneme for a given concept correlates with the number of phonemes in the word. Moreover, the bits per word should be similar across languages.

We consider the relation between the average bits per phoneme of a held-out portion of a language's lexicon, as measured by our best language model, and the average length of the words in that language. We present the results in Figures 2 and 3 and in Table 3. We find a strong correlation under the LSTM LM (Spearman's $\rho = -0.744$ with $p < 10^{-19}$). At the same time, we see only a weak correlation under conventional measures of phonotactic complexity, such as vowel inventory size (Spearman's $\rho = -0.162$ $with$ $p = 0.098$). In Figure 4, we plot the kernel density estimate and histogram densities (both 10 and 100 bins) of word-level complexity (bits per word).

5.2 Study 2: Possible Confounds for Negative Correlations

One possible confound for these results is that phonemes later in a word may in general have higher probability given the previous phonemes than those earlier in the string. This sort of positional effect was demonstrated in Dutch (van Son and Pols, 2003), where position in the word accounted for much of the variance in segmental information.[11] To ensure that we are not simply replicating such a positional effect across many languages, we performed several additional analyses.

Truncated Words. First, we calculated the bits-per-phoneme for just the first three positions in the word, and then looked at the correlation between this word-onset bits per phoneme and the average (full) word length in phoneme segments. In other words, for the purpose of calculating bits-per-phoneme, we truncated all words to a maximum of three phonemes, and in such a way explicitly eliminated the contribution of positions later in any word. Using the LSTM model, this yielded a Spearman correlation of $\rho = -0.469$ ($p < 10^{-7}$), in contrast to $\rho = -0.744$ without such truncation (reported in Table 3). This suggests that there is a contribution of later positions to the effect presented in Table 3 that we lose by eliding them, but that even in the earlier positions of the word we are seeing a trade-off with full average word length.

Correlation with phoneme position. We next looked to measure a position effect directly, by calculating the correlation between word position and bits for that position across all languages. Here we find a Spearman correlation of $\rho = -0.429$ ($p < 10^{-200}$), which again supports the contention that later positions in general require fewer bits to encode. Nonetheless, this correlation is

[11] We briefly note that the van Son and Pols (2003) study did not make use of a train/dev/test split of their data, but rather simply analyzed raw relative frequency over their Dutch corpus. As a result, all positions beyond any word onset that is unique in their corpus would have probability 1, leading to a more extreme position effect than we would observe using regularization and validating on unseen forms.

Model	Complexity		Diff
	Orig	Art	
trigram:			
German	3.703	3.708	$0.005(0.13\%)$
Dutch	3.607	3.629	$0.022(0.58\%)^{\dagger}$
LSTM:			
German	3.230	3.268	$0.038(1.18\%)^{\dagger}$
Dutch	3.161	3.191	$0.030(0.95\%)^{\dagger}$

Table 5: Complexities for original and artificial languages when removing final-obstruent devoicing. † represents an statistically significant difference with $p < 0.05$

still weaker than the per-language word length correlation (of $\rho = -0.744$).

Per-Word Correlations. We also calculated the correlation between word length and bits per phoneme across all languages (without averaging per language here). The Spearman correlation between these factors—at the word level using all languages—is $\rho = -0.312$ ($p < 10^{-19}$). Analyzing each language individually, there is an average Spearman's $\rho = -0.257$ ($p < 10^{-19}$) between bits per phoneme and word length. The minimum negative (i.e., highest magnitude) correlation of any language in the set is $\rho = -0.607$. These per word correlations are reported in the upper half of Table 4.

Permuted ''Language'' Correlations. Finally, to determine if our language effects perhaps arise due to the averaging of word lengths and bits per phoneme for each language, we ran a permutation test on languages. We shuffle words (with their precalculated bits per phoneme values) into 106 sets with the same size as the original languages—thus creating fake ''languages''. We take the average word length and bits per phoneme in each of these fake languages and compare the correlation—returning to the ''language'' level this time—with the original correlation. After running this test for 10^4 permutations, we found no shuffled set with an equal or higher Spearman (or Pearson) correlation than the real set. Thus, with a strong confidence ($p < 10^{-4}$) we can state there is a language level effect. Average and minimum negative correlations for these ''fake'' languages (as well as the real set for ease of comparison) are presented in the lower half of Table 4.

