The Harmonic Structure of Information Contours

Eleftheria Tsipidi[1] **Samuel Kiegeland**[1] **Franz Nowak**[1] **Tianyang Xu**[2]
Ethan Gotlieb Wilcox[3] **Alex Warstadt**[4] **Ryan Cotterell**[1] **Mario Giulianelli**[1]
[1]ETH Zürich [2]TTIC [3]Georgetown [4]UCSD
{eleftheria.tsipidi, samuel.kiegeland, franz.nowak,
ryan.cotterell, mario.giulianelli}@inf.ethz.ch,
sallyxu@ttic.edu, ethan.wilcox@georgetown.edu, awarstadt@ucsd.edu

Abstract

The uniform information density (UID) hypothesis proposes that speakers aim to distribute information evenly throughout a text, balancing production effort and listener comprehension difficulty. However, language typically does not maintain a strictly uniform information rate; instead, it fluctuates around a global average. These fluctuations are often explained by factors such as syntactic constraints, stylistic choices, or audience design. In this work, we explore an alternative perspective: that these fluctuations may be influenced by an implicit linguistic pressure towards periodicity, where the information rate oscillates at regular intervals, potentially across multiple frequencies simultaneously. We apply harmonic regression and introduce a novel extension called time scaling to detect and test for such periodicity in information contours. Analyzing texts in English, Spanish, German, Dutch, Basque, and Brazilian Portuguese, we find consistent evidence of periodic patterns in information rate. Many dominant frequencies align with discourse structure, suggesting these oscillations reflect meaningful linguistic organization. Beyond highlighting the connection between information rate and discourse structure, our approach offers a general framework for uncovering structural pressures at various levels of linguistic granularity.

https://github.com/rycolab/harmonic-surprisal

1 Introduction

Studying the rate at which speakers transmit information has been a long-standing topic of interest in linguistics and cognitive science (Shannon, 1948; Genzel and Charniak, 2002; Bell et al., 2003; Xu and Reitter, 2018; Giulianelli and Fernández, 2021, *inter alia*). From an information-theoretic perspective, effective communication involves striking a balance between a rate sufficiently low for the receiver to successfully decode the intended message and yet sufficiently high for the sender to reduce their effort (Zipf, 1949; Clark and Wilkes-Gibbs, 1986; Aylett, 1999; Aylett and Turk, 2004; Gibson et al., 2019). In this context, information is often quantified as Shannon surprisal, i.e., the negative log probability, of the unit being communicated within its context. As evident from visualizing surprisal values across a text—see, for example, Fig. 1 for token-level surprisals estimated with a Transformer language model—information rate fluctuates harmonically throughout the discourse.

To predict and explain fluctuations in the surprisal of units, prior work has examined their relationship with a unit's position within elements of the discourse structure, such as paragraphs (Genzel and Charniak, 2003), topic episodes (Qian and Jaeger, 2011; Xu and Reitter, 2016), and dialogue-specific contextual units (Giulianelli et al., 2021; Maës et al., 2022). While these studies independently establish links between aspects of discourse context and information rate, a comprehensive framework for investigating when and how a unit's position within its contextual structure affects its information remains an open question. More broadly, no overarching theory yet accounts for harmonic structure in global information contours.

One influential framework for understanding the relationship between information rate and *local* context is the **uniform information density (UID)** hypothesis (Fenk and Fenk, 1980; Aylett and Turk, 2004; Levy and Jaeger, 2006; Meister et al., 2021). This hypothesis posits that, within the constraints of grammar, speakers tend to distribute information as evenly as possible. UID accounts for production choices within narrow syntactic and discourse contexts; in particular, in environments that allow for multiple alternative formulations, speakers favor those that achieve greater information uniformity (Jaeger, 2010; Mahowald et al., 2013; Torabi Asr and Demberg, 2015). However, when UID is stretched to larger

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31636–31659
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

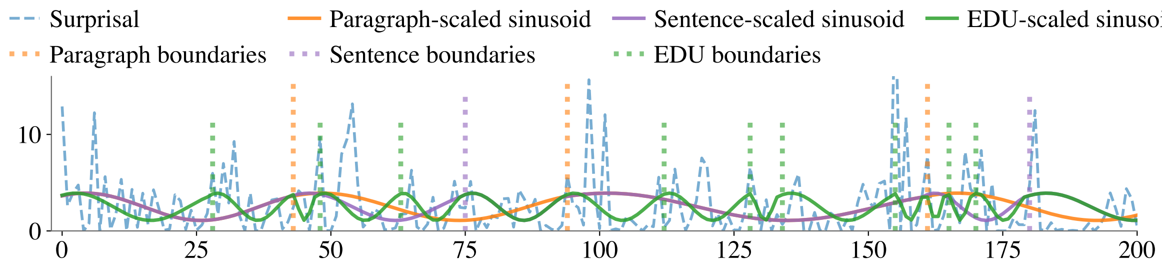

Figure 1: **Illustration of Harmonic Regression on Surprisal Contours**. Surprisal contours, unit boundaries, and first-order sinusoids for the first 200 tokens from a Wall Street Journal article (document `wsj_1111` in the English RST Discourse Bank). Time scaling (§3.2) is applied according to the lengths of elementary discourse units (EDUs), sentences, and paragraphs. Here, we set the coefficients of the sinusoids to 1 for illustrative purposes. See Fig. 3 and App. H.1 for realistic decompositions.

contextual units, its explanatory power weakens.

