A surprisal–duration trade-off across and within the world's languages

Tiago Pimentel[ð] **Clara Meister**[ʕ] **Elizabeth Salesky**[ɳ] **Simone Teufel**[ð]
Damián Blasi[ə,θ,ɾ] **Ryan Cotterell**[ð,ʕ]
[ð]University of Cambridge [ʕ]ETH Zürich [ɳ]Johns Hopkins University [ə]Harvard University
[θ]Max Planck Institute for the Science of Human History [ɾ]Higher School of Economics
tp472@cam.ac.uk clara.meister@inf.ethz.ch esalesky@jhu.edu
sht25@cl.cam.ac.uk dblasi@fas.harvard.edu ryan.cotterell@inf.ethz.ch

Abstract

While there exist scores of natural languages, each with its unique features and idiosyncrasies, they all share a unifying theme: enabling human communication. We may thus reasonably predict that human cognition shapes how these languages evolve and are used. Assuming that the capacity to process information is roughly constant across human populations, we expect a surprisal–duration trade-off to arise both across and within languages. We analyse this trade-off using a corpus of 600 languages and, after controlling for several potential confounds, we find strong supporting evidence in both settings. Specifically, we find that, on average, phones are produced faster in languages where they are less surprising, and vice versa. Further, we confirm that more surprising phones are longer, on average, in 319 languages out of the 600. We thus conclude that there is strong evidence of a surprisal–duration trade-off in operation, both across and within the world's languages.

1 Introduction

During the course of human evolution, countless languages have evolved, each with unique features. Despite their stark differences, however, it is plausible that shared attributes in human cognition may have placed constraints on how each language is implemented. These constraints, in turn, may lead to compensations and trade-offs in the world's languages. For instance, if we assume a channel capacity (Shannon, 1948) in human's ability to process language (as posited by Frank and Jaeger, 2008), we may make predictions about these trade-offs. Additionally, if we assume this capacity to be uniform across human populations, these trade-offs will extend cross-linguistically.

Within languages, there is a direct connection between this channel capacity assumption and the uniform information density hypothesis (UID; Fenk and Fenk, 1980; Aylett and Turk, 2004; Levy and Jaeger, 2007), which predicts that speakers smooth the information rate in a linguistic signal so as to keep it roughly constant; by smoothing their information rate, natural languages can stay close to a (hypothetical) channel capacity. Across languages, a unified channel capacity allows us to derive a specific instantiation of the compensation hypothesis (Hockett, 1958), with information density (measured in, e.g., bits per phone) being compensated by utterance speed (in, e.g., milliseconds per phone). We may thus predict a trade-off between surprisal[1] and duration both within and across the world's languages.

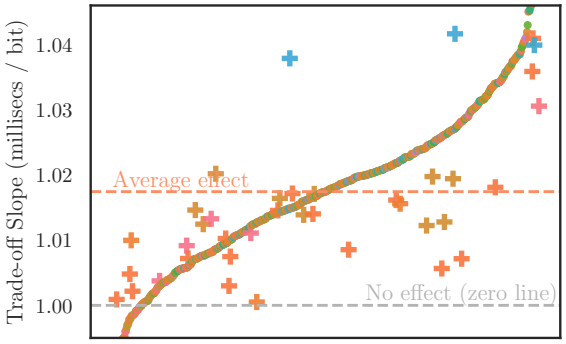


Figure 1: Surprisal–duration trade-off slopes. The y-axis presents a multiplicative effect, duration is multiplied by y per bit of information. Sorted dots represent languages in Unitran; '+' are languages in Epitran.

This trade-off has been studied amply within high resource languages (Genzel and Charniak, 2002; Bell et al., 2003; Mahowald et al., 2018, *inter alia*). Cross-linguistically, however, this trade-off has received comparatively little attention, with a few notable exceptions such as Pellegrino

[1]Surprisal is defined as the negative log-probability of an event, e.g. observing a phone given its prior context.

et al. (2011) and Coupé et al. (2019). Several factors have inhibited cross-linguistic studies of this kind. Arguably, the most prominent is the sheer lack of data necessary to investigate the phenomenon. While massively cross-linguistic data abounds in the form of wordlists (Wichmann et al., 2020; Dellert et al., 2020), surprisal is a context-dependent measure and, therefore, isolated word types are not enough for this analysis. Further, as we have a specific hypothesis for why this trade-off should arise (humans' information processing capacity), we are not interested in simply finding *any* correlation between surprisal and duration. Several confounds could drive such a correlation, but most of these are either trivially true or uninteresting from our perspective. Therefore, a thorough analysis of this trade-off needs to control for these potential confounds.

In this work, we investigate the surprisal–duration trade-off by analysing a massively multi-lingual dataset of more than 600 languages (Salesky et al., 2020). We present an experimental framework, controlling for several possible confounds, and evaluate the surprisal–duration trade-off at the phone level. We find evidence of a trade-off across languages: languages with more surprising phones compensate by making utterances longer. We also confirm mono-lingual trade-offs in 319 languages, out of 600;[2] within these languages, more surprising phones are pronounced with a significantly longer duration. This is the most representative evidence of the uniform information density hypothesis to date. Moreover, we did not find evidence of a single language where the opposite effect is in operation (i.e. where more informative phones are shorter). Given these collective results, we conclude there is strong evidence for a surprisal–duration trade-off both across and within the world's languages.

2 Surprisal and Duration

Cross-linguistic comparisons of information rate go back at least 50 years. In a study comparing phonemes per second, Osser and Peng (1964) found no statistical difference between the speech rate of English and Japanese native speakers. In a similar study, den Os (1985, 1988) compared Dutch and Italian and found no difference in terms of syllables per second, although Italian was found to be somewhat slower in phones per second. Such cross-linguistic comparisons, however, are not straightforward, since the range of speech rate can vary widely within a single language, depending on sentence length (Fonagy and Magdics, 1960) and type of speech (e.g. storytelling vs interview; Kowal et al., 1983). In a meta-analysis of these studies, Roach (1998) concludes that carefully assembled speech databases would be necessary to answer this question. In this line, Pellegrino et al. (2011) recently analysed the speech rate of 8 languages using a semantically controlled corpus. They found strong evidence towards non-uniform speech rates across these languages.

This result is not surprising, however, given that natural languages vary widely in their phonology, morphology, and syntax. Despite these differences, researchers have hypothesised that there exist compensatory relationships between the complexity of these components (Hockett, 1958; Martinet, 1955). For instance, a larger phonemic diversity could be compensated by shorter words (Moran and Blasi, 2014; Pimentel et al., 2020) or a larger number of irregular inflected forms could lead to less complex morphological paradigms (Cotterell et al., 2019). Such a compensation can be thus seen as a type of balance, where languages compromise reliability versus effort in communication (Zipf, 1949; Martinet, 1962). One natural avenue for creating this balance would be a language's information rate. If this were kept roughly constant, the needs of both speakers (who prefer shorter utterances) and listeners (who value easier comprehension) could be accommodated. Speech rate would then be compensated by information density, resulting in a form of surprisal–duration trade-off. Indeed, Pellegrino et al. (2011) and Coupé et al. (2019) present initial evidence of this trade-off across languages.