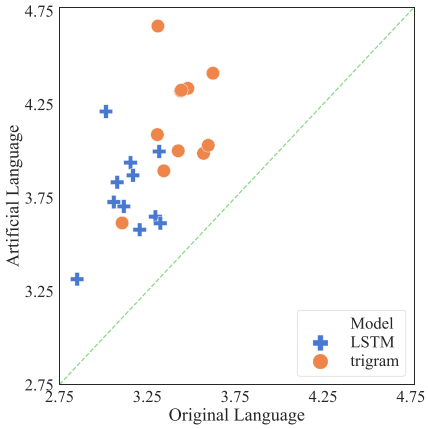

Figure 5: Complexities for natural and artifical languages when removing vowel harmony. A paired permutation test showed all differences present statistical difference with $p < 0.01$.

5.3 Study 3: Constraining Languages Reduces Phonotactic Complexity

Final-obstruent devoicing and vowel harmony reduce the number of licit syllables in a language, hence reducing the entropy. To determine the magnitude that such effects can have on the measure for our different model types, we conduct two studies. In the first, we remove final-obstruent devoicing from the German and Dutch languages in NorthEuraLex, as discussed in §4.3. In the second study, we remove vowel harmony from 10 languages that have it,[12] as also explained in §4.3.

After deriving two artificial languages without obstruent devoicing from both German and Dutch, we used 10-fold cross validation to train models for each language. The statistical relevance of differences between normal and artificial languages was analyzed using paired permutation tests between the pairs. Results are presented in Table 5. We see that the n-gram can capture this change in complexity for Dutch, but not for German. At the same time, the LSTM shows a statistically significant increase of ≈ 0.034 bits per phoneme when we remove obstruent devoicing from both languages. Figure 5 presents a similar impact on complexity from vowel harmony removal, as evidenced by the fact that all points fall above the equality line. Average complexity increased by ≈ 0.62 bits per phoneme (an approximate 16% entropy increase), as measured by our LSTM models.

[12]The languages with vowel harmony are: bua, ckt, evn, fin, hun, khk, mhr, mnc, myv, tel, and tur.

	Correlation	
Measure	Pearson r	Spearman ρ
Number of:		
phonemes	-0.214	-0.095
vowels	-0.383	-0.367
consonants	-0.147	-0.092
Bits/phoneme:		
unigram	-0.267	-0.232
trigram	-0.621	-0.520
LSTM	-0.778	-0.526

Table 6: Pearson and Spearman correlation between complexity measures and word length in phoneme segments averaged across language families.

	Spearman ρ		
Family	**LSTM**	**Vowels**	**# Langs**
Dravidian	-1.0^{*}	-0.894	4
Indo-European	-0.662^{*}	-0.218	37
Nakh-Daghestanian	$-0.771^{\dagger}$	-0.530	6
Turkic	$-0.690^{\dagger}$	$-0.773^{\dagger}$	8
Uralic	-0.874^{*}	$0.363^{\dagger}$	26

* Statistically significant with $p < 0.01$
† Statistically significant with $p < 0.1$

Table 7: Spearman correlation between complexity measures and average word length per language family. Phonotactic complexity in bits per phoneme presents very strong intra-family correlation with word length in three of the five families. Size of vowel inventory presents intra-family correlation in Turkic and Uralic.

In both of these artificial language scenarios, the LSTM models appeared more sensitive to the constraint removal, as expected.

5.4 Study 4: Negative Trade-off Persists Within and Across Families

Moran and Blasi (2014) investigated the correlation between the number of phonological units in a language and its average word length across a large and varied set of languages. They found that, although these measures of phonotactic complexity (number of vowels, consonants or phonemes in a language) are correlated with word length when measured *across* a varied set of languages, such a correlation usually does not hold *within* language families. We hypothesize that this is due to their measures being rather coarse approximations to phonotactic complexity, so that only large changes in the language would show significant correlation given the noise. We also hypothesize that our complexity measure is less noisy, hence should be able to yield significant correlations both within and across families.

Results in Table 3 show a strong correlation for the LSTM measure, while they show a weak one for conventional measures of complexity. As stated before, Moran and Blasi (2014) found that vowel inventory size shows a strong correlation to word length on a diverse set of languages, but, as mentioned in §4.2, our dataset is more limited than desired. To test if we can mitigate this effect we average the complexity measures and word length per family (instead of per language) and calculate the same correlations again. These results are presented in Table 6 and show that when we average these complexity measures per family we indeed find a stronger correlation between vowel inventory size and average word length, although with a higher null hypothesis probability (Spearman's $\rho = -0.367$ with $p = 0.111$). We also see our LSTM based measure still shows a strong correlation (Spearman's $\rho = -0.526$ with $p = 0.017$).