At a *global* level, the UID hypothesis has been taken to imply that each linguistic unit contributes a constant amount of information throughout a discourse, which corresponds to a fully rational use of the communication channel (Genzel and Charniak, 2002). Empirical findings, however, show that surprisal curves over discourse are rarely static (Xu and Reitter, 2016; Giulianelli and Fernández, 2021; Verma et al., 2023); instead, surprisal fluctuates within a bounded range; see also Fig. 1. This suggests the pressure for uniformity is counterbalanced by competing functional constraints on communication beyond grammaticality, which become increasingly influential in longer stretches of text and dialogue. Modulating surprisal within discourse can indeed serve various functions, such as adhering to aesthetic or stylistic conventions (Lewis, 1894), maintaining listener engagement and supporting comprehension (Cervantes and Gainer, 1992) reducing the cognitive demands of real-time production (Bergey and DeDeo, 2024), and enhancing task success in cooperative interactions (Yee et al., 2024). The way in which these global pressures shape the harmonic structure of information contours is not yet well understood.

To fill this gap, this paper introduces the **harmonic surprisal (HS)** hypothesis. We hypothesize that surprisal contours can be globally described as a mixture of periodic trends, and that the lengths of the different periods align with structural units of varying granularity. The HS hypothesis is a refinement of the recently introduced **structured context (SC)** hypothesis (Tsipidi et al., 2024). Both hypotheses posit that a unit's surprisal can be predicted from the unit's position within the contextual structure. However, the HS hypothesis establishes a specific relationship between position and surprisal rate—one that is governed by a periodic function. To operationalize the HS hypothesis, we propose a simple modification of harmonic regression. Harmonic regression is a variant of linear regression that models dependent variables as a combination of sinusoidal components; it offers a convenient way of examining whether surprisal contours exhibit periodic trends. Our modification allows us to embed hypotheses about relevant contextual structures directly into the statistical modeling procedure. Specifically, scaling the sinusoidal predictors by the length of a given structural unit, we are able to test whether the periodic trends of surprisal contours align with the boundaries of that unit. In doing so, this approach enables us to uncover interpretable harmonic structures that reflect the underlying contextual organization.

Building on Tsipidi et al.'s (2024) findings, we focus our analysis on sentences, paragraphs, and elementary units of the rhetorical discourse structure (Mann and Thompson, 1988). However, the flexibility of the proposed harmonic regression approach also allows us to observe the role of smaller units, whose influence on surprisal fluctuations is modeled by sinusoidal components with higher frequencies. Our analyses on English, Spanish, German, Dutch, Basque, and Brazilian Portuguese texts provide consistent evidence for periodicity in surprisal contours. This evidence for periodicity is particularly pronounced when we time-scale predictors to align with the boundaries of elementary discourse units (EDUs), with first-order sinusoids—those exactly corresponding to the EDU spans—having the highest amplitude. Overall, our findings indicate that discourse structure influences surprisal dynamics in text, with periodic patterns in surprisal contours emerging in alignment with discourse structure constituents.

2 Information Contours

The standard information-theoretic approach to analyzing the rate at which speakers transmit information in text and speech is to track **information contours**—time series representing per-unit information throughout the linguistic signal (Genzel and Charniak, 2002; Keller, 2004; Xu and Reitter, 2016; Giulianelli and Fernández, 2021). While alternative measures of information may be suitable to such analyses (Rabovsky et al., 2018; Aurnhammer and Frank, 2019; Giulianelli et al., 2023, 2024b,c; Meister et al., 2024; Li and Futrell, 2024, *inter alia*), this work adheres to the classical information-theoretic model, using surprisal (Shannon, 1948) as a measure of information. The following sections introduce key notation and concepts, along with several prominent hypotheses regarding the functional pressures shaping surprisal contours.

2.1 Alphabets, Strings, and Documents

An **alphabet** Σ is a non-empty set of symbols. A **string** $\boldsymbol{w}$ over alphabet Σ is a finite sequence of symbols $\boldsymbol{w} = w_1 \cdots w_N$, where $w_1, \ldots, w_N \in \Sigma$. The string's length N is denoted as $|\boldsymbol{w}|$ and the empty string as ε. The set of all strings composed of symbols in Σ is denoted as Σ^*. Given a string $\boldsymbol{w}$ of length $|\boldsymbol{w}| \geq t$, $\boldsymbol{w}_{<t}$ is the string composed of the first $t-1$ symbols of $\boldsymbol{w}$. We write $\boldsymbol{w} \preceq \boldsymbol{w}'$ if $\boldsymbol{w}$ is a prefix of $\boldsymbol{w}'$ and denote the concatenation of two strings $\boldsymbol{w}$, $\boldsymbol{w}'$ as $\boldsymbol{w}\boldsymbol{w}'$. We define a **document**, such as a full text or dialog, as a string $\boldsymbol{w} \in \Sigma^*$.[1]

2.2 Language Models

Given an alphabet Σ, a **language model** p is a probability distribution over strings Σ^* composed of symbols from the alphabet. The **prefix probability** under language model p is defined as

$$\overrightarrow{p}(\boldsymbol{w}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{w}' \in \Sigma^*} \mathbb{1}\left\{\boldsymbol{w} \preceq \boldsymbol{w}'\right\} p(\boldsymbol{w}'). \tag{1}$$

Eq. (1) is the probability that a string has $\boldsymbol{w}$ as prefix. The prefix probability can be used to define the conditional prefix probability of a target string $\boldsymbol{w}'$ given its preceding context $\boldsymbol{w}$:

$$\overrightarrow{p}(\boldsymbol{w}' \mid \boldsymbol{w}) \stackrel{\text{def}}{=} \frac{\overrightarrow{p}(\boldsymbol{w}\boldsymbol{w}')}{\overrightarrow{p}(\boldsymbol{w})}, \tag{2a}$$

$$\overrightarrow{p}(\text{EOS} \mid \boldsymbol{w}) \stackrel{\text{def}}{=} \frac{p(\boldsymbol{w})}{\overrightarrow{p}(\boldsymbol{w})}. \tag{2b}$$

Eq. (2b) is the conditional prefix probability of the end-of-string event EOS given a context, i.e., the probability that, if the language model p is to generate the string $\boldsymbol{w}$, then it will only generate $\boldsymbol{w}$ and not continue it in the manner of $\boldsymbol{w}\boldsymbol{w}'$.

Every language model can be expressed in autoregressive form by decomposing the probability of a string as the product of conditional probabilities of each of its symbols, followed by EOS:

$$p(\boldsymbol{w}) = \overrightarrow{p}(\text{EOS} \mid \boldsymbol{w}) \prod_{t=1}^{|\boldsymbol{w}|} \overrightarrow{p}(w_t \mid \boldsymbol{w}_{<t}), \tag{3}$$

where each conditional distribution $\overrightarrow{p}(\cdot \mid \boldsymbol{w}< t)$ is a probability distribution over $\Sigma \cup \{\text{EOS}\}$.

Modeling the Human Language Model. Many modern language models are defined via the product in Eq. (3), with each conditional prefix probability derived from a parametric model, such as a neural network. In this paper, we use neural network models as a proxy for a particular hypothetical construct model, i.e., the human language model, which is unknown.

2.3 Surprisal

Given a document $\boldsymbol{w} = w_1 \cdots w_T$, the surprisal of a unit w_t given its preceding context $\boldsymbol{w}_{<t}$ is defined as the negative logarithm of the unit's conditional probability:

$$\iota(w_t; \boldsymbol{w}_{<t}) \stackrel{\text{def}}{=} -\log \overrightarrow{p}(w_t \mid \boldsymbol{w}_{<t}), \tag{4}$$

where $\overrightarrow{p}$ is the prefix probability of a language model p. Note that the surprisal at the beginning of the string is given by $\iota(w_1) \stackrel{\text{def}}{=} -\log \overrightarrow{p}(w_1 \mid \varepsilon)$ and at the end of the string by $\iota(\text{EOS}; \boldsymbol{w}) \stackrel{\text{def}}{=} -\log \overrightarrow{p}(\text{EOS} \mid \boldsymbol{w})$. Here, for simplicity, we assume that the set of units of interest corresponds to the alphabet of the language model; however, this need not be the case (Giulianelli et al., 2024a; Vieira et al., 2024). The **surprisal contour** of a document $\boldsymbol{w} = w_1 \cdots w_T$ is defined as

$$\boldsymbol{\iota}_{\boldsymbol{w}} = [\iota(w_1), \iota(w_2; w_1), \ldots, \iota(\text{EOS}; \boldsymbol{w})]^\top. \tag{5}$$

The Role of Surprisal in Psycholinguistics. Beyond measuring the information content of linguistic units, surprisal plays an important role in psycholinguistic theory as a measure of processing effort in human language comprehension. In particular, surprisal theory posits that the effort incurred

[1] To encode a text or a dialog as a string, sentence breaks, paragraph breaks, turn transitions, and other markers of conventional document structure must be expressible in Σ^*.

by a comprehender in processing a unit is in a logarithmic relationship with its contextual probability or, equivalently, that it is proportional to the unit's surprisal. This relationship has been confirmed empirically by a large body of work using neural and behavioral measurements of processing effort in reading and listening (Fernandez Monsalve et al., 2012; Smith and Levy, 2013; Frank et al., 2015; Goodkind and Bicknell, 2018; Shain et al., 2020, 2024; Schrimpf et al., 2021; Wilcox et al., 2020, 2023; Wallbridge et al., 2022; Xu et al., 2023; Huber et al., 2024, *inter alia*). Speakers' modulation of surprisal has also been shown to explain a variety of phenomena in language production (Bell et al., 2003; Aylett and Turk, 2004, 2006; Levy and Jaeger, 2006; Frank and Jaeger, 2008; Jaeger, 2010; Futrell, 2023; Yee et al., 2024). In addition to examining the relationship between individual surprisal values $\iota(w_t; \boldsymbol{w}_{<t})$ and linguistic constructs or phenomena, another insightful approach in psycholinguistics is to analyze the broader dynamics of surprisal contours $\boldsymbol{\iota_w}$. The following sections discuss key hypotheses about the functional pressures that influence the shape of global surprisal contours.

2.4 Uniform Information Density

One of the most prominent hypotheses regarding the shape of surprisal contours is the uniform information density (UID) hypothesis (Fenk and Fenk, 1980; Aylett and Turk, 2004; Levy and Jaeger, 2006). UID has been proposed as a constraint across multiple levels of the linguistic hierarchy, affecting consonant deletion (Cohen Priva, 2015), syllable duration (Aylett and Turk, 2004), word abbreviation (Mahowald et al., 2013), syntactic reduction (Levy and Jaeger, 2006), and discourse as a whole (Genzel and Charniak, 2002). When extended to global contexts, the UID hypothesis can be expressed as follows.

Hypothesis 2.1 (Uniform Information Density; UID)**.** *Subject to the constraints of the grammar, speakers optimize their linguistic signals such that the surprisals $\boldsymbol{\iota_w}$ are distributed as uniformly as possible throughout a document $\boldsymbol{w}$.*

There are several ways to operationalize the uniformity of the information contour $\boldsymbol{\iota_w}$. Uniformity can be expressed either through local variance, where the surprisal of adjacent units is evenly distributed, or through global variance, where surprisal tends to regress toward a global mean (Collins, 2014). Comparing these two operationalizations, Meister et al. (2021) and Giulianelli and Fernández (2021) find stronger evidence for information uniformity on a global scale—whether considering words or entire utterances as linguistic units—supporting the notion that, at the discourse level, UID is better understood as a regression toward a mean information rate. Nonetheless, the pressure for surprisal to regress to a mean does not fully account for the variability and internal structure of surprisal contours. Specifically, one might expect fluctuations in surprisal contours not merely to represent noise around the global mean but instead to reflect linguistic structures of varying granularity, ranging from collocations and syntactic constructions to broader discourse organization.

2.5 The Structured Context Hypothesis

Regular fluctuations in surprisal values have been observed in empirical studies at nearly every level for which UID has been claimed, including the character level (Elman, 1990), the syntactic level (Slaats et al., 2024), and the discourse level (Xu and Reitter, 2016, 2018; Giulianelli and Fernández, 2021; Maës et al., 2022; Verma et al., 2023). Taken together, these studies indicate that there are pressures beyond UID influencing the shape of surprisal contours and that the deviations of surprisal away from a global base rate follow a pattern that is predictable from the unit's position within its containing structural units. Tsipidi et al. (2024) propose the structured context (SC) hypothesis as a refinement of the UID hypothesis. The SC hypothesis explicitly describes the relationship between surprisal contours and the hierarchical structure of a document.

Hypothesis 2.2 (Structured Context; SC)**.** *The components $\iota(w_t; \boldsymbol{w}_{<t})$ in the surprisal contour $\boldsymbol{\iota_w}$ of a document $\boldsymbol{w}$ are (partially) determined by the position of w_t within the hierarchy of $\boldsymbol{w}$'s constituent structural units.*

In other words, this hypothesis posits that we should be able to predict surprisal contours from features that describe the hierarchical structure of language at different levels. Tsipidi et al. (2024) test this at the discourse level using features such as the relative position of w_t within a higher-level structure (e.g., a sentence or a rhetorical discourse unit) and the position of structural units that contain w_t within their parent node as predictors in a linear model. They find that both shallower and deeper types of hierarchical features are significant predictors of surprisal contours. However, the predictive power of their models is moderate, leaving open the possibility that refinements to the structural features and our assumptions about their relationships to surprisal could better explain fluctuations in information contours.

2.6 The Harmonic Surprisal Hypothesis

To offer a more precise account of global surprisal fluctuations, we propose a refinement of the UID and SC hypotheses by restricting the set of possible explanations to those that inherently capture the oscillatory nature of surprisal contours.

Hypothesis 2.3 (Harmonic Surprisal; HS)**.** *The components $\iota(w_t; \boldsymbol{w}_{<t})$ in the surprisal contour $\boldsymbol{\iota}_{\boldsymbol{w}}$ of a document $\boldsymbol{w}$ vary periodically, with periods that correspond to the boundaries of structural units within $\boldsymbol{\iota}_{\boldsymbol{w}}$.*

This can be understood as stating that the predictability of surprisal contours posited by the SC hypothesis arises from a certain degree of regularity in surprisal fluctuations and that such regularity can be better described as a mixture of periodic patterns. Our experiments examine whether the periods contributing to this mixture correspond to the span of discourse units. Furthermore, the reason we refer to Hypothesis 2.3 as an elaboration of UID, rather than a competing hypothesis, is that the existence of periodic structure in $\boldsymbol{\iota}_{\boldsymbol{w}}$ does not contradict that surprisal values should be evenly distributed locally, e.g., with adjacent surprisals $\iota(w_t; \boldsymbol{w}_{<t})$ and $\iota(w_{t+1}; \boldsymbol{w}_{<t+1})$ of similar magnitude, nor does it contradict a global notion of uniformity where surprisal values tend to accumulate around the mean surprisal in $\boldsymbol{\iota}_{\boldsymbol{w}}$.

3 Harmonic Regression Modeling

Hypothesis 2.3 motivates the search for a statistical method that can automatically discover periodic structure in the information contour $\boldsymbol{\iota}_{\boldsymbol{w}}$. In this work, we adopt a time series modeling perspective, specifically using **harmonic regression**, a parameterization of linear regression that incorporates sinusoids as independent variables. We define the global per-unit surprisal as our dependent variable and BPE-segmented tokens as our base units w_t. Harmonic regression is based on the principle that any periodic function can be approximated using a sum of sine and cosine functions. Beyond detecting periodicity, harmonic regression enables us to explore whether any fluctuations in surprisal we find align with the structural units believed to influence them. For other approaches to surprisal contour modeling, see App. C.

3.1 Harmonic Regression

Harmonic regression models a periodic function $f(t)$ as a linear combination of sine and cosine components at integer multiples of its fundamental frequency:

$$f(t) = \beta_0 + \sum_{k=1}^{K} \left(\beta_{1,k} \cdot \sin\left(\frac{k2\pi t}{T}\right) + \beta_{2,k} \cdot \cos\left(\frac{k2\pi t}{T}\right) \right), \quad (6)$$

where K is the order of the model, i.e., the number of harmonic components, T is the length of the longest period, and $\frac{k}{T}$ is the frequency. The coefficient β_0 controls the vertical shift of the series while $\beta_{1,k}$ and $\beta_{2,k}$ scale the contribution of the sine and cosine of the harmonic component k, yielding the model parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{2K+1}$,

$$\boldsymbol{\beta} = [\beta_0, \beta_{1,1}, \ldots, \beta_{1,K}, \beta_{2,1}, \ldots, \beta_{2,K}]^\top. \quad (7)$$

The amplitude of the k^{th} harmonic component is given by $A_k = \sqrt{\beta_{1,k}^2 + \beta_{2,k}^2}$, capturing the strength of that frequency component. The parameters are then estimated by minimizing the ordinary least-squares objective. Harmonic regression identifies the combination of sinusoids that best predicts the shape of a surprisal curve in a fully unsupervised manner. However, it does not offer a way to examine the influence of linguistic structures (e.g., paragraphs or sentences) that we might a priori expect to be predictive of periodic surprisal patterns. To test hypotheses about such structures, we introduce an additional scaling mechanism in the time domain of the signal.

3.2 Time Scaling

Time scaling adjusts the period of the sinusoids in the harmonic regression to account for the span of different structural elements containing the base unit w_t. This results in a modified summation term for the k^{th} harmonic component (cf. Eq. (6)):

$$\beta_{1,k} \cdot \sin\left(\frac{k2\pi t}{U_t}\right) + \beta_{2,k} \cdot \cos\left(\frac{k2\pi t}{U_t}\right), \quad (8)$$

where U_t is the length of the structural unit containing w_t. When considering the entire document $\boldsymbol{w}$, i.e., the coarsest unit containing w_t, and normalizing by its length $|\boldsymbol{w}|$, this corresponds to adjusting periods to the relative rather than absolute position of w_t. This serves as our reference condition.

Time scaling can be applied to linguistic structures of varying granularity, from syntactic constructions and multi-word expressions to larger discourse structures. In the present work, we focus on discourse structural predictors—in particular, paragraphs, sentences, and elementary discourse units (EDUs, i.e., the smallest meaningful units of discourse in rhetorical structure theory; Mann and Thompson, 1988). See §4.1 for more details on these predictors. For every w_t, we scale the periods of the sinusoids by the length U_t (measured as the number of tokens) of w_t's containing paragraph, sentence, and EDU. Furthermore, for each granularity level, we set K to the length of the longest unit in the training set of each cross-validation fold. This ensures that our harmonic components represent periods ranging from one base unit w_t to the full length of the longest unit. For an example of first-order sinusoids scaled to our three discourse structures of interest, see Fig. 1. Time scaling allows us to test our HS hypothesis by observing whether structure-specific sinusoids are significant predictors of surprisal contours.

3.3 Feature Selection and Significance

We fit linear models including baseline features and the aforementioned time-scaled sinusoids using 10-fold cross-validation.[2] We perform feature selection using L_1 regularization (see App. D), and we then refit the model using only the features with non-zero coefficients in the regularized fit. We use one-way ANOVA to assess statistical significance for the remaining features by comparing each harmonic order (sine and cosine) against a baseline model that includes only the non-sinusoid baseline features. For more details, see App. E.

4 Data

We test for periodic structure in surprisal contours across six languages: English, Spanish, German, Dutch, Basque, and Brazilian Portuguese. Surprisal contours are obtained using Transformer-based language models as estimators.

4.1 Datasets

Following Tsipidi et al. (2024), we base our analysis on datasets annotated according to Rhetorical Structure Theory (RST), a widely recognized formalism for analyzing discourse structure which originated from early work in text generation and later developed into a linguistic theory. We use six RST-annotated corpora covering English, Spanish, German, Dutch, Basque, and Brazilian Portuguese (Carlson et al., 2001; Carlson and Marcu, 2001; da Cunha et al., 2011; Stede, 2004; Stede and Neumann, 2014; van der Vliet et al., 2011; Redeker et al., 2012; Iruskieta et al., 2013; Cardoso et al., 2011). RST segments texts into recursively nested spans linked by rhetorical relations, and its basic units of analysis are elementary discourse units (EDUs)—the smallest communicative segments in a discourse tree, which convey complete propositions and serve as the fundamental building blocks of larger logical and rhetorical structures.[3]

In addition to RST-based discourse segmentation, we also consider sentences and paragraphs as conventional prose structures. Each dataset is thus processed by segmenting documents into paragraphs, sentences, and EDUs. Tab. 3 summarizes the total counts of these units, while Tab. 4 provides an overview of the mean and median number of EDUs per sentence, paragraph, and document; both tables and additional information on the datasets can be found in App. B.

4.2 Surprisal Estimation

We compute global per-token surprisal according to Eq. (4) for every document in the respective dataset. For each language, we select a dedicated open-weight LLM, fine-tuned on data in that language.[4]

5 The Harmonics in Surprisal Contours

We start by evaluating the overall predictive power of harmonic regression for surprisal contours. We fit five models: one for each type of structure—i.e., sinusoids scaled to the document, EDU, sentence, or paragraph level—and a maximal model that includes sinusoids scaled by all structures simultane-

[2]Baseline features include the number of characters in w_t, previous timestep surprisal $\iota(w_{t-1}; \boldsymbol{w}_{<t-1})$, relative position of w_t in $\boldsymbol{w}$, and boolean feature vectors indicating whether w_t is within windows of 1, 2, and 4 tokens distance from a structural boundary. The latter are included to test whether harmonic features capture periodicity beyond what can be explained by changes in surprisal at structural boundaries.

[3]EDUs often align with clauses, but there are also instances where an EDU may contain more than a single clause, such as clauses where the subject or object of the main verb is also a clause (Carlson and Marcu, 2001).

[4]For an overview of the models, see Tab. 6.

	English	Spanish	German	Dutch	Basque	Brazilian Portuguese
Baseline	$9.91 \pm .43$	$14.63 \pm .47$	$12.43 \pm .25$	$9.32 \pm .79$	$9.00 \pm .55$	$9.62 \pm .81$
Document-scaled	$9.92 \pm .44$	$\mathbf{13.52 \pm .38}$*	$12.29 \pm .26$*	$9.60 \pm .81$	$9.17 \pm .52$	$9.80 \pm .78$
EDU-scaled	$\mathbf{9.46 \pm .40}$*	$13.83 \pm .45$*	$\mathbf{11.31 \pm .29}$*	$\mathbf{8.73 \pm .74}$*	$\mathbf{8.67 \pm .55}$*	$\mathbf{9.07 \pm .83}$*
Sentence-scaled	$9.55 \pm .42$*	$14.17 \pm .45$*	$11.56 \pm .30$*	$8.92 \pm .72$*	$8.85 \pm .56$*	$9.32 \pm .84$*
Paragraph-scaled	$9.73 \pm .43$*	$14.40 \pm .41$*	$12.23 \pm .27$*	$9.31 \pm .75$	$9.24 \pm .53$	$9.55 \pm .85$
All	$9.37 \pm .40$*	$13.09 \pm .35$*	$11.37 \pm .33$*	$9.22 \pm .74$	$9.08 \pm .53$	$9.42 \pm .80$*

Table 1: Mean and standard deviation of validation MSE across ten cross-validation folds for each harmonic regression model and language. Bolded values indicate the lowest MSEs, excluding the maximal (All) model. Asterisks (*) denote models that significantly outperform the baseline according to a one-sided paired t-test ($p < 0.001$); see App. F for further details on significance testing.

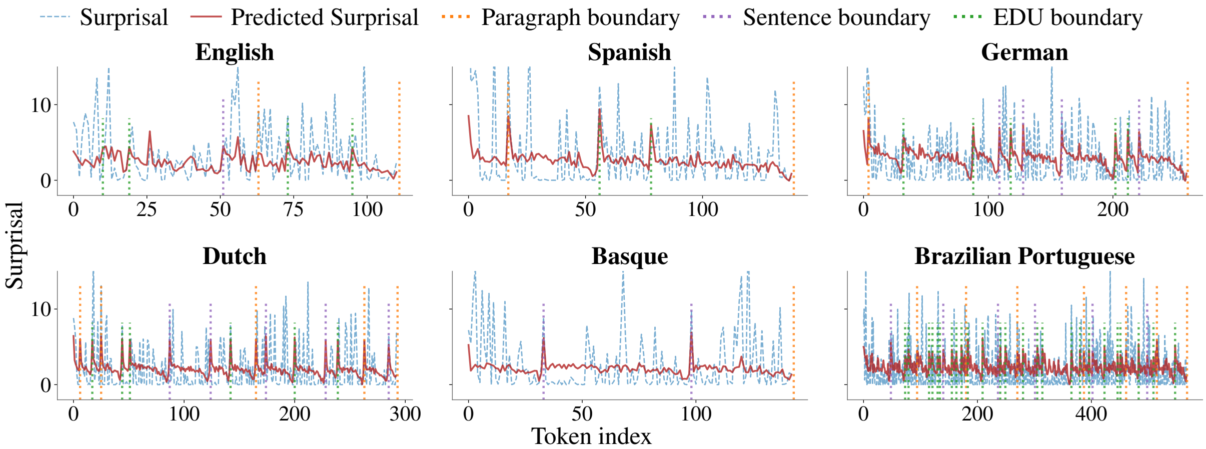

Figure 2: **Predicted vs. Observed Surprisal Curves for EDU-scaled Sinusoids**. Each panel shows predictions for one document: English (`wsj_0605`), Spanish (`as00007`), German (maz-1818), Dutch (`AD14_CarpeDiem`), Basque (GMB0002-GS), Brazilian Portuguese (`D2_C38_Estadao`). For other scaling methods, see App. H.2.

ously. All models include the baseline predictors, as described in §3.3. Predictive power is reported as the mean validation MSE over 10-fold cross-validation, with results summarized in Tab. 1.

We find that scaling periods by EDU length leads to the lowest MSE across all languages except for Spanish. For the other five languages, EDU-scaling yields similar or better MSE than the maximal model (All).[5] We test for statistical significance against a model that includes only baseline predictors, and find that all models incorporating EDU- or sentence-scaled sinusoids significantly outperform this baseline. To further provide a visual assessment of the best model's fit, in Fig. 2, we present the predicted surprisal derived from the EDU-scaled harmonic regression model. The predicted curves reflect the overall pattern of the observed surprisal, aligning particularly closely at unit boundaries.

We note that the differences between models are moderate, and their MSE still leaves considerable room for improvement. However, the HS hypothesis does not posit that periodic regularities stemming from discourse structure are the *sole* determinants of surprisal contours; thus, a perfect fit for every individual surprisal value is not expected. Rather, our goal is to identify periodic regularities and investigate their connection to discourse structural elements.

5.1 Contribution of Individual Sinusoids

To identify which periods most influence the shape of surprisal contours, we analyze the amplitudes estimated by the maximal harmonic regression model. These amplitudes, which reflect the strength of each frequency (see §3.1), highlight the contributions of different harmonic components across structure types. Tab. 2 presents the mean amplitudes of the most dominant sinusoids, averaged across all cross-validation folds. Subscripts denote the number of cross-validation folds in which a sinusoid is significant according to the ANOVA test ($p < 0.001$; see §3.3 and App. E for more details).

[5] In App. G, we repeat the experiments with randomly permuted surprisal values, which results in higher MSE and no notable differences between scaling methods.

English								Spanish							
Document		EDU		Sentence		Paragraph		Document		EDU		Sentence		Paragraph	
k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k
1	0.235_{10}	1	0.370_{10}	4	0.171_{10}	9	0.037_{10}	1	0.422_{10}	1	0.364_{10}			126	0.058_{7}
		2	0.330_{10}	5	0.151_{10}			4	0.323_{10}	2	0.313_{10}			535	0.054_{10}
		3	0.241_{10}	10	0.144_{10}			5	0.293_{10}	4	0.264_{10}			150	0.053_{8}

German								Dutch							
Document		EDU		Sentence		Paragraph		Document		EDU		Sentence		Paragraph	
k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k
4	0.165_{10}	1	0.599_{10}	10	0.101_{10}	11	0.087_{10}	5	0.153_{10}	1	0.470_{10}	3	0.198_{10}	10	0.095_{10}
5	0.140_{10}	2	0.515_{10}	12	0.063_{10}	56	0.077_{10}	6	0.135_{10}	2	0.308_{10}	1	0.141_{10}	66	0.089_{10}
6	0.137_{10}	3	0.380_{10}	13	0.059_{10}	345	0.066_{9}	7	0.103_{6}	3	0.247_{10}			8	0.074_{10}

Basque								Brazilian Portuguese							
Document		EDU		Sentence		Paragraph		Document		EDU		Sentence		Paragraph	
k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k	k	A_k
7	0.099_{10}	1	0.260_{10}	189	0.053_{0}	25	0.066_{10}	24	0.091_{10}	2	0.389_{10}	5	0.129_{10}	27	0.046_{10}
6	0.098_{10}	2	0.196_{10}	112	0.043_{10}	651	0.054_{9}	14	0.065_{4}	3	0.312_{10}	4	0.084_{10}	34	0.041_{10}
8	0.093_{10}	5	0.122_{10}			119	0.041_{7}	20	0.059_{3}	4	0.274_{10}	3	0.082_{10}		

Table 2: Mean amplitudes (A_k) of the three most dominant sinusoids that persist through feature selection in all ten cross-validation folds. Subscripts indicate the number of folds in which each sinusoid is also statistically significant according to the ANOVA test (see §3.3). Mean amplitudes for additional harmonic orders are reported in Tab. 13. Fewer than three values indicate that fewer than three sinusoids persisted through feature selection across all folds.

EDU-scaled sinusoids, particularly those with lower orders ($k \in [1, 2, 3, 4]$), show the highest amplitudes in all languages except Spanish, where they rank second after document-scaled ones. The results for sentence-scaled sinusoids are mixed, while scaling by paragraph length leads to lower amplitudes compared to document scaling. Notably, all EDU-scaled sinusoids that remain after feature selection are consistently significant across folds (see also Tab. 13 for additional harmonics), indicating reliable predictive strength—an effect not seen for any other structure type.

Overall, our results reveal periodicity in surprisal contours, particularly at the EDU level. This yields evidence that EDUs play an important role in determining the information structure of discourse, corroborating results in (Tsipidi et al., 2024) while refining the form of the functional relationship between a unit's information and its position.

6 Surprisal at Discourse Unit Boundaries

So far, we have observed significant periodicity in surprisal contours at the EDU level, with predicted harmonic curves closely aligning with discourse boundaries—points where surprisal tends to peak.

Moreover, boundary features emerge as the strongest individual predictors of surprisal, exhibiting the highest coefficients in both our baseline and maximal linear models (see Tab. 7 and Tab. 8). These findings motivate a closer examination of the relationship between surprisal peaks, discourse unit boundaries, and periodicity. Specifically, we aim to understand the extent to which the contribution of harmonic components to the shape of surprisal contours is explained by their alignment with boundary peaks, as opposed to reflecting additional structure in the distribution of information—the latter interpretation being supported by the significance of harmonic component effects even after accounting for baseline predictors (as shown in §5.1).

To investigate the relationship between surprisal peaks and unit boundaries, we calculate the mean surprisal of tokens within one- and two-token windows immediately *before* and *after* paragraph, sentence, and EDU boundaries, and compare these values to the mean surprisal of tokens located farther from any boundary (i.e., all other tokens). We find that tokens preceding boundaries exhibit lower mean surprisal compared to non-boundary tokens, while those following boundaries considerably exhibit higher surprisal (see Tab. 12 for a comparison of surprisal values before, after, and away from boundaries). For example, in German, the single token before a paragraph boundary has a mean surprisal of 1.47 ± 2.69, while the one after has 7.75 ± 5.31. At sentence and EDU boundaries, surprisal shifts from 1.07 ± 1.77 to 7.06 ± 3.73 and from 1.30 ± 1.87 to 6.39 ± 3.79, respectively. This overall

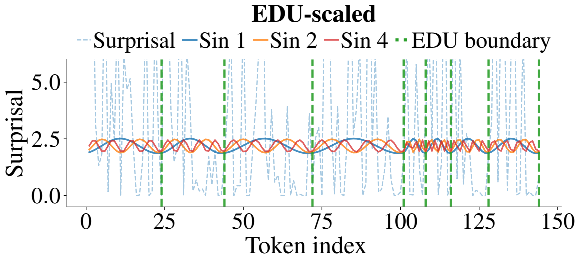

Figure 3: **Harmonic Structure in Surprisal Contours**. Top three most dominant sinusoids (EDU-scaled) in the maximal model for a Spanish text (doc `ma00059a`). Amplitudes signify the contribution to the overall variation, with higher amplitudes indicating a larger effect.

trend holds across languages and boundary types.

To explore how this relates to periodicity, we focus on EDU boundaries, since they are the most predictive structure and subsume both sentence and paragraph boundaries. We visualize surprisal contours alongside unit boundaries and the highest-amplitude sinusoids. Fig. 3 shows this for a Spanish document, with further examples for other languages in App. H.1. These visualizations reveal that the most prominent EDU-scaled sinusoids tend to intersect discourse boundaries at their troughs—points where the curve declines prior to the boundary and rises immediately after. Taken together with the results presented in §5, our findings suggest that information is systematically modulated around discourse boundaries—decreasing before and increasing after—and that surprisal exhibits periodicity which not only reflects this modulation but also extends beyond the immediate influence of boundary proximity.

These findings naturally give rise to further questions about the underlying mechanisms: Why is information organized to peak at discourse boundaries? And why is EDU-specific periodicity particularly prominent? We can offer some preliminary speculations. Prior work on UID assumes a constant channel capacity, but there is substantial evidence that processing effort increases toward the end of syntactic phrases—due to greater memory demands and integration costs (Just and Carpenter, 1980; Gibson, 1998, 2000; Keith Rayner and Duffy, 2000)—as well as toward the end of narrative events (Speer and Zacks, 2005; Radvansky and Copeland, 2010). These findings imply that channel capacity decreases at the end of units and increases at their beginning. An optimal speaker would therefore modulate the information rate across transitions. Furthermore, context informativity tends to drop at the beginning of new structural units, as these often introduce new referents, topics, or discourse relations (Genzel and Charniak, 2002). This, too, would result in increased surprisal around boundaries. Nevertheless, the presence of numerous higher-frequency sinusoids with comparatively high amplitudes suggests the existence of meaningful structure at sub-EDU levels. Investigating the role of these finer-grained units is a promising direction for future research.

7 Conclusion

We proposed a refinement of two hypotheses that aim to explain the structure of global information contours: the established Uniform Information Density hypothesis, which is agnostic to the nature and granularity of the units and structures being analyzed, and the more recent Structured Context hypothesis, which links per-unit surprisal rates to the unit's position within the discourse structure. The Harmonic Surprisal hypothesis gives more color to the phenomena that are not explained by the UID hypothesis, which remains neutral on why surprisal deviates from the mean (or the upper bound of channel capacity). It also makes stronger predictions than the SC hypothesis by introducing a periodic functional relationship between surprisal and discourse constituents.

Our harmonic regression analysis of surprisal contours across six languages reveals clear periodic patterns, especially under EDU-based time scaling. Within discourse units, information is not evenly distributed: surprisal reliably decreases before and increases after discourse boundaries. Although our focus is on global discourse-level trends, we also detect significant higher-order components in surprisal contours, indicating the presence of periodicity at smaller scales—such as syntactic units, words, with their subword tokens. Future work could apply our time-scaled harmonic regression framework to these finer-grained structures to investigate their potential role in shaping surprisal dynamics.

In conclusion, this contribution furthers the broader program of understanding the processing, aesthetic, and conventional constraints that shape the linguistic exchange of information.

Limitations

One limitation of our study is the diversity of languages on which we test our hypotheses. Although Basque offers a typologically distinct language with a subject-object-verb (SOV) word order, the remaining languages are Indo-European, with a subject-verb-object (SVO) order. Future work could broaden the scope by including a more diverse set of languages. Moreover, as already acknowledged in §7, while our methodology is, in principle, applicable to structural elements at any scale, this study focused specifically on discourse-level constituents: EDUs, sentences, and paragraphs. Future work could extend our investigation by applying time-scaled harmonic regression to smaller linguistic structures—such as syntactic constructions, collocations, or even individual words—to explore whether similar periodic patterns emerge at those levels. Similarly, surprisal contours could be examined using base units of varying granularity, ranging from coarser units like full clauses or sentences to more fine-grained levels such as individual characters.

Ethics