# Rethinking Phonotactic Complexity

**Tiago Pimentel**[†], **Brian Roark**[*] **and Ryan Cotterell**[°]
[†]Kunumi and Department of Computer Science, Universidade Federal de Minas Gerais
[*]Google [°]Department of Computer Science, Johns Hopkins University
tiago.pimentel@kunumi.com, roarkbr@gmail.com, ryan.cotterell@jhu.edu

## 1 Introduction and overview

One prevailing view on system-wide phonological complexity is that increases in complexity in one aspect (e.g., phonemic inventory) is offset by reductions in complexity in other aspects (e.g., phonotactics). Underlying this claim – the so-called "compensation hypothesis" (Martinet, 1955; Moran and Blasi, 2014) – is the intuition that languages are generally speaking of roughly equivalent complexity, i.e., no language is overall inherently more complex than others. This has been hypothesized to be the result of natural processes of historical language change, and is sometimes attributed to a potential linguistic universal of equal "communicative capacity" (Pellegrino et al., 2011).

Methods for making such intuitions and hypotheses objectively measurable and/or testable have been of interest for some time, though existing measures are typically relatively coarse. For example, correlations between the size of vowel and consonant inventories have been extensively studied, with mixed results – see, e.g., Moran and Blasi (2014) for a review. Increases in phonemic inventory size are also thought to negatively correlate with word length measured in phonemes. 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). Syllable inventories and syllable-based measures of phonotactic complexity – e.g., highest complexity syllable type in Maddieson (2006) – are also used as variables when looking for evidence of complexity compensation in phonological systems. Even when moving beyond the segment to larger possible structures, however, complexity is generally measured in terms of inventory size. Note, additionally, that by examining negative correlations between word length and inventory size within the context of complexity compensation, the word length in phonemes is also being taken implicitly as a measure of complexity.

In this paper, we take an information theoretic view of phonotactic complexity, and derive a measure that permits straightforward cross-linguistic comparison: bits per phoneme. When given a word, represented as a sequence of phonemic segments, and a statistical model trained on a sample of words from the language, we can measure the bits per phoneme (which, as we know from Brown et al. (1992), is an upper bound on the actual value) and compare across languages. Using a collection of approximately 1000 "basic" concept words across more than a hundred languages, we demonstrate a very high negative correlation between bits per phoneme and the average length of words measured in phonemes. Conventional segmental inventory measures demonstrated relatively poor correlation with word length.

## 2 Data, methods and experiments

### 2.1 NorthEuraLex

We experiment on 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. For the results reported in this paper, we omitted Mandarin, since 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.

We split the data at the concept level. We create 10 random train-dev-test splits where the training portion has 812 concepts, the dev portion has 101 concepts and the test portion has 103 concepts. We then create language-specific sets with the language-specific words for the concept to be rendered.



*Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 365-366.
New York City, New York, January 3-6, 2019

| Measure | Correlation | |
|---|---|---|
| | Pearson $r$ | Spearman $\rho$ |
| 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 1: Pearson and Spearman rank correlation coefficients between complexity measures and average sentence length in phoneme segments.

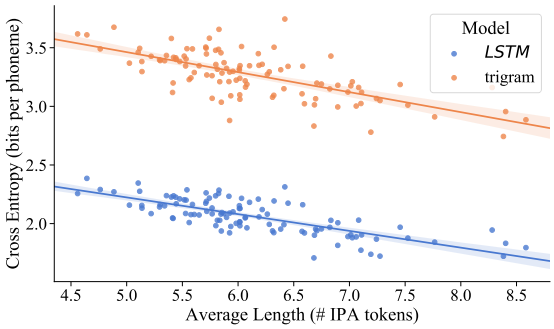


Figure 1: Bits-per-phoneme vs average sentence length under both a trigram and an LSTM language model.

### 2.2 Models

We train three models for measuring bits per phoneme: a unigram model; a smoothed trigram model; and an LSTM language model. In each case, the model is defined over sequences of symbols from an IPA vocabulary (plus end-of-word symbol), and is estimated on the training data and tuned on the development data. Bits-per-phoneme and correlations are calculated on the test set. The unigram model estimates the probability of each phone in the sequence simply as the relative frequency of that phone in the training set. The trigram model is estimated as the deleted interpolation (Jelinek, 1980) of the trigram, bigram and unigram relative frequency estimates, with the mixture parameters estimated on the dev set. Finally, the LSTM language model over IPA symbols is akin to a character-level LSTM, which has been shown to be state of the art for character-level language modeling (Merity et al., 2018). We train the LSTM with an attention mechanism.

### 2.3 Experiments

In addition to measuring bits-per-phoneme with our three models, we also measure the size of the phoneme inventories, as well as the number of vowels and the number of consonants, as additional potential measures of phonological complexity. For each of these variables, we calculated both the Pearson correlation and Spearman rank correlation coefficients, and present them in Table 1.

In addition to the correlations, we plot the trigram and LSTM bits-per-phoneme for each language versus the average length (in IPA tokens) of words in the collection, for each language in Figure 1. The plot demonstrates that the LSTM is, indeed, achieving much lower bits-per-phoneme than the trigram model for every language in the collection, i.e., the increased modeling capacity yields improved models. In addition, however, the improved modeling yields an improved correlation with the length of the string.

## 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.

Johannes Dellert and Gerhard Jäger. 2017. Northeuralex (version 0.9).

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

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

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

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

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

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.