We now analyze these correlations intra families, for all family languages in our dataset with at least 4 languages. These results are presented in Table 7. Our LSTM based phonotactic complexity measure shows strong intra-family correlation with average word length for all five analyzed language families ($-0.662 \geq \rho \geq -1.0$ with $p < 0.1$). At the same time, vowel inventory size only shows a negative statistically significant correlation within Turkic.

5.5 Study 5: Explicit Feature Representations Do Not Generally Improve Models

Table 3 presents strong correlations when using an LSTM with standard one-hot lookup embedding. Here we train LSTMs with three different phoneme embedding models: (1) a typical Lookup embedding, in which each Phoneme has an associated embedding; (2) a phoneme features based embedding, as explained in §4.1; (3) the concatenation of the Lookup and the Phoneme embedding. We also train these models both using independent models for each language, and with independent

Model	Complexity	Spearman ρ
n-Grams:		
unigram	4.477	−0.222
trigram	3.270	−0.672
Independent Embeddings:		
Lookup	2.976	−0.744
Phoneme	2.992	−0.741
Lookup + Phoneme	2.975	−0.752
Shared Embeddings:		
Lookup	2.988	−0.743
Phoneme	2.977	−0.744
Lookup + Phoneme	2.982	−0.740

Table 8: Average cross-entropy across all languages and the correlation between complexity and average word length for different models.

models, but sharing embedding weights across languages.

We first analyze these model variants under the same lens as used in Study 1. Table 8 shows the correlations between the complexity measure resulting from each of this models and the average number of phonemes in a word. We find strong correlations for all of them ($-0.740 \geq \rho \geq -0.752$ with $p < 10^{-18}$). We also present in Table 8 these models' cross entropy, averaged across all languages. At least for the methods that we are using here, we derived no benefit from either more explicit featural representations of the phonemes or by sharing the embeddings across languages.

We also investigated scenarios using less training data, and it was only in very sparse scenarios (e.g., using just 10% of the training used in our standard trials, or 81 example words) where we observed even a small benefit to explicit feature representations and shared embeddings.

6 Conclusion

We have presented methods for calculating a well-motivated measure of phonotactic complexity: bits per phoneme. This measure is derived from information theory and its value is calculated using the probability distribution of a language model. We demonstrate that cross-linguistic comparison is straightforward using such a measure, and find a strong negative correlation with average word length. This trade-off with word length can be seen as an example of complexity compensation or perhaps related to communicative capacity.

Acknowledgments

We thank Damián E. Blasi for his feedback on previous versions of this paper and the anonymous reviewers, as well as action editor Eric Fosler-Lussier, for their constructive and detailed comments—the paper is much improved as a result.

References

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

E. Colin Cherry, Morris Halle, and Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language*, pages 34–46.

Noam Chomsky and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.

Jeffry A. Coady and Richard N. Aslin. 2003. Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, 30(2):441–469.

Jeffry A. Coady and Richard N. Aslin. 2004. Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, 89(3):183–213.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192. Vancouver, Canada. Association for Computational Linguistics.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? 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 536–541. New Orleans, Louisiana. Association for Computational Linguistics.

Christophe Coupé, Yoon 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.

Thomas M. Cover and Joy A. Thomas. 2012. *Elements of Information Theory*, John Wiley & Sons.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

Johannes Dellert. 2015. Compiling the Uralic dataset for NorthEuraLex, a lexicostatistical database of northern Eurasia. In *First International Workshop on Computational Linguistics for Uralic Languages*.

Johannes Dellert. 2017. *Information-Theoretic Causal Inference of Lexical Flow*. Ph.D. thesis, University of Tübingen.

Johannes Dellert and Gerhard Jäger. 2017. NorthEuraLex (version 0.9). `http://northeuralex.org/`

Richard Futrell, Adam Albright, Peter Graff, and Timothy J. O'Donnell. 2017. A generative model of phonotactics. *Transactions of the Association for Computational Linguistics*, 5:73–86.

Matthew Goldrick and Meredith Larson. 2008. Phonotactic probability influences speech production. *Cognition*, 107(3):1155–1164.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*.

Sharon Goldwater, Mark Johnson, and Thomas L. Griffiths. 2006. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, pages 459–466.

Kyle Gorman. 2013. Generative phonotactics. Ph.D. thesis, University of Pennsylvania.

Joseph Greenberg. 1966. *Language universals, with special reference to feature hierarchies.* Mouton, The Hague.

Joseph H. Greenberg, Charles A. Ferguson, and Edith Moravcsik, editors. 1978. *Universals of Human Language. Vol. 2: Phonology*, Stanford University Press.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8.

Kathleen Currie Hall, Elizabeth Hume, T. Florian Jaeger, and Andrew Wedel. 2018. The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, 4(s2).

Morris Halle. 1959. *The Sound Pattern of Russian*, Mouton, The Hague.

Morris Halle. 1978, Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan, and G. Miller, editors, *Linguistic Theory and Psychological Reality*, pages 294–303, The MIT Press, Cambridge, MA.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Charles Francis Hockett. 1955. *A manual of phonology*, Waverly Press, Baltimore, MD.

Charles Francis Hockett. 1958. *A course in modern linguistics*, Macmillan, New York.

Elizabeth Hume and Frédéric Mailhot. 2013. The role of entropy and surprisal in phonologization and language change. In Alan C.L. Yu, editor, *Origins of sound change: Approaches to phonologization*, pages 29–47, Oxford University Press, Oxford.

Hemant Ishwaran and Lancelot F. James. 2003. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, pages 1211–1235.

Gerhard Jäger. 2007. Maximum entropy models and stochastic optimality theory. *Architectures, Rules, and Preferences: Variations on Themes* by Joan W. Bresnan. Stanford: CSLI, pages 467–479.

Frederick Jelinek. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice, 1980*.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the 11th Language Resources and Evaluation Conference*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Björn Lindblom and Ian Maddieson. 1988. Phonetic universals in consonant systems. *Language, Speech, and Mind*, pages 62–78.

Ian Maddieson. 2006. Correlating phonological complexity: Data and validation. *Linguistic Typology*, 10(1):106–123.

Ian Maddieson. 2009. Calculating phonological complexity, François Pellegrino, Ioana Chitoran, Egidio Marsico, and Christophe Coupe, editors, *Approaches to phonological complexity*, pages 85–110. Mouton de Gruyter, Berlin, Germany.

Ian Maddieson and Sandra Ferrari Disner. 1984. *Patterns of Sounds*, Cambridge University Press.

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

André Martinet. 1955. *Économie des changements phonétiques*, Éditions A. Francke S. A.

John McWhorter. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, 5(2):125–66.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*.

Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Matti Miestamo. 2006. On the feasibility of complexity metrics. In *FinEst Linguistics, Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, pages 11–26.

Matti Miestamo. 2008, Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemaki, and Fred Karlsson, editors, *Language complexity: Typology, contact, change*, pages 23–41. John Benjamins, Amsterdam, The Netherlands.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Steven Moran and Damián Blasi. 2014, Cross-linguistic comparison of complexity measures in phonological systems, Frederick J. Newmeyer and Laurel B. Preston, editors, *Measuring grammatical complexity*, pages 217–240. Oxford University Press, Oxford, UK.

Steven Moran, Daniel McCloy, and Richard Wright, editors . 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Daniel Nettle. 1995. Segmental inventory size, word length, and communicative efficiency. *Linguistics*, 33:359–367.

François Pellegrino, Ioana Chitoran, Egidio Marsico, and Christophe Coupé. 2011. A cross-language perspective on speech information rate. *Language*, 87(3):539–558.

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.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Steven T. Piantadosi, Harry J. Tily, and Edward Gibson. 2009. The communicative lexicon hypothesis. In *The 31st Annual Meeting of the Cognitive Science Society (CogSci09)*, pages 2582–2587.

Uriel Cohen Priva and T. Florian Jaeger. 2018. The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard*, 4(s2).

Ryan K. Shosted. 2006. Correlating complexity: A typological approach. *Linguistic Typology*, 10(1):1–40.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages pages 2951–2959.

R.J.J.H. van Son and Louis C.W. Pols. 2003. Information structure and efficiency in speech production. In *Eighth European Conference on Speech Communication and Technology (Eurospeech)*.

Richard Stanley. 1967. Redundancy rules in phonology. *Language*, pages 393–436.

Holly L. Storkel. 2001. Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, 44(6):1321–1337.

Holly L. Storkel. 2003. Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research*, 46(6):1312–1323.

Holly L. Storkel, Jonna Armbrüster, and Tiffany P. Hogan. 2006. Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, 49(6):1175–1192.

Holly L. Storkel and Jill R. Hoover. 2010. An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken American English. *Behavior Research Methods*, 42(2):497–506.

Holly L. Storkel and Su-Yeon Lee. 2011. The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2):191–211.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Nikolaï Sergeyevich Trubetzkoy. 1938. *Grundzüge der phonologie*, Van den Hoeck & Ruprecht, Gottigen, Germany.

Michael S. Vitevitch and Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3):374–408.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, MIT Press, Cambridge, MA.