Using Information Theory to Characterize Prosodic Typology: The Case of Tone, Pitch-Accent and Stress-Accent

Anonymous ACL submission

Abstract

This paper argues that the relationship between lexical identity and prosody—one well-studied parameter of linguistic variation—can be characterized using information theory. We predict that languages that use prosody to make lexical distinctions should exhibit a higher mutual information between word identity and prosody, compared to languages that don't. We test this hypothesis in the domain of pitch, which is used to make lexical distinctions in tonal languages, like Cantonese. We use a dataset of speakers reading sentences aloud in ten languages across five language families to estimate the mutual information between the text and their pitch curves. We find that, across languages, pitch curves display similar amounts of entropy. However, these curves are easier to predict given their associated text in the tonal languages, compared to pitch- and stress-accent languages, and thus the mutual information is higher in these languages, supporting our hypothesis. Our results support perspectives that view linguistic typology as gradient, rather than categorical.

1 Introduction

One central tension in linguistics is between linguistic *universality* and *diversity*. The world contains some 7,000 languages (Ethnologue, 2023), each with its unique and idiosyncratic lexicon, phonological inventory, and grammar. At the same time, linguistic properties are shared between sets of related languages (Croft, 2002), and some features appear, or covary, across languages, giving rise to the hypothesis that human language is governed by a set of universal principles (Greenberg, 2005). Major advances in the study of language have been made through the introduction of frameworks that can describe both the typological variation observed between languages, as well as the universal consistencies observed across languages. Examples of such frameworks are the principles and parameters approach for syntactic structure (Chomsky, 1993; Culicover, 1997) and Optimality Theory for phonological systems (Prince and Smolensky, 2004).

One promising candidate for this type of framework is information theory (Shannon, 1948). Studies have argued that information-theoretic approaches can explain universal principles in languages, including the distribution of word lengths (Zipf, 1949; Piantadosi et al., 2011; Pimentel et al., 2023), the organization of semantic systems (Kemp et al., 2018; Zaslavsky et al., 2018, 2021), word orders (Dyer et al., 2021) as well as language processing phenomena (Futrell et al., 2020; Wilcox et al., 2023). However, information-based approaches are less widely used to describe typological variation (although c.f., Futrell et al., 2020; Pimentel et al., 2020; Socolof et al., 2022). In this paper, we take one well-studied crosslinguistic parameter—whether or not a language has lexical tone—and argue that it can be characterized information-theoretically, as the amount of mutual information between a lexical item (i.e., a word) and the pitch curve associated with that word. Our goal is to demonstrate how an information-based approach can be used to characterize crosslinguistic variation, as well as to showcase how NLP methods can be used to formally quantify properties that are debated in the formal linguistics literature (e.g., whether a given language is a tonal language).

The domain we are interested in is prosody, or the melody of speech. A word's prosody is transmitted through several unique features including its duration, energy (perceived as loudness), and fundamental frequency (perceived as pitch). Pitch, specifically, is the main focus of our study. Crucially, the role that pitch plays varies across languages: In tonal languages such as Vietnamese, Mandarin, and Yoruba, all or most syllables carry one of several discrete pitch contours which differentiate between lexical items; similarly, pitch-accent languages, such as Swedish or Japanese

have an inventory of pitch contours that are lexically contrastive, but they are typically not present on every word; in stress-accent languages such as English and Italian pitch does not differentiate between lexical identity at all, playing other roles like providing cues for stress placement, or indicating whether or not a sentence is a question. However, there is an ongoing debate about the validity of this simple three-way distinction, with some arguing for a typological continuum over multiple prosodic features (see, e.g., Hyman, 2006).

We use pitch as a case study to demonstrate how a typological parameter can be recast in information-theoretic terms. We hypothesize that given a lexical item's identity, it should be easier to predict its pitch curve in tonal compared to non-tonal languages; this is because in tonal languages pitch is used to distinguish lexical identity. Information-theoretically, this means there should be more mutual information between lexical identity and pitch in tonal languages, such as Cantonese, than in non-tonal languages, such as English.

To test this hypothesis, we use a pipeline (Wolf et al., 2023) originally developed in English to measure the mutual information between prosody and written text; where text is used as a proxy for lexical identity. We make several technical contributions to this pipeline, enabling it to produce more accurate MI estimates across languages. We measure mutual information for ten typologically distinct languages: English, French, Italian, German, Swedish (Indo-European) Mandarin, Cantonese (Sino-Tibeten), Japanese (Japonic), Thai (Kra–Dai), and Vietnamese (Austroasiatic). These languages are traditionally classified as either stress-accent, pitch-accent, or tonal. We find that, across languages, pitch curves display similar amounts of entropy, suggesting that pitch itself conveys similar amounts of information in each language. However, these curves are easier to predict given their associated text in the tonal languages, compared to pitch- and stress-accent languages, and thus the mutual information is higher in these languages, supporting our hypothesis. Interestingly, the mutual information does not follow a multimodal distribution, which would classify languages into clearly distinct categories. Rather they show a continuum of values, in line with recent work arguing for a gradient, rather than a categorical approach to prosodic typology (Hyman, 2006) and linguistic typology more broadly (Pimentel et al., 2020; Levshina et al., 2023; Baylor et al., 2024).

2 Prosodic Typology

In this section, we provide a formal framework for describing linguistic typologies based on prosodic features. We start by outlining our notation: We assume that each natural language consists of lexical items, w, drawn from a vocabulary $\mathcal{W}$. We use W to denote a lexical-item-valued random variable. By "lexical item" we mean the sense of dictionary definitions—each value of W is associated with a unique semantic meaning, rather than with a particular orthographic representation. However, as we do not have direct access to lexical identities in a large corpus, we will relax this in our experiments and work instead with orthographic words, which we use as a proxy for lexical identities. In addition, we define $\mathbf{p}$, as a real-valued vector that represents some prosodic feature for a given word. Although in our subsequent experiments, $\mathbf{p}$ refers only to the pitch curve, for the present we will use $\mathbf{p}$ for prosody as a whole, including other features, such as average acoustic energy or duration. We denote a prosody-valued random variable as $\mathbf{P}$.

What does it mean for a language to have contrastive tone, stress, or length? In linguistics textbooks, this is often defined through minimal pairs, by showing that there are systematic correspondences between lexical identity and the prosodic feature of interest. For example Yip (2002) illustrates the notion of a tonal language by giving an example of the syllable *[yau]* in Cantonese. If pronounced with a high-rising tone, this syllable means *paint*, however, if pronounced with a low-level tone, it means *again*. Based on such examples, we propose the following definition:

Definition 1 *A language ℓ is typologically a $\mathbf{p}$-language if, in ℓ, prosodic feature $\mathbf{p}$ provides information about lexical identity.*

That is, if a language is a $\mathbf{p}$-language, then knowing the prosodic value, $\mathbf{p}$, of a particular lexical item, w, should make that word easier to predict. As an example, in Cantonese, if we know a word has a high-rising tone, then it will be easier to predict that word's meaning, compared to a situation where we don't know the pitch at all.[1]

Based on this definition, we propose that one natural way to describe prosodic typologies is through the lens of information theory. Under informa-

[1]We acknowledge that "providing information about" lexical identity is a less stringent requirement than, say *determining* lexical identity. We adopt this definition, in part, because it is more conducive to measuring, experimentally.

tion theory, if a variable (e.g., $\mathbf{p}$) makes another variable (e.g., w) easier to predict, we say that it contains information about it. We can thus say that a $\mathbf{p}$-language should be one where pitch conveys information about lexical information, written as:

$$\mathrm{MI}(\mathbf{P}; \mathrm{W}) > 0 \tag{1}$$

That is, the MUTUAL INFORMATION (MI) between $\mathbf{p}$ and w is greater than zero. Conversely, in non $\mathbf{p}$-languages, where $\mathbf{p}$ does not determine lexical identity, the mutual information will be roughly equivalent to zero, i.e., $\mathrm{MI}(\mathbf{P}; \mathrm{W}) \approx 0$. Note that because mutual information is symmetric, in $\mathbf{p}$-languages, we also predict that lexical identity reduces uncertainty about prosodic features, which is what we empirically test in the following sections.

2.1 Predictions: Tone, Stress and Pitch-accent Languages

The prediction outlined in eq. (1) is limited, in that it only makes a binary classification: MI should be positive in $\mathbf{p}$-languages, and equal to zero in non $\mathbf{p}$-languages. However, in real life, we expect that things are more complicated. Rather than a single distinction, one might expect to find more nuanced differences between languages. This should be the case especially when it comes to pitch—the focus of our study—as existing typologies already separate languages into (at least) three categories based on the relationship between pitch and lexical identity. We therefore outline three more concrete hypotheses, concerning the mutual information, MI of a language's lexical identity (W) and pitch ($\mathbf{P}$):

Hypothesis 1 *Typological Ordering Hypothesis: Languages will display the following ordering of average* MI *within linguistic typological groups: tonal languages » pitch-accent languages » stress-accent languages.*

In addition, we formulate two competing hypotheses that correspond to different approaches toward linguistic typology:

Hypothesis 2 *Categorical Prediction: Languages will display a categorical distinction in MI, separated into modes corresponding to typological group.*

Hypothesis 3 *Gradient Prediction: Languages will display a gradient in* MI *on a gradual continuum. Differences between languages within a typological group can be as large as differences between groups.*

To explore these hypotheses, we improve an existing pipeline for estimating MI, the details of which we will turn to in section 3.

2.2 A Type- or Token-level Prediction?

It is important to note the nature of the information we treat here. In particular, we could define the MI above in two ways: at the type or token level. These would quantify categorically different linguistic properties. A type-level $\mathrm{MI}(\mathbf{P}; \mathrm{W})$ measures how predictable a novel word's pitch is given its lexical identity; it would thus quantify if $\mathbf{p}$-values are systematically assigned to words based on their meaning or orthography. As lexicons' form–meaning assignments are largely arbitrary (a property known as the arbitrariness of the sign; Saussure, 1916; Dautriche et al., 2017; Pimentel et al., 2019), we would expect such type-level MI to be small in both $\mathbf{p}$- and non-$\mathbf{p}$-languages. A token-level $\mathrm{MI}(\mathrm{W}; \mathbf{P})$, on the other hand, quantifies how well $\mathbf{p}$ disambiguates *known* words in a language, and should thus have significantly different values in $\mathbf{p}$- and non-$\mathbf{p}$-languages. We thus focus on this MI's token-level definition here.

3 Methods

The prediction in eq. (1) is about *lexical items*, however, we do not have direct access to these in the multilingual corpora we use for this study. Rather, we have access to textual representations, i.e., orthographic words, which often correspond to lexical items. In the rest of this paper, therefore, we take W to be a random variable corresponding to either a piece of *text* or an orthographic word. Furthermore, as we are specifically interested in pitch, from here on out $\mathbf{P}$ is a random variable that represents the parameterization of a pitch curve, specifically, as opposed to just a general prosodic feature. We discuss how we represent $\mathbf{P}$ and W at greater length in Section 3.3 and Section 3.4.

3.1 Estimating Mutual Information

We estimate the mutual information between prosody and text, following the proposal from Wolf et al. (2023). Wolf et al. estimate this quantity by first decomposing MI as the difference between two entropies, and separately estimating each term

$$\begin{aligned} \mathrm{MI}(\mathbf{P}, \mathrm{W}) &= \mathrm{H}(\mathbf{P}) - \mathrm{H}(\mathbf{P} \mid \mathrm{W}) && \text{(2a)} \\ &\approx \mathrm{H}_\theta(\mathbf{P}) - \mathrm{H}_\theta(\mathbf{P} \mid \mathrm{W}) && \text{(2b)} \end{aligned}$$

As represented by eq. (2b), we estimate the MI as the difference between two *cross entropies*,

$\mathrm{H}_\theta(\cdot)$. The cross-entropy is defined as the expectation of $-\log p_\theta(\mathbf{p})$ or $-\log p_\theta(\mathbf{p} \mid \mathrm{w})$, given the ground-truth distribution $p(\mathbf{p})$ or $p(\mathbf{p}, \mathrm{w})$, respectively. Following Wolf et al. (2023), we use *re-distributive sampling* (Tibshirani and Efron, 1993; Beirlant et al., 1997) to estimate these quantities. Given model p_θ, we select a set of held-out test samples from our dataset, and then estimate each quantity as the average negative log probability (i.e., surprisal) of these test items:

$$\mathrm{H}_\theta(\mathbf{P}) \approx \frac{1}{N}\sum_{n=1}^{N} \log \frac{1}{p_\theta(\mathbf{p}^n)} \tag{3a}$$

$$\mathrm{H}_\theta(\mathbf{P} \mid \mathrm{W}) \approx \frac{1}{N}\sum_{n=1}^{N} \log \frac{1}{p_\theta(\mathbf{p}^n \mid \mathrm{w}^n)} \tag{3b}$$

Where $\mathbf{p}^n$ and w^n are the n^{th} text/pitch pair in our test set. In order to make this estimation, we need to learn a probability distribution $p_\theta(\mathbf{p})$ and $p_\theta(\mathbf{p} \mid \mathrm{w})$. We do so with the following methods.

3.1.1 Estimating $p_\theta(\mathbf{p})$

Following Wolf et al. (2023) we estimate the unconditional distribution with a Gaussian Kernel Density Estimate, KDE (Parzen, 1962; Sheather, 2004). Bandwidth is optimized via 10-fold cross-validation, using the training and validation data, selecting from Scott's rule, Silverman's rule, and fixed values. We implement this with `SciPy` (Virtanen et al., 2020). After selecting the optimal bandwidth, we fit the KDE on the training data and compute eq. (3a) on the held-out test data.

3.1.2 Estimating $p_\theta(\mathbf{p} \mid \mathrm{w})$

Wolf et al. (2023) estimate this conditional distribution by using a neural network to *learn* the parameterization, ϕ of a predictive distribution $\mathcal{Z}(\cdot)$ that captures the desired conditional probability distribution, $p_\theta(\mathbf{p} \mid \mathrm{w})$. In their setup, the predictive distribution is always either a Gaussian, Gamma or Laplace distribution. This, however, leads to a discrepancy between the expressivity of the distribution learned for the conditional and unconditional distributions, $p_\theta(\mathbf{p} \mid \mathrm{w})$ and $p_\theta(\mathbf{p})$. The KDEs used to model $p_\theta(\mathbf{p})$ construct *non-parametric* distributions from the bottom-up, summing together many Gaussians and having a number of parameters that grows with K, the number of samples in the training dataset; this distribution can thus be increasingly complex given larger training datasets. However, the learned conditional distribution, $p_\theta(\mathbf{p} \mid \mathrm{w})$, is fit as a *parametric* distribution $\mathcal{Z}$ (Gaussian or Gamma), and is thus constrained independently of the training dataset size. Therefore, the two distributions are *not* apples to apples. In particular, the greater expressivity of the unconditional distribution $p_\theta(\mathbf{p})$ means that, in practice, eq. (2b) is likely to underestimate the true mutual information and can often be negative. To fix this problem, we use two different methods for estimating the conditional probability distribution, which we outline below.

Language	Tag	Type	Family	Hours	Tokens	Types	Speakers
German	DE	SA	Indo-Euro.	8.6	47819	13519	338
English	EN	SA	Indo-Euro.	7.8	47670	10930	557
French	FR	SA	Indo-Euro.	7.4	27974	8062	260
Italian	IT	SA	Indo-Euro.	8.7	39413	10937	1641
Japanese	JA	PA	Japonic	6.4	54866	6434	896
Swedish	SV	PA	Indo-Euro.	6.6	38761	8002	461
Vietnamese	VI	Tonal	Austroasiatic	5.9	37838	2468	130
Thai	TH	Tonal	Kra-Dai	6.8	42153	4315	1749
Cantonese	YUE	Tonal	Sino Tibetan	6.5	37380	6753	747
Mandarin	ZH	Tonal	Sino Tibetan	7.9	36729	12547	1723

Table 1: Overview of the languages and dataset used in this study. SA= Stress Accent, PA= Pitch Accent.

Conditional KDEs: For this method, we partition the dataset by orthographic word type and fit a different KDE for each partition. We use two different estimation procedures: In the first, C-KDE-ALL, we use the whole dataset for bandwidth selection, training, and entropy estimation, which we do using Monte Carlo sampling. In the second, C-KDE-SPLIT, we use 70% the dataset for bandwidth selection and training and estimate entropy using redistributive sampling on the held-out portion. As the accuracy of the estimate decreases with fewer samples, we discard words with frequency lower than a certain threshold, λ. We conducted several pilot experiments with $\lambda = \{20, 30, 40, 50, 60\}$ and found that the qualitative nature of the results did not change. In section 4, we present results for $\lambda = 40$.

Mixture Density Networks (MDNs): For our second method, we employ a MIXTURE DENSITY NETWORK (MDN) (Bishop, 1994). MDNs are very similar to KDE estimators insofar as the final conditional probability is the sum of several Gaussian kernels. However, the means and variances of these Gaussians are *learned* by a neural network, LM_θ, with parameters θ, given input w. In addition, the network also learns a set of weights $\mathbf{w}$ that govern the mixture of the individual Gaussians.

The conditional distribution is therefore:

$$p_\theta(\mathbf{p} \mid \mathrm{w}) = \sum_{k=1}^{K} w^k(\mathrm{w};\theta)\mathcal{N}(\mathbf{p} \mid \mu^k(\mathrm{w};\theta), \sigma^k(\mathrm{w};\theta)) \quad (4)$$

where w^k is the weight μ^k is the mean and σ^k is the variance of the k^{th} Gaussian kernel. To ensure the properness of the distribution, the weights (which must sum to one) are the output of a softmax function, and standard deviations (which must be positive) are the output of a soft-plus function. We use 20 kernels per dimension. The MDN itself consists of a Multi-Layer Perceptron, where the number of layers and hidden units are hyperparameters tuned during training, and by default, we use a five-layer MLP with 512 hidden units per layer.

3.2 Dataset

We use the Common Voice dataset (Ardila et al., 2020), a multilingual corpus that contains paired text–audio samples from contributors reading individual sentences out loud.[2] Samples are rated by other contributors who assign them either a thumbs-up or a thumbs-down. The validated portion of the dataset that we use includes only sentences whose first two ratings are up-votes. We select data from ten languages, across five different language families, representing a range of stress-accent, pitch-accent, and tonal languages. The details of each language are given in Table 1. We sample 5000 sentences per language for consistency, based on the language with the fewest validated sentences. In order to extract word-level prosodic features we align each sentence's audio to its text at the word level using the Montreal Forced Aligner (McAuliffe et al., 2017). For our Sino-Tibetan languages (Mandarin and Cantonese) we use two different tokenization or word-grouping schemes. In one both MFA alignment and NLP tokenization use characters as input units (this is tagged with `(chr)` in figures), and in the other MFA aligns audio to words, and NLP tools tokenize sentences into words using their default tokenizer.

3.3 Representation of Pitch

Representing the pitch curve of a word presents substantial challenges: We want to find a relatively low-dimensional representation space, but one that can still capture the complexities of pitch contours across languages, which may, for example, contain rising and falling elements on a single word. To do so, we use the preprocessing methods given in Suni et al. (2017) to extract the fundamental frequency, f_0 from the raw waveforms from each aligned word segment, and to remove outliers. We apply interpolation to create a smooth f_0 curve across moments where no pitch is being produced, for example during unvoiced consonants. Once it has been extracted, we resample the f_0 curves to 100 points and parameterize them with the first four coefficients of a discrete cosine transform (DCT). The objective of our prosodic pipeline, therefore, is to estimate the four coefficients of the DCT pitch representation.

3.4 Representation of Text

Although our prediction about prosodic typologies concerns lexical items, we estimate mutual information between pitch and *text*. For our conditional KDEs we simply condition on orthographic words in the dataset. However, for our MDNs, we represent text using the representational space of pretrained word embeddings and language models. In our experiments, we use three different models, corresponding to different amounts of context:

No context (fastText): We use fastText representations (Bojanowski et al., 2017) to estimate $p(\mathbf{p} \mid \mathrm{w})$. To do so, we simply feed the fastText embedding as the input into the MDN network. As the fastText embeddings provide non-contextual representations of word forms, in conjunction with the conditional KDEs, we treat them as closest to instantiating the "lexical identity" over which our prediction is based. Therefore, we predict that the differences in MI between pitch-accent, stress-accent, and tonal languages will be the strongest for this textual representation and for conditional KDEs.

Previous Context (mGPT) and Bidirectional Context (mBERT): Additionally, we estimate $p(\mathbf{p} \mid \mathrm{w})$ using representations from mGPT (Shliazhko et al., 2024) a multilingual autoregressive language model, largely based on the GPT-2 architecture, as well as a multilingual version of BERT, mBERT (Devlin et al., 2019). For both models, we use hidden representations as inputs to our MDN network. During training, we fine-tune the combined model, not just the MDN network. When words are tokenized into multiple parts, we use the representation of the final token.

[2]The dataset is released under a Creative Commons Attribution Share-Alike license.

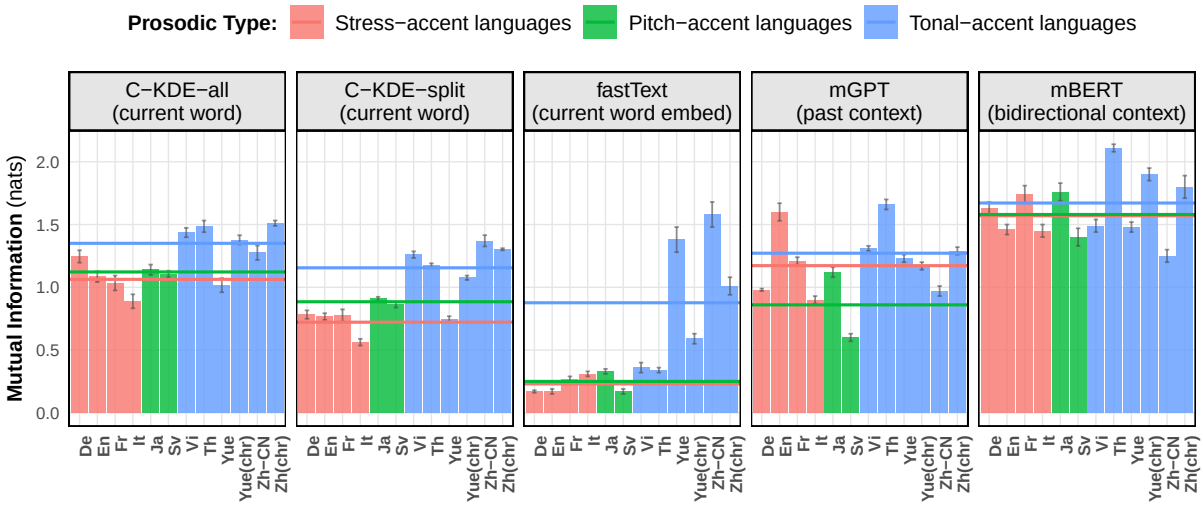


Figure 1: **Main Results:** Mutual information between pitch and text across languages. Lines show within typological group averages. Error bars show standard deviations from Monte Carlo resampling (C-KDE-all, C-KDE-split) or 5-fold cross-validation (fastText, mGPT, mBERT). We find that tonal languages have higher MI on average compared to stress-accent and pitch-accent languages.

As both mGPT and mBERT have access to context, they are capable of disambiguation between different senses of a word (Chawla et al., 2021). On one hand, this may make their representations closer to the "lexical identity" over which our theoretical prediction is made. On the other hand, because these models have access to context, they also likely represent non-lexical properties of the context that affect pitch. For example, although English is not a tonal language, sentence-final punctuation (e.g., question marks) can provide strong cues to pitch. Therefore, there may be nonzero MI between pitch and mBERT representations, even for non-tonal languages. In addition, because of their longer context, these models can be thought of as estimating MI between pitch and the *sentence*, as opposed to pitch and the word. We, therefore, make two predictions: First, because of the increased representational capacity of our neural LMs, we expect *higher mutual information* when they are used for estimation. Second, because our conditional KDE and fastText embeddings more closely resemble lexical identity, we expect *greater differences in* MI between tonal and non-tonal languages when these methods are used.

4 Results

4.1 Main Results

Mutual Information: The results of our experiment are visualized in fig. 1, with our different representations of text across the different facets. Horizontal bars show within typological group averages. The data support the typological ordering hypothesis: We observe higher MI in tonal languages compared to non-tonal languages, for all of our estimation methods. Additionally, we find evidence supporting the tonal » pitch-accent » stress-accent hierarchy, especially for our C-KDE and fastText models. The ordering is not present for mGPT, where stress-accent languages have higher average MI than pitch-accent languages, or in mBERT, where stress- and pitch-accent languages have almost identical MI.

Following the logic outline in section 3.4, we observe the greatest separation between tonal and non-tonal languages when using estimation techniques that do not take context into account (i.e., fastText and C-KDE). While estimation methods that incorporate longer context tend to have higher mutual information on average, these methods collapse the difference between typological groups. For example, using mBERT, we find the highest average MI of any model, but we also find almost no difference between tonal and stress-accent languages, in terms of group averages. We suspect that this is because mBERT, with its bidirectional context, is capable of representing non-lexical information that can be useful for predicting pitch even in non-tonal languages, for example, whether a given sentence is a question.

Interestingly, even though prosodic type behavior is consistent across models (i.e., tonal languages

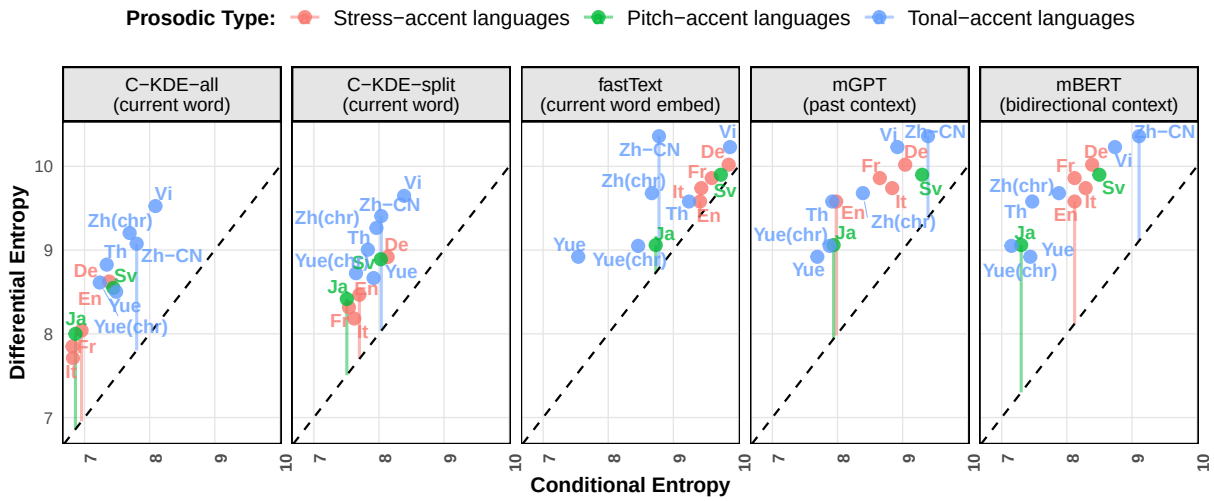


Figure 2: **Main Results Separated into Conditional/Differential Entropy:** Dashed line shows the $y = x$ line. Points show individual languages. Colored lines for En, Ja, and Zh visualize the MI of these languages, which is the points' vertical distance from the dotted line.

always have the highest MI), within each prosodic type models show variability. For example, our conditional KDE methods suggest that German is the stress-accent language with the highest MI between pitch and lexical item. However, when using mGPT, we find the highest MI for English, and in mBERT, French. One possibility, here, is that the different ways we represent context between these models lead to different amounts of MI. We return to this point in the larger context of our *gradient* vs. *categorical* hypotheses in the discussion.

Conditional and Differential Entropy: To zoom in on these data further, fig. 2 shows the same results broken down into conditional and differential entropy. The difference between these two is the MI, shown in figure 1 and visualized here as the vertical distance to the $x = y$ line, which is plotted for English, Japanese, and Mandarin.[3] Overall, we observe a relatively narrow range for both differential entropy (ranging from 8–10.5 bits) and conditional entropy (ranging from 7–10 bits) across languages. These data support recent studies showing that information-theoretic properties of human language exist within a narrow bandwidth (Bentz et al., 2017; Wilcox et al., 2023; Pimentel et al., 2020)

[3]We find higher entropy in our word embedding models, compared to our C-KDE models. We believe that because we excluded words that occurred fewer than $\lambda = 40$ times, this dataset was free of many low-frequency words whose pitch was potentially difficult to predict. Therefore, this difference may be an artifact of our methods, and not necessarily reflective of the C-KDE being an overall better estimation technique.

When looking at conditional entropy instead of mutual information, we observe more consistency at the language level. For all methods, Vietnamese, Chinese, and German have higher entropy (both conditional and differential), and Japanese, Cantonese, and (to some extent) English have lower entropy. The overall amount of entropy present in a language does not follow typological patterns or even the complexity of a language's tonal system. Cantonese, which is traditionally analyzed as having nine tones, always has lower entropy values than Mandarin, which is typically analyzed as having only four. However, other factors like average word length and isolating vs. agglutinating could be factors.

4.2 Follow-up Experiment: Effect of Subword Tokenization

One difference between C-KDE, fastText and our neural-network-based estimation techniques (mBert and mGPT) is that the latter two use subword tokenization schemes. For words that have multiple tokens, we used the embedding of the last token in the word during estimation. It's possible that this skews or biases our results. Additionally, the number of single-token words varies across languages within our multilingual models, with English having more single-token words than the other languages. To investigate how this may impact our results, we took each of our initial datasets, and subsetted them to include only words with k or fewer tokens. We then re-ran our MI estimation proce-

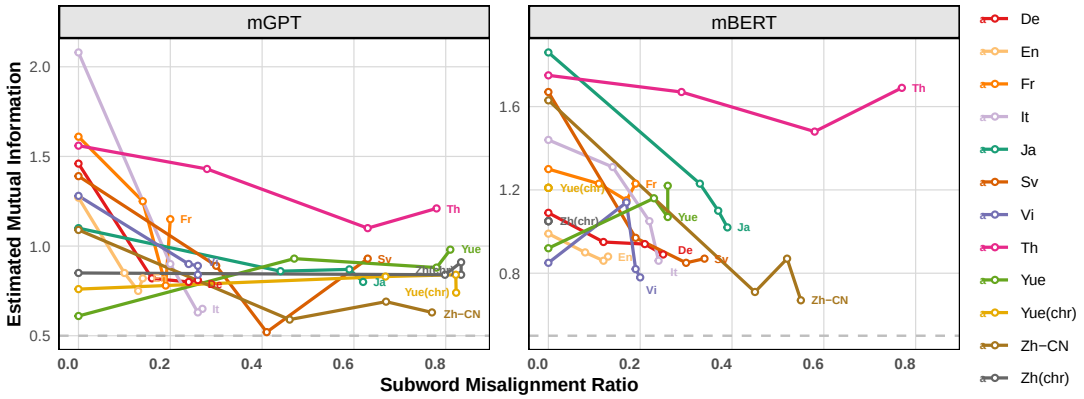


Figure 3: **Impact of tokenization on** MI **estimation:** x-axis shows the proportion of words in our dataset tokenized into more than one token. Subsampling data to include only words with one token changes the estimated MI.

dure using only mBERT and mGPT. This resulted in datasets that were balanced in terms of tokens-per-word, but not in terms of total dataset size.

The results are visualized in fig. 3. We can see that as the percentage of multi-token words decreases, the MI estimation changes, suggesting that, indeed, this impacts our results. However, the overall picture of the results remains the same—there is no clear separation between tone, pitch-accent and stress-accent languages using these models. Howe the tokens-per-word ratio decreases, the MI increases for most (although not all) languages, suggesting that the MI estimates in fig. 1 are slight underestimates. For additional presentation of these data see appendix A.

5 Discussion

Our experiments supported the typological ordering hypothesis, namely that tonal languages have higher MI between pitch and text, followed by pitch-accent and stress-accent languages. The ordering of languages according to this prediction is relatively clean, especially for the tonal vs. non-tonal distinction. Among the C-KDE estimates, where we expect the separation to be the strongest, we found only one tonal language (Cantonese, word level) with a lower MI than any stress-accent language. And with fastText, we found that all tonal languages had higher MI than all stress-accent. Finally, we generally found that pitch-accent languages fell between tonal and stress-accent languages, as expected.

What do our results say about the status of categorical vs. gradient typological theories? On one hand, they could be construed to support the categorical prediction. Using fastText, we can find a single amount of mutual information (0.34 nats) that separates all tonal from non-tonal languages. At the same time, our results demonstrate interesting gradient differences both between and within prosodic types. Firstly, it's not the case that languages are clearly separated into different modes based on typological type. For example, in our fastText models, there is far more variation in MI *within* tonal languages (ranging from 0.36–1.58 nats) than *between* tonal vs. stress-accent groups (0.23 vs. 0.88 nats). Based on these considerations, we conclude that our data are more closely aligned with the gradient prediction as outlined in section 2.1.

We close by discussing the relationship between our definition of a p-language and Greenberg's notion of an implicational universal (Greenberg, 2005). While implicational universals result in mutual information between linguistic properties, we argue that it is not possible to reduce such universals to MI alone. To take one example, a well-studied implicational universal holds that VSO languages always have prepositions (as opposed to *post*positions). This implies that there is mutual information between a language's word order and its adposition placement. However, if the implication were reversed—VSO implies postpositions—the amount of MI would remain unchanged. Importantly, implicational universals specify *how* features of a language covary, not just that they *do* covary. Zooming out, we can say that implicational universals and p-languages are a larger class of linguistic variation that implies MI between linguistic features. Further characterizing how mutual information relates to known typological features is an important direction for future research.

Limitations

One limitation of this work has to do with our dataset: First, the dataset is relatively small, with just $5,000$ sentences per language. Second, we did not control for the number of unique speakers in the dataset, meaning that some languages have over-representation from a single or handful of individuals. For example, our Thai data includes samples from 1749 speakers, whereas our Vietnamese data includes samples from just 130 speakers. One other shortcoming of our dataset is that while our pitch-accent and tonal languages include data from multiple language families, our stress-accent data comes entirely from Indo-European languages. Finally, our dataset did not control for content, meaning the distribution of concepts and therefore words could vary substantially between different languages. While collecting high-quality audio-text-aligned data across multiple languages is a difficult undertaking, assembling such a dataset could be the basis for future research.

Ethics Statement

We foresee no obvious ethical problems with this research. Furthermore, we do not foresee any obvious risks with this research.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from universal dependencies. *Preprint*, arXiv:2402.01513.

Jan Beirlant, Edward J Dudewicz, László Györfi, Edward C Van der Meulen, et al. 1997. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6).

Christopher M Bishop. 1994. Mixture density networks.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Avi Chawla, Nidhi Mulay, Vikas Bishnoi, Gaurav Dhama, and Anil Kumar Singh. 2021. A comparative study of transformers on word sense disambiguation. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*, pages 748–756. Springer.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

Peter W Culicover. 1997. *Principles and parameters: An introduction to syntactic theory*. Oxford University Press.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T. Piantadosi. 2017. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8):2149–2169.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dyer, Richard Futrell, Zoey Liu, and Greg Scontras. 2021. Predicting cross-linguistic adjective order with information gain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 957–967, Online. Association for Computational Linguistics.

Ethnologue. 2023. Ethnologue: Languages of the world.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Joseph H. Greenberg. 2005. *Language Universals: With Special Reference to Feature Hierarchies*. De Gruyter Mouton, Berlin, New York.

Larry M Hyman. 2006. Word-prosodic typology. *Phonology*, 23(2):225–257.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4:109–128.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan

Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

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, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255, Singapore. 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.

Alan Prince and Paul Smolensky. 2004. Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in Phonology: A reader*, pages 1–71.

Ferdinand de Saussure. 1916. *Course in General Linguistics*. Columbia University Press, New York. English edition of June 2011, based on the 1959 translation by Wade Baskin.

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

Simon J Sheather. 2004. Density estimation. *Statistical Science*, pages 588–597.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Michaela Socolof, Jacob Louis Hoover, Richard Futrell, Alessandro Sordoni, and Timothy J. O'Donnell. 2022. Measuring morphological fusion using partial information decomposition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 44–54, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 57(1):1–436.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Moira Yip. 2002. *Tone*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

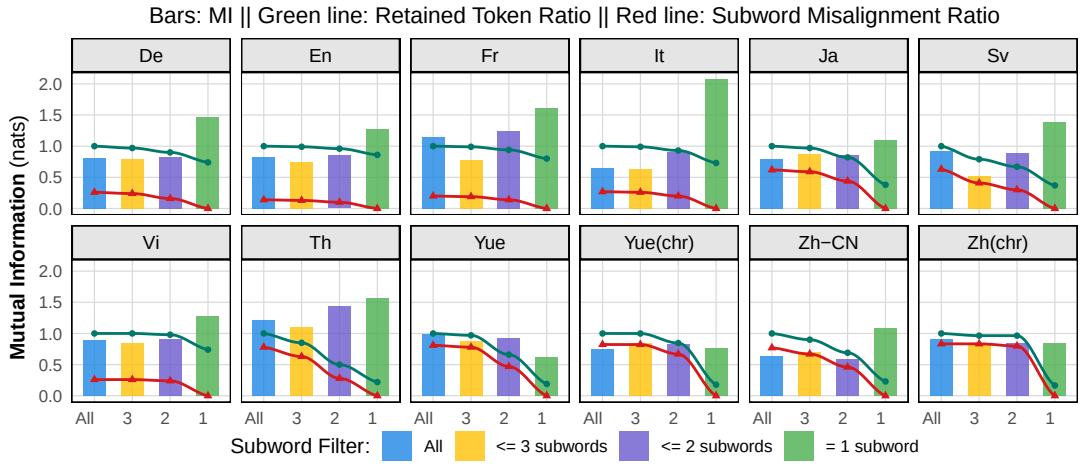
Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. 2021. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

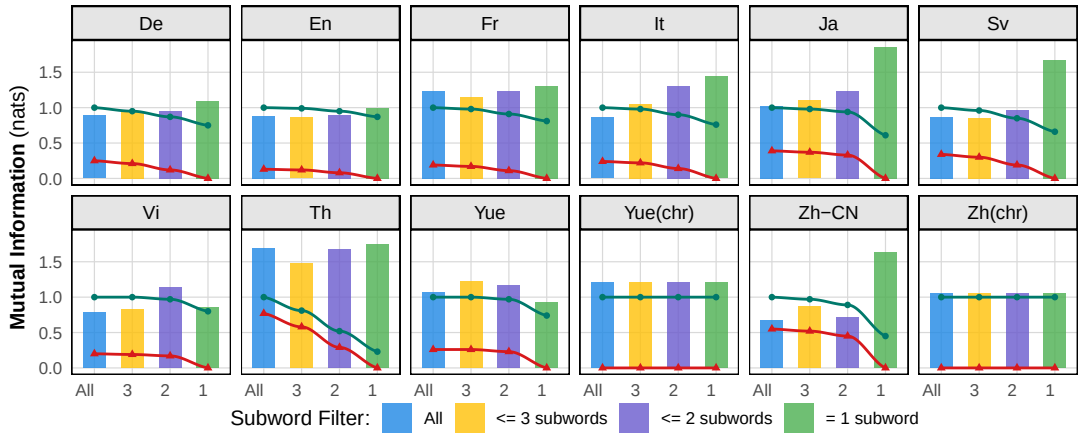
A Subword Tokenization Follow-up Analysis

In fig. 4, filtering out multi-token words (blue to green bars) increases MI, showing that subword tokenization misalignment adds noise. Cantonese (Yue) is an exception, likely due to its many single-character words.

Retained tokens (green) and misalignment (red) decrease similarly, but languages like English,



(a) mGPT



(b) mBERT

Figure 4: **Fine-Grained Analysis of Subword Tokenization Effects on MI Estimation in mGPT and mBERT.** The x-axis represents subword filtering levels: "All" (no filtering), "3" (subsetted words with at most 3 subword tokens), "2" (at most 2 tokens), and "1" (only single-token words). Bars show estimated MI, the green line represents the retained token ratio after subsetting, and the red line represents the misalignment ratio in the retained data.

French, and German keep more data, while Chinese, Thai, and Swedish lose more, resulting in cleaner but smaller datasets for MI estimation.

Languages also vary in initial misalignment (red lines). English has the least, while Chinese and Thai have more, leading to larger MI gains after filtering and suggesting that MI is likely underestimated in our data for these languages with mGPT and mBERT model representation.

B Hyperparameter and Hyperparameter search

We perform a hyperparameter search using 5-fold cross-validation to tune the fastText MDN model. The search space includes:

- Learning rate: 0.01, 0.001
- Dropout: 0.2, 0.5
- Hidden layers: 15, 20, 30
- Hidden units: 512, 1024

Models are trained for a maximum of 50 epochs using the AdamW optimizer with weight decay (L2 regularization = 0.001) and early stopping (patience = 3) based on validation loss. The best hyperparameters are selected based on average performance across the 5 folds, and evaluated on the test set.

For mGPT (`ai-forever/mGPT`) and mBERT (`bert-base-multilingual-cased`) MDNs, we fine-tune using AdamW (weight decay = 0.1), a learning rate of 5.0×10^{-5} with ReduceLROnPlateau (factor = 0.1, patience = 2), batch size 16 (effective 64), gradient clipping at 1.0, dropout of 0.1 (applied to the MLP head), and early stopping (patience = 3). For mGPT, we fine-tune only the last eight transformer layers, freezing the rest for

efficiency, resulting in 612M trainable parameters (out of 1.4B total). For mBERT, all layers are fine-tuned, with 177M trainable parameters.