Analogously, the UID hypothesis posits that, within a language, users balance the amount of information per linguistic unit with the duration of its utterance. This hypothesis has been used to explain a range of experimental data in psycholinguistics, including syntactic reduction (Levy and Jaeger, 2007) and contractions, such as *are* vs *'re* (Frank and Jaeger, 2008). While this theory is somewhat under-specified with respect to its causal mechanisms, as we argue in Meister et al. (2021), one of its typical interpretations is that users are maximising a communicative channel's capacity

[2]The original number of languages was 635, but after removing those with quality issues, we end up with 600. This process is explained in §4.

(Frank and Jaeger, 2008; Piantadosi et al., 2011). If we assume this channel's capacity to be constant across languages, we may derive a cross-linguistic version of UID. Such a hypothesis would predict, for instance, that speakers of languages with less informative phones will make them faster. Under this specific interpretation, our study can be seen as evidence of UID as a cross-linguistic phenomenon.

3 Measuring Surprisal

To formalise our approach, we first present a standard measure of information content: **surprisal**. In the context of natural language, surprisal (Hale, 2001) measures the Shannon information content a linguistic unit conveys in context, which can be measured as its negative log-probability:

$$\mathrm{H}(S_t = s_t \mid \boldsymbol{S}_{<t} = \boldsymbol{s}_{<t}) = -\log p(s_t \mid \boldsymbol{s}_{<t}) \tag{1}$$

In this equation, $\boldsymbol{S}$ is a sentence-level random variable, with instances $\boldsymbol{s} \in \mathcal{S}^*$, and t indexes a position in the sentence. Accordingly, we define $\mathcal{S}$ as the set of phones in a given phonetic alphabet, and we use $\boldsymbol{s}_{<t}$ to indicate the context in which phone s_t appears.

Unfortunately, this surprisal is not readily available, since we would need access to the true distribution $p(s_t \mid \boldsymbol{s}_{<t})$ to compute it. We will use an approximation $p_{\boldsymbol{\theta}}(s_t \mid \boldsymbol{s}_{<t})$ instead, i.e. a phone-level model with estimated parameters $\boldsymbol{\theta}$.

3.1 Approximating $p(s_t \mid \boldsymbol{s}_{<t})$.

While much of the original psycholinguistic work on surprisal estimated $p_{\boldsymbol{\theta}}$ using n-gram models (Levy and Jaeger, 2007; Coupé et al., 2019, *inter alia*), recent work has shown that a language model's psychometric predictive power correlates directly with its quality, measured by its cross-entropy in held-out data (Goodkind and Bicknell, 2018; Wilcox et al., 2020). We will thus make use of LSTMs in this work, since they have been shown to outperform n-grams on phone-level language modelling tasks (Pimentel et al., 2020). We first encode each phone s_t into a high-dimensional lookup embedding $\mathbf{e}_t \in \mathbb{R}^{d_1}$, where d_1 is its embedding size. We then process these embeddings using an LSTM (Hochreiter and Schmidhuber, 1997), which outputs contextualised hidden state vectors:

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{h}_{t-1}, \mathbf{e}_{t-1}) \in \mathbb{R}^{d_2} \tag{2}$$

where the initial hidden state $\mathbf{h}_0$ is the zero vector and the initial phone s_0 is a start-of-sentence symbol. The hidden states are then linearly transformed and projected onto $\Delta^{|\mathcal{S}|+1}$, the probability simplex, via a softmax to compute the desired distribution:[3]

$$p_{\boldsymbol{\theta}}(s_t \mid \boldsymbol{s}_{<t}) = \mathrm{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b}) \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{(|\mathcal{S}|+1)\times d_2}$ and $\mathbf{b} \in \mathbb{R}^{(|\mathcal{S}|+1)}$ are learnable parameters. We optimise the parameters by minimising our model's cross-entropy with a training set, which corresponds to minimising the following objective

$$\mathrm{H}_{\boldsymbol{\theta}}(S_t \mid \boldsymbol{S}_{<t}) = -\sum_{n=1}^{N}\sum_{t=1}^{|\boldsymbol{s}^{(n)}|} \log p_{\boldsymbol{\theta}}(s_t^{(n)} \mid \boldsymbol{s}_{<t}^{(n)}) \tag{4}$$

where we assume $\{\boldsymbol{s}^{(n)}\}_{n=1}^{N}$ are sampled from the true distribution $p(\cdot)$.

To avoid overfitting to this training set, we then estimate the cross-entropy with a validation set $\{\widehat{\boldsymbol{s}}^{(m)}\}_{m=1}^{M}$, where we stop training once this validation cross-entropy stops decreasing. Note that minimising the cross-entropy is equivalent to minimising the Kullback–Leibler divergence between two distributions. Further, if we have access to a larger number of samples M, we assume this cross-entropy estimate will give us a tight approximation to the cross-entropy between $p_{\boldsymbol{\theta}}$ and the true distribution. Thus, the lower this cross-entropy, the closer we may assume our model is to the true $p(\cdot)$, and the better we should expect our surprisal estimates to be.

Hyper-parameter choices. We implement our phone-level LSTM language models with two hidden layers, an embedding size of 64 and a hidden size of 128. We further use a dropout of 0.5 and a batch size of 64. We train our phone-level LSTM models using AdamW (Loshchilov and Hutter, 2019) with its default hyper-parameters in PyTorch (Paszke et al., 2019). We evaluate our models on a validation set every 100 batches, stopping training when we see no improvement for five consecutive evaluations. We split each language's data (described in §4) into train-dev-test sets using an 80-10-10 split, using sentences as our delimiters. We thus do not separate phone data points from the same sentence. We use the first two splits to train and validate our models, while the test set is held out and used throughout our analysis.

[3] The dimension of the probability simplex is $|\mathcal{S}| + 1$ to account for an end-of-sentence symbol.

4 Data

We use the VoxClamantis dataset for our analysis (Salesky et al., 2020). This dataset is derived from spoken readings of the Bible[4] and spans more than 600 languages from 70 language families, as shown in Fig. 2.[5] This dataset offers us a semantically controlled setting for our experiments, as it is composed of translations of a single text, the Bible.

This dataset contains automatically generated phone alignments and derived phonetic measures for all its languages (with both phone duration, and vowels' first and second formant frequencies). On average, there are approximately 9,000 utterances (or 20 hours of speech) per language, making it the largest dataset of its kind. Phone labels were generated using grapheme-to-phoneme (G2P) tools and time aligned using either multilingual acoustic models (Wiesner et al., 2019; Povey et al., 2011) or language-specific acoustic models (Black, 2019; Anumanchipalli et al., 2011). VoxClamantis offers its phonetic measurements under three G2P models, which trade-off language coverage and quality. We will focus on two:[6]

- **Epitran (Mortensen et al., 2018).** This is a collection of high quality G2P models based on language-specific rules. Phonetic measurements produced with Epitran are available for a collection of 39 doculects[7] from 29 languages (as defined by ISO codes) in 8 language families.

- **Unitran (Qian et al., 2010).** This is a naïve and deterministic G2P model, but its derived measurements are available for all languages in VoxClamantis. While Unitran is particularly error-prone for languages with opaque orthographies (Salesky et al., 2020), we filter out the languages with lower-quality alignments (as we detail below). The original dataset has 690 doculects from 635 languages in 70 language families.

In order to study the trade-off hypothesis we require two measurements: phone durations and phone-level surprisals. As mentioned above, phone durations are readily available in VoxClamantis. Phone-level surprisals, on the other hand, are not, so we employed phone-level language models in order to estimate them (as detailed in §3). Given both these values, we can now perform our cross-linguistic analysis. First, though, we will describe some data quality checks.

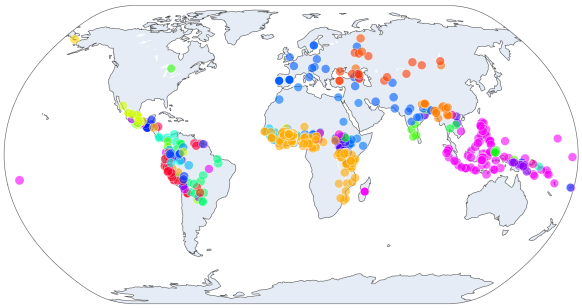


Figure 2: The languages of the VoxClamantis corpus geo-located and coloured by language family.

Filtering Unitran. The phone and utterance alignments for the VoxClamantis dataset were automatically generated and may be noisy due to both of these processes. The labels from the Unitran G2P also contain inherent noise due to their deterministic nature. Accordingly, we filter the data using the mean Mel-Cepstral Distortion (MCD) as an implicit quality measure for the alignments. MCD is an edit-distance metric which evaluates the distance between some reference speech and speech synthesised using the alignments (Kubichek, 1993). We use the utterance-level MCD scores from the CMU Wilderness dataset (Black, 2019), removing all utterances with an MCD score higher than 7. This leaves us with 647 doculects from 600 languages in 69 language families.

5 Design Choices

There are several critical design choices that must be made when performing a cross-linguistic analysis of this nature. While some may at first seem inconsequential, they can have a large impact on down-stream results. Specifically, we assume that there is a surprisal–duration trade-off which is caused by a capacity to process information, which should be roughly constant across human populations. We must thus control for other potential sources for this trade-off, which we deem to be uninteresting in this work.

Phone-level Analysis. While there are good reasons for performing this analysis at the syllable- or

[4]These texts were crawled from `bible.is` and utterance-aligned by Black (2019) for the CMU Wilderness dataset.

[5]A list of all languages can be found in App. C.

[6]We set Wikipron (Lee et al., 2020) alignments aside because we could not obtain word position information for them.

[7]The term doculect refers to a dialect as recorded in a specific document, in this case a Bible reading.

word-level, we believe phones are advantageous for out study. Greenberg (1999), for instance, shows syllables are less prone than phones to be completely deleted in casual speech; syllables would thus allow more robust estimates of speech duration. Nonetheless, languages that allow for more complex (and long) syllabic structures will naturally have more valid syllables. A larger number of syllables, in turn, will cause each syllable to be less predictable on average.[8] Therefore, more complex syllables will be both longer and unpredictable. Studying syllables can thus lead to trivial trade-offs which mainly reflect the methodology employed. A similar argument can be made against word-level analyses.[9] Performing this type of analysis at the phone-level should alleviate this effect, making it the more appropriate choice.

Articulatory Costs. Whereas the range of effort used to produce individual phones may be smaller than in other linguistic hierarchies, there is still a considerable variation in the cost associated with each phone's articulation. For instance, Zipf (1935) argued that a phone's articulatory effort was related to its frequency. If this is indeed the case, a direct analysis of surprisal–duration pairs that does not control for articulatory effort could also lead to a trivial trade-off: the long and effortful phones will be less frequent and likely to be more unpredictable, having higher surprisals. To account for each phone's articulatory cost, we use mixed effects models in our analysis, and include phone identity as random intercept effects.

Word-initial and Word-final Lengthening. There is ample evidence showing that, across languages, word-initial and word-final segments are lengthened during production (Fougeron and Keating, 1997; White et al., 2020). Another property is that word-initial positions carry more information than word-final ones, which has been well-studied in both psycholinguistics and information-theory. From a psycholiguistic perspective, it seems word-initial segments are more important for word recognition (Bagley, 1900; Fay and Cutler, 1977; Bruner and O'Dowd, 1958; Nooteboom, 1981). Under an information-theoretic analysis, it has been observed that earlier segments in a word are more surprising than later ones (van Son and Pols, 2003; King and Wedel, 2020; Pimentel et al., 2021). Word-initial segments are both lengthened and more surprising, potentially for unrelated reasons. An analysis which does not control for such word-positioning is thus doomed to find trivial correlations. To account for this word-initial and word-final lengthening, we include three word position fixed effects (initial, middle, or final) in our mixed effects models.

Sentential Context. The amount of context that a model conditions on when estimating probabilities will undoubtedly have an impact on a study of this nature. For example, a model that cannot look back beyond the current word, such as the one employed by Coupé et al. (2019), can by definition only condition on the previous phones in the same word. Arguably, a cognitively motivated surprisal–duration trade-off should estimate surprisal using a phone's entire sentential context and not only the prior context inside a specific word. In this work, we make use of LSTMs (as described in §3), which can model long context dependencies (Khandelwal et al., 2018).

6 Generalised Mixed Effects

Throughout our experiments, we will use mixed-effects models; we provide a brief introduction here (see Wood (2017) for a longer exposition). Classical linear regressions models can be written as:

$$y_i = \boldsymbol{\phi}^\intercal \mathbf{x}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2_{\texttt{err}}) \tag{5}$$

where y_i is the target variable, $\mathbf{x}_i \in \mathbb{R}^d$ is the model's input and $\boldsymbol{\phi} \in \mathbb{R}^d$ a learned weight vector. Further, the error (or unexplained variance) term ϵ_i is assumed to be normally distributed and independent and identically distributed (i.i.d.) across data instances. Such an i.i.d. assumption, however, may not hold. In our analysis, for instance, multiple phones come from each of our analysed languages; it is thus expected that such co-language phones share dependencies in how their ϵ_i are sampled. Mixed-effects models allow us to model such dependencies through the use of random effects. Formally, for an instance $\mathbf{x}_i$ from a specific language ℓ_i, we model:

$$y_i = \boldsymbol{\phi}^\intercal \mathbf{x}_i + \omega_{\ell_i} + \epsilon_i, \quad \begin{matrix} \omega_{\ell_i} \sim \mathcal{N}(0, \sigma^2_\omega) \\ \epsilon_i \sim \mathcal{N}(0, \sigma^2_{\texttt{err}}) \end{matrix} \tag{6}$$

[8] Probabilities must sum to 1. This finite probability mass means average probability must go down with more classes.

[9] Concatenative languages, for instance, would have both longer and less predictable words. Take the German word *Hauptbahnhof* which can be translated into English as *central train station*. Predicting this single (and long) German word is equivalent to predicting three words in English.

where ω_{ℓ_i} is a random effect and ϕ is now termed a fixed effect. Here, ω_{ℓ_i} is an intercept term which is assumed to be shared across all instances of language ℓ_i, and σ^2_ω is directly learned from the data. Similarly, we can add random slope effects:

$$y_i = \boldsymbol{\phi}^\intercal \mathbf{x}_i + \boldsymbol{\beta}_{\ell_i}^\intercal \mathbf{x}_i + \omega_{\ell_i} + \epsilon_i, \quad \begin{array}{l} \boldsymbol{\beta}_{\ell_i} \sim \mathcal{N}(0, \Sigma_{\boldsymbol{\beta}}) \\ \omega_{\ell_i} \sim \mathcal{N}(0, \sigma^2_\omega) \\ \epsilon_i \sim \mathcal{N}(0, \sigma^2_{\texttt{err}}) \end{array} \tag{7}$$

where each $\boldsymbol{\beta}_{\ell_i} \in \mathbb{R}^d$ is a language-specific random slope and $\Sigma_{\boldsymbol{\beta}}$ is a (learned) covariance matrix. Furthermore, our assumption that error terms are normally distributed may not hold in this setting. Phone durations, for instance, cannot be negative and are positively skewed, making a log-linear model more appropriate:

$$\log(y_i) = \boldsymbol{\phi}^\intercal \mathbf{x}_i + \boldsymbol{\beta}_{\ell_i}^\intercal \mathbf{x}_i + \omega_{\ell_i} + \epsilon_i \tag{8}$$

where $\boldsymbol{\beta}_{\ell_i}$, ω_{ℓ_i}, and ϵ_i are still distributed as in eq. (7). This is similar to modelling the original ϵ_i terms as coming from a log-normal distribution. We note though, that under this model our effects become multiplicative (as opposed to additive): an increase of δ unit in the right side will make the value of y_i be multiplied by e^δ. We will use `lme4`'s (Bates et al., 2015) notation to represent these models. Under this notation, a parenthesis represents a random effect and parameters are left out. We thus re-write eq. (8) as:

$$\log(y) = 1 + \mathbf{x} + (1 + \mathbf{x} \mid \texttt{language}) \tag{9}$$

7 Experiments and Results[10]

In this section, we will first analyse the surprisal–duration trade-off in individual languages. We will then perform an analysis with our full data, studying the trade-off both within and across languages with a single model. Finally, in our last experiment we will average phone information per language to analyse a purely cross-linguistic trade-off.

7.1 Individual Language Analyses

We first analyse languages individually, verifying if more surprising phones have on average a longer duration. With this in mind, we estimate a generalised mixed effects model for each language. We control for each phone's articulatory costs by adding phone identity as a random effect. Additionally, we include fixed effects to control for word position effects, adding separate intercepts for word-initial and word-final positions. Finally, we consider a fixed effect relating surprisal and word positions. At word-initial positions, for instance, the connection between surprisal and duration could potentially be stronger or weaker.[11] This leaves us with the following relationship:

$$\log(\texttt{duration}) = 1 + \texttt{surprisal} + \texttt{position} + \texttt{surprisal} \cdot \texttt{position} + (1 \mid \texttt{phone}) \tag{10}$$

In this parametrisation, a trade-off between surprisal and duration will emerge as a positive and significant `surprisal` slope. Analogously, an inverse trade-off will emerge as a negative and significant slope, since we use two-tailed statistical tests.[12] Out of the 39 doculects in Epitran, 30 present statistically significant positive slopes ($^{23}/_{29}$ languages, and $^{8}/_{8}$ families; meaning at least one language showed a significant effect per family). On Unitran (which we recall is a noisier dataset), $^{326}/_{647}$ doculects presented significantly positive slopes ($^{319}/_{600}$ languages, and $^{53}/_{69}$ families). Additionally, we find no language in either dataset with significantly negative slopes: we either find evidence for the trade-off or we have no association whatsoever.

The trade-off strength, as measured by the surprisal–duration slopes, can be seen in Fig. 1 (on first page) and Fig. 3. As noted above, by

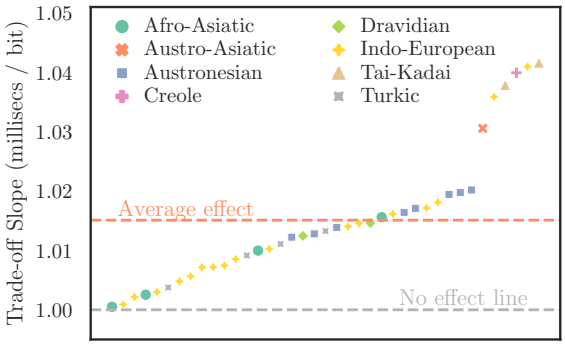


Figure 3: Language-specific trade-off slopes in Epitran from the mixed effects model in eq. (10). The y-axis represents a multiplicative effect, duration is multiplied by y per extra bit of phone information.

[10] Our code is available at https://github.com/rycolab/surprisal-duration-tradeoff.

[11] We analyse the impact of both these effects, phone identity and word position, in App. A.

[12] Statistical significance was assessed under a confidence level of $\alpha < 0.01$ and we used Benjamini and Hochberg (1995) corrections for multiple tests whenever necessary.

predicting a linear change in logarithmic scale, our effects become multiplicative instead of additive. The average multiplicative slope we get across all the analysed languages in both datasets is roughly 1.02, meaning that each added bit of information multiplies duration by 1.02. We believe this should serve as strong support for our hypothesis of a trade-off within languages. Moreover, to the best of our knowledge, this is the most representative study of the UID hypothesis to date, as measured by the number and typological diversity of analysed languages.

7.2 Aggregated Cross-linguistic Analysis

Following the previous study, we now run a cross-linguistic analysis by aggregating all the languages within a single model. We add the same controls as before, but further nest the phone random effects per language (meaning we create one random effect per phone–language pair). We also include random language-specific intercepts and slopes. Formally,

$$\begin{aligned}\log(\texttt{duration}) = 1 &+ \texttt{surprisal} + \texttt{position} \\ &+ \texttt{surprisal} \cdot \texttt{position} \\ &+ (1 + \texttt{surprisal} + \texttt{surprisal} \cdot \texttt{position} \\ &\qquad \mid \texttt{language}) \\ &+ (1 \mid \texttt{language} : \texttt{phone}) \end{aligned} \tag{11}$$

After estimating this generalised mixed effects model, we find statistically significant cross-linguistic trade-off effects in both datasets. The multiplicative slope is roughly 1.02 in both datasets, again meaning each extra bit of information multiplies the duration by this value ($\phi = 1.023$ in Unitran and $\phi = 1.015$ in Epitran).[13] We further analyse the per-language trade-off slopes, which can be seen in Fig. 4. These language-specific slopes are calculated by summing the fixed effect of the surprisal term with its random effects per language. We see a similar trend in this figure as in Fig. 3, with most of the analysed doculects having a positive surprisal–duration trade-off.

7.3 Cross-linguistic Trade-offs

Our previous experiment in §7.2 makes use of language-specific random effects. These effects allow the model to potentially represent *within*-language trade-off effects, while correcting for cross-linguistic differences by using the model parameters. It therefore cannot serve as confirmation of a trade-off across the world's languages by itself, only as additional evidence for it. In this section, we do not use language-specific random effects; instead, we average surprisal within a language for each phone–position tuple. We then train the following mixed effects model:

$$\begin{aligned}\texttt{duration} = 1 &+ \texttt{surprisal} + \texttt{position} \\ &+ \texttt{surprisal} \cdot \texttt{position} + (1 \mid \texttt{phone}) \end{aligned} \tag{12}$$

This equation is identical to the one in eq. (10), but now we model the language–phone–position tuples, instead of a language's individual phones.[14] Additionally, since we are aggregating results per tuple for this analysis, the central limit theorem tells us our model's residuals should be roughly Gaussian. We thus use linear mixed effects models, instead of the generalised log-linear ones. By analysing this model we find a significantly positive surprisal–duration additive slope, of $\phi = 1.5$ milliseconds per bit ($\phi = 1.54$ in Unitran and $\phi = 1.52$ in Epitran). This confirms the expected cross-linguistic trade-off: languages with more surprising phones really have longer durations, even after controlling for word positions and phone-specific articulatory costs.

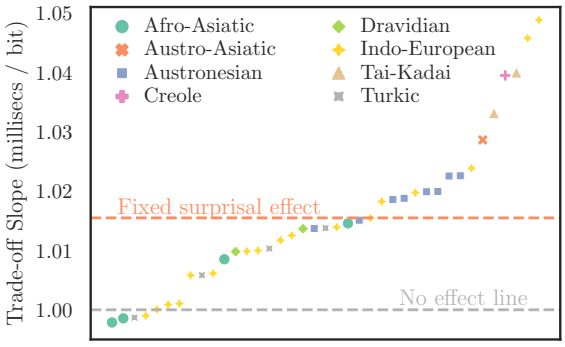


Figure 4: Language-specific trade-off slopes in Epitran from the mixed effects model in eq. (11). The y-axis represents a multiplicative effect, duration is multiplied by y per extra bit of phone information.

8 Discussion

The pressure towards a specific information rate (potentially set at a specific cognitive channel ca-

[13] For the Epitran data, we performed this analysis while also adding language family effects and found similar results. However, we could not repeat this experiment for Unitran as the model was too memory intensive.

[14] We note that phone labels may not always align exactly across languages here, due to possible differences between VoxClamantis' G2P label sets. This may introduce noise into this analysis. It is reassuring, though, that the previous analyses with phones as language-specific effects lead to similar conclusions.

pacity) has been posited as an invariant across languages. Directly testing such a claim is perhaps impossible, as data alone cannot prove its universality. Moreover, providing meaningful evidence towards this phenomenon requires a careful and comprehensive cross-linguistic analysis, which we attempt to perform in this work. In comparison to similar studies, such as those by Pellegrino et al. (2011) and Coupé et al. (2019), we employ more sophisticated techniques to measure a linguistic unit's (in our case, a phone's) information content. Moreover, we also employ more rigorous strategies for analysing the surprisal–duration relationship, controlling for several potential confounds. By introducing these improvements, we attain a more detailed understanding of the role of information in language production, both across and within languages.

Experimentally, we find that, after controlling for other artefacts, the information conveyed by a phone in context has a modest but significant relationship with phone duration. We see that this relationship is consistently positive across a number of investigated settings, despite being small in magnitude, meaning that more informative phones are on average longer. Additionally, using two-tailed tests at $\alpha < 0.01$ throughout our experiments, we find no language with a significant negative relationship between phone surprisal and duration.

Limitations and Future Work. In this work, we implemented a careful evaluation protocol to study the relationship between a phone's surprisal and duration in a representative set of languages. To perform our study in such a large number of languages, however, we rely on the automatically aligned phone measurements from VoxClamantis, which contain noise from various sources. Future work could investigate if biases in the dataset generation protocol could impact our results. Further, VoxClamantis data is derived from readings of the Bible. Future studies could extend our analysis to other settings, such as conversational data.

9 Conclusion

In this work, we have provided the widest cross-linguistic investigation of phone surprisal and duration to date, covering 600 languages from over 60 language families spread across the globe. We confirm a surprisal–duration trade-off both across these analysed languages and within a subset of 319 of them, covering 53 language families. While there exist arguments against some of our design choices, our overarching conclusion is remarkably consistent across our analyses: the presence of a surprisal–duration trade-off is significant in language production. In other words, both across and within languages, phones carrying more information are longer, while phones carrying less information are produced faster.

References

Gopala Krishna Anumanchipalli, Kishore Prahallad, and Alan W Black. 2011. Festvox: Tools for creation and analyses of large speech corpora. In *Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia*.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

William Chandler Bagley. 1900. The apperception of the spoken sentence: A study in the psychology of language. *The American Journal of Psychology*, 12(1):80–130.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Jerome S. Bruner and Donald O'Dowd. 1958. A note on the informativeness of parts of words. *Language and Speech*, 1(2):98–101.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

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).

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.

David Fay and Anne Cutler. 1977. Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, 8(3):505–520.

August Fenk and Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß? *Zeitschrift für Experimentelle und Angewandte Psychologie*, 27(3):400–414.

Ivan Fonagy and Klara Magdics. 1960. Speed of utterance in phrases of different lengths. *Language and Speech*, 3(4):179–192.

Cécile Fougeron and Patricia A. Keating. 1997. Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, 101(6):3728–3740.

Austin F. Frank and T. Florain Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Steven Greenberg. 1999. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2-4):159–176.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

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

Charles Francis Hockett. 1958. *A Course in Modern Linguistics*. Macmillan, New York.

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.

Adam King and Andrew Wedel. 2020. Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind*, pages 1–12.

Sabine Kowal, Richard Wiese, and Daniel C. O'Connell. 1983. The use of time in storytelling. *Language and Speech*, 26(4):377–392.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

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 mining with WikiPron. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA). Resources downloadable from `https://github.com/kylebgorman/wikipron`.

Roger P. Levy and Tim Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, pages 849–856.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

André Martinet. 1962. *A Functional View of Language*, volume 196. Clarendon Press.

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*. Association for Computational Linguistics.

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.

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 (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Sieb G. Nooteboom. 1981. Lexical retrieval from fragments of spoken words: Beginnings vs endings. *Journal of Phonetics*, 9(4):407–424.

Els den Os. 1985. Perception of speech rate of Dutch and Italian utterances. *Phonetica*, 42(2-3):124–134.

Els den Os. 1988. *Rhythm and tempo of Dutch and Italian: A contrastive study*. Dr. Elinkwijk.

Harry Osser and Frederick Peng. 1964. A cross cultural study of speech rate. *Language and Speech*, 7(2):120–125.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8024–8035. Curran Associates, Inc.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, pages 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.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 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.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Ting Qian, Kristy Hollingshead, Su-youn Yoon, Kyoung-young Kim, and Richard Sproat. 2010. A Python toolkit for universal transliteration. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Peter Roach. 1998. Some languages are spoken more quickly than others. In Laurie Bauer and Peter Trudgill, editors, *Language myths*, pages 150–8. Penguin Books.

Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner. 2020. A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Rob J. J. H. van Son and Louis C.W. Pols. 2003. How efficient is speech? In *Proceedings of the Institute of Phonetic Sciences*, volume 25, pages 171–184.

Laurence White, Silvia Benavides-Varela, and Katalin Mády. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics*, 81:100982.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2020. The ASJP database (version 19). Accessed on April 2, 2020.

Matthew Wiesner, Oliver Adams, David Yarowsky, Jan Trmal, and Sanjeev Khudanpur. 2019. Zero-shot pronunciation lexicons for cross-language acoustic model transfer. In *Proceedings of IEEE Association for Automatic Speech Recognition and Understanding (ASRU)*.

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Meeting of the Cognitive Science Society*, page 1707–1713.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. MIT Press, Cambridge.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge.

A Confound analysis

In this section, we analyse the word position and articulatory cost confounds mentioned in §5, as they could have an impact on a surprisal–duration trade-off analysis. We first investigate the parameters from our mixed effects models containing word positioning effects. The word-initial and word-final intercepts are significantly positive in all 39 languages of our mono-lingual Epitran analysis (represented by eq. (10)) and in both cross-linguistic experiments (eqs. (11) and (12)). The intercepts for word-initial positions average at 67 milliseconds, while the word-final ones average at 32, providing new evidence for this word boundary lengthening effect. Since word position is correlated with surprisal, this boundary lengthening phenomenon could pose as a source of bias in our results, had we not controlled for it.

We now explore the potential bias introduced by phone-specific articulatory costs. As mentioned in §5, languages with larger phonetic inventory sizes may be more inclined to use marked phones, which have longer duration. While this correlation between inventory size and unit cost would be particularly problematic for larger linguistic units (e.g. syllables) it can also affect our phone-level analysis. In fact, we take the Spearman correlation between a language's inventory size (in number of unique phones) and its average phone duration, finding a positive correlation of $\rho = .28$. The average surprisal–duration Spearman correlation across languages is $\rho = 0.45$. As inventory size and surprisal are strongly correlated across languages, we find that pure inventory effects may be driving a large part of the analysed correlation.

To analyse how strongly both confounds would reflect in the main effect if left unaccounted for, we rerun our previous analyses, but without effects for either `position`, phone, or both. We do so for Epitran only. The resulting estimated trade-off effects are given in Tab. 1. We indeed see that these confounds are typically absorbed by the fixed surprisal effect in all three settings. Notably, without confound control we would find supposedly significant results in all analysed languages, and a 10 times stronger cross-linguistic effect, all of which are in fact spurious.

Controls		Trade-off Slope ϕ			
		Mono-lingual		Cross-linguistic	
Phone	Position	eq. (10)	# Sign	eq. (11)	eq. (12)
✓	✓	1.02	30	$1.02^{\ddagger}$	$1.52^{\ddagger}$
✓	✗	1.02	37	$1.03^{\ddagger}$	$0.93^{\ddagger}$
✗	✓	1.03	33	$1.02^{\ddagger}$	15.75
✗	✗	1.04	39	$1.04^{\ddagger}$	$15.73^{\dagger}$

Table 1: Comparison of trade-off (in milliseconds per bit) found when not conditioning on potential confounds. # Sign represents the number of significant languages ($\alpha < 0.01$) in a mono-lingual analysis.

B Additive Effects

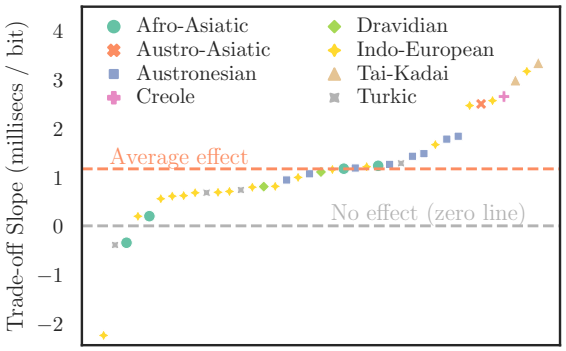


Figure 5: Additive slope of the model in eq. (10).

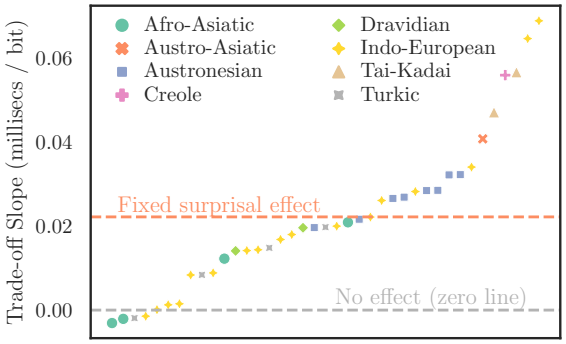
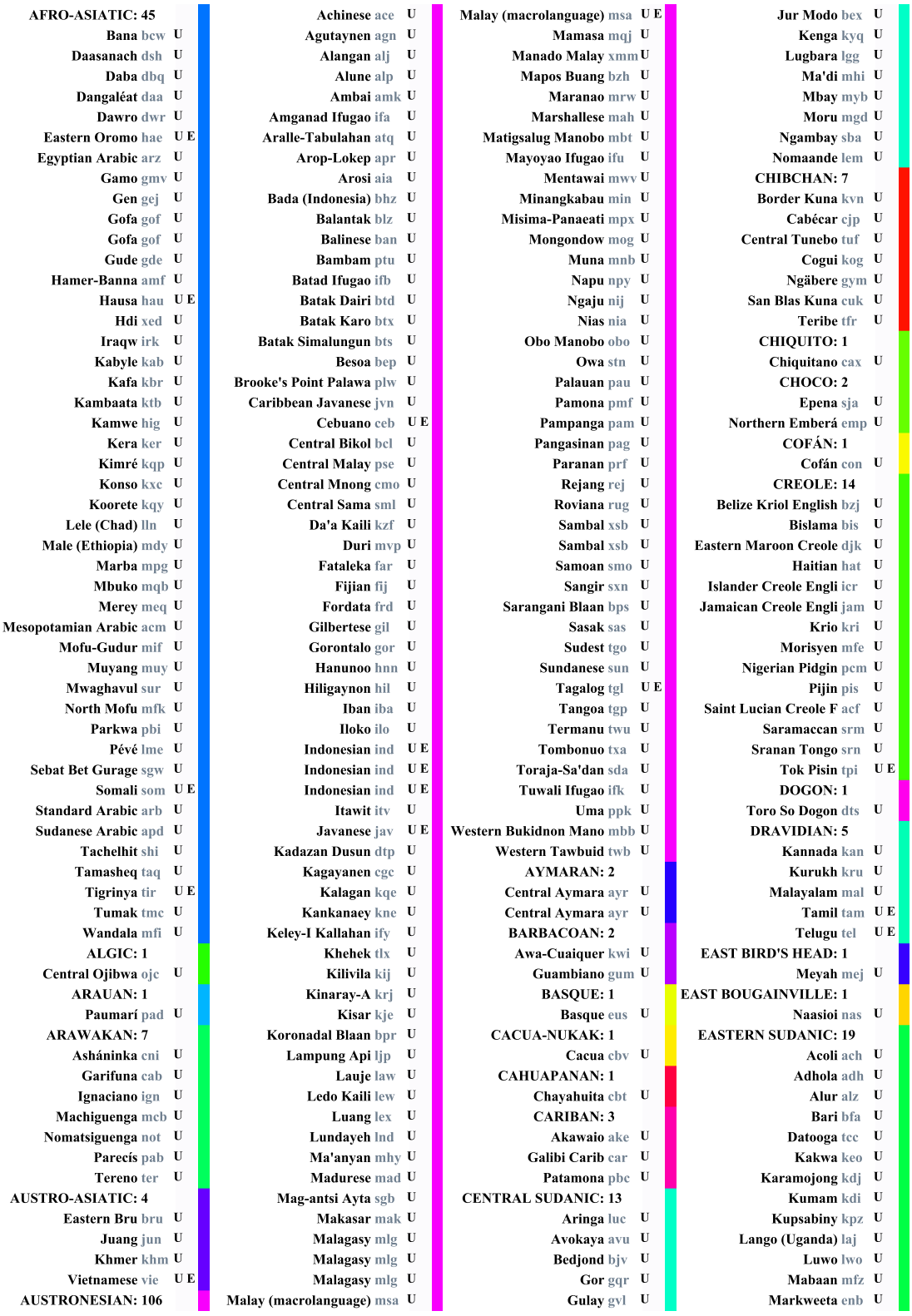


Figure 6: Additive slope of the model in eq. (11).

C Languages

The languages used in our analyses are listed below, grouped by language family, along with their three character ISO 639-3 code, and the grapheme-to-phoneme schemes for which phone alignments are available for that language in the VoxClamantis dataset – Unitran: **U**, Epitran: **E** (Salesky et al., 2020). ISO codes for which there are multiple languages listed may represent dialects or other sub-language variations and/or multiple available Bible versions for which data is available.



Murle mur U
Nuer nus U
Sabaot spy U
Shilluk shk U
Southwestern Dinka dik U
Teso teo U
ESKIMO-ALEUT: 1
Central Siberian Yupi ess U
GUAHIBAN: 3
Cuiba cui U
Guahibo guh U
Guayabero guo U
GUAICURUAN: 1
Toba tob U
HMONG-MIEN: 1
Hmong Daw mwwU
HUAVEAN: 1
San Mateo Del Mar Hua huv U
HUITOTOAN: 3
Bora boa U
Minica Huitoto hto U
Murui Huitoto huu U
INDO-EUROPEAN: 40
Albanian sqi U
Awadhi awa U
Bengali ben U E
Bengali ben U E
Bengali ben U E
Caribbean Hindustani hns U
Chhattisgarhi hne U
Dari prs U
English eng U
Fiji Hindi hif U
French fra U
French fra U
Hindi hin U E
Iranian Persian pes U
Latin lat U
Magahi mag U
Maithili mai U
Malvi mup U
Marathi mar U E
Northern Kurdish kmr U E
Oriya (macrolanguage) ori U
Ossetian oss U
Polish pol U E
Portuguese por U
Portuguese por U
Portuguese por U
Portuguese por U
Romanian ron U E
Russian rus U E
Sinte Romani rmo U
Spanish spa U E
Spanish spa U E
Spanish spa U E
Spanish spa U E
Spanish spa U E
Swedish swe U E
Swedish swe U E
Tajik tgk U E
Urdu urd U
Vlax Romani rmy U
JIVAROAN: 4
Achuar-Shiwiar acu U
Aguaruna agr U
Huambisa hub U
Shuar jiv U
KHOE-KWADI: 1
Southern Samo sbd U
KOMAN: 1
Uduk udu U
KORDOFANIAN: 1
Moro mor U
LOWER SEPIK-RAMU: 1
Aruamu msy U
MACRO-GE: 1
Kayapó txu U
MANDE: 13
Bambara bam U
Bissa bib U
Boko (Benin) bqc U
Busa bqp U
Dyula dyu U
Dyula dyu U
Kuranko knk U
Loko lok U
Mandinka mnk U
Mende (Sierra Leone) men U
Northern Bobo Madaré bbo U
Susu sus U
Xaasongaxango kao U
MASCOIAN: 1
Enxet enx U
MATACOAN: 1
Maca mca U
MAYAN: 42
Achi acr U
Aguacateco agu U
Chol ctu U
Chortí caa U
Chuj cac U
Chuj cac U
Huastec hus U
Ixil ixl U
Ixil ixl U
Ixil ixl U
K'iche' quc U
K'iche' quc U
K'iche' quc U
K'iche' quc U
K'iche' quc U
Kaqchikel cak U
Kaqchikel cak U
Kaqchikel cak U
Kaqchikel cak U
Kaqchikel cak U
Kekchí kek U
Kekchí kek U
Mam mam U
Mam mam U
Mam mam U
Mam mam U
Mopán Maya mop U
Popti' jac U
Popti' jac U
Poqomchi' poh U
Poqomchi' poh U
Q'anjob'al kjb U
Tektiteko ttc U
Tz'utujil tzj U
Tzeltal tzh U
Tzeltal tzh U
Tzotzil tzo U
Tzotzil tzo U
Western Kanjobal knj U
Yucateco yua U
MISUMALPAN: 1
Mískito miq U
MIXE-ZOQUE: 3
Coatlán Mixe mco U
Highland Popoluca poi U
Quetzaltepec Mixe pxm U
MONGOLIC: 2
Halh Mongolian khk U
Kalmyk xal U
NAKH-DAGHESTAN.: 2
Avaric ava U
Chechen che U
NIGER-CONGO: 159
Abidji abi U
Adele ade U
Adioukrou adj U
Akan aka U
Akebu keu U
Akoose bss U
Anufo cko U
Avatime avn U
Bafut bfd U
Bandial bqj U
Bekwarra bkv U
Bete-Bendi btt U
Biali beh U
Bimoba bim U
Bokobaru bus U
Bomu bmq U
Buamu box U
Buli (Ghana) bwu U
Bum bmv U
Cameroon Mambila mcu U
Central-Eastern Niger fuq U
Cerma cme U
Cerma cme U
Chopi cce U
Chumburung ncu U
Delo ntr U
Denya anv U
Ditammari tbz U
Djimini Senoufo dyi U
Duruma dug U
Eastern Karaboro xrb U
Ekajuk eka U
Ewe ewe U
Ewe ewe U
Farefare gur U
Farefare gur U
Fon fon U
Gikyode acd U
Giryama nyf U
Gitonga toh U
Gogo gog U
Gokana gkn U
Gourmanchéma gux U
Gwere gwr U
Hanga hag U
Haya hay U
Ifè ife U
Ivbie North-Okpela-Ar atg U
Izere izr U
Jola-Fonyi dyo U
Jola-Kasa csk U
Jukun Takum jbu U
Kabiyè kbp U
Kagulu kki U
Kako kkj U
Kasem xsm U
Kasem xsm U
Kenyang ken U
Kim kia U
Kim kia U
Koma kmy U
Konkomba xon U
Kono (Sierra Leone) kno U
Koonzime ozm U
Kouya kyf U
Kukele kez U
Kunda kdn U
Kuo xuo U
Kusaal kus U
Kutep kub U
Kutu kdc U
Kuwaataay cwt U
Kwere cwe U
Lama (Togo) las U
Lelemi lef U
Lobi lob U
Lokaa yaz U
Lukpa dop U
Lyélé lee U
Machame jmc U
Mada (Nigeria) mda U
Makaa mcp U
Makhuwa vmw U
Malawi Lomwe lon U
Malba Birifor bfo U
Mamara Senoufo myk U
Mampruli maw U
Mankanya knf U
Masaaba myx U
Meta' mgo U
Miyobe soy U
Moba mfq U
Moba mfq U
Mochi old U
Mossi mos U
Mossi mos U
Mumuye mzm U
Mundani mnf U
Mwan moa U
Mwani wmwU
Mündü muh U
Nafaanra nfr U
Nande nnb U
Nateni ntm U

Nawdm nmz U
Ndogo ndz U
Ngangam gng U
Nigeria Mambila mzk U
Nilamba nim U
Ninzo nin U
Nkonya nko U
Noone nhu U
Northern Dagara dgi U
Ntcham bud U
Nyabwa nwb U
Nyakyusa-Ngonde nyy U
Nyankole nyn U
Nyaturu rim U
Nyole nuj U
Nyoro nyo U
Nzima nzi U
Obolo ann U
Oku oku U
Paasaal sig U
Plapo Krumen ktj U
Pokomo pkb U
Pular fuf U
Rigwe iri U
Rundi run U
Saamia lsm U
Sango sag U
Sekpele lip U
Selee snw U
Sena seh U
Shambala ksb U
Sissala sld U
Siwu akp U
Soga xog U
South Fali fal U
Southern Birifor biv U
Southern Bobo Madaré bwq U
Southern Dagaare dga U
Southern Nuni nnw U
Southwest Gbaya gso U
Supyire Senoufo spp U
Talinga-Bwisi tlj U
Tampulma tpm U
Tharaka thk U
Tikar tik U
Timne tem U
Toura (Côte d'Ivoire) neb U
Tsonga tso U
Tumulung Sisaala sil U
Tuwuli bov U
Tyap kcg U
Vengo bav U
Vunjo vun U
West-Central Limba lia U
Yocoboué Dida gud U

OTO-MANGUEAN: 27

Atatláhuca Mixtec mib U
Ayutla Mixtec miy U
Central Mazahua maz U
Chicahuaxtla Triqui trs U
Diuxi-Tilantongo Mixt xtd U
Jalapa De Díaz Mazate maj U
Jamiltepec Mixtec mxt U
Lalana Chinantec cnl U
Lealao Chinantec cle U
Magdalena Peñasco Mix xtm U
Mezquital Otomi ote U
Nopala Chatino cya U
Ozumacín Chinantec chz U
Peñoles Mixtec mil U
Pinotepa Nacional Mix mio U
San Jerónimo Tecóatl maa U
San Jerónimo Tecóatl maa U
San Juan Atzingo Popo poe U
San Marcos Tlacoyalco pls U
San Pedro Amuzgos Amu azg U
Santa María Zacatepec mza U
Sochiapam Chinantec cso U
Southern Puebla Mixte mit U
Tepetotutla Chinantec cnt U
Tezoatlán Mixtec mxb U
Usila Chinantec cuc U
Yosondúa Mixtec mpm U

PANOAN: 4

Cashinahua cbs U
Panoan Katukína knt U
Sharanahua mcd U
Shipibo-Conibo shp U

PUINAVE: 1

Puinave pui U

PÁEZAN: 1

Páez pbb U

QUECHUAN: 22

Ayacucho Quechua quy U
Cajamarca Quechua qvc U
Cañar Highland Quichu qxr U
Cusco Quechua quz U
Huallaga Huánuco Quec qub U
Huamalíes-Dos de Mayo qvh U
Huaylas Ancash Quechu qwh U
Huaylla Wanca Quechua qvw U
Inga inb U
Lambayeque Quechua quf U
Margos-Yarowilca-Laur qvm U
Napo Lowland Quechua qvo U
North Bolivian Quechu qul U
North Junín Quechua qvn U
Northern Conchucos An qxn U
Northern Pastaza Quic qvz U
Panao Huánuco Quechua qxh U
San Martín Quechua qvs U
South Bolivian Quechu quh U
South Bolivian Quechu quh U
Southern Pastaza Quec qup U
Tena Lowland Quichua quw U

SINO-TIBETAN: 24

Achang acn U
Akeu aeu U
Akha ahk U
Bawm Chin bgr U
Eastern Tamang taj U
Falam Chin cfm U
Hakka Chinese hak U
Kachin kac U
Khumi Chin cnk U
Kulung (Nepal) kle U
Lahu lhu U
Lashi lsi U
Lolopo ycl U
Mandarin Chinese cmn U
Maru mhx U
Min Nan Chinese nan U
Mro-Khimi Chin cmr U
Newari new U
Pwo Northern Karen pwww U
Sherpa xsr U
Sunwar suz U
Tedim Chin ctd U
Yue Chinese yue U
Zyphe Chin zyp U

SULKA: 1

Sulka sua U

TACANAN: 2

Ese Ejja ese U
Tacana tna U

TAI-KADAI: 4

Lao lao U E
Northern Thai nod U
Tai Dam blt U
Thai tha U E

TARASCAN: 1

Purepecha tsz U

TICUNA: 1

Ticuna tca U

TOL: 1

Tol jic U

TOR-ORYA: 1

Orya ury U

TOTONACAN: 4

Coyutla Totonac toc U
Highland Totonac tos U
Pisaflores Tepehua tpp U
Tlachichilco Tepehua tpt U

TRANS-NEW GUINEA: 12

Anjam boj U
Awa (Papua New Guinea awb U
Ese mcq U
Gwahatike dah U
Huli hui U
Ipili ipi U
Kuman (Papua New Guin kue U
Kyaka kyc U
Lower Grand Valley Da dni U
Lower Grand Valley Da dni U
Nalca nlc U
South Tairora omw U

TUCANOAN: 11

Desano des U
Guanano gvc U
Koreguaje coe U
Macuna myy U
Piratapuyo pir U
Secoya sey U
Siona snn U
Siriano sri U
Tucano tuo U
Tucano tuo U
Tuyuca tue U

TUPIAN: 8

Aché guq U
Eastern Bolivian Guar gui U
Guajajára gub U
Guarayu gyr U
Kayabí kyz U
Paraguayan Guaraní gug U
Urubú-Kaapor urb U
Western Bolivian Guar gnw U

TURKIC: 18

Bashkir bak U
Chuvash chv U
Crimean Tatar crh U
Gagauz gag U
Gagauz gag U
Kara-Kalpak kaa U
Karachay-Balkar krc U
Kazakh kaz U E
Khakas kjh U
Kumyk kum U
Nogai nog U
North Azerbaijani azj U E
Southern Altai alt U
Tatar tat U
Turkish tur U E
Turkish tur U E
Tuvinian tyv U
Uighur uig U

URALIC: 3

Finnish fin U
Komi-Zyrian kpv U
Udmurt udm U

URARINA: 1

Urarina ura U

URU-CHIPAYA: 1

Chipaya cap U

UTO-AZTECAN: 15

Central Huasteca Nahu nch U
Eastern Huasteca Nahu nhe U
El Nayar Cora crn U
Guerrero Nahuatl ngu U
Highland Puebla Nahua azz U
Isthmus-Mecayapan Nah nhx U
Isthmus-Mecayapan Nah nhx U
Mayo mfy U
Northern Oaxaca Nahua nhy U
Northern Puebla Nahua ncj U
Santa Teresa Cora cok U
Sierra Negra Nahuatl nsu U
Southeastern Puebla N npl U
Western Huasteca Nahu nhw U
Zacatlán-Ahuacatlán-T nhi U

WEST PAPUAN: 3

Galela gbi U
Tabaru tby U
Tobelo tlb U

YANOMAM: 1

Sanumá xsu U

ZAMUCOAN: 1

Chamacoco ceg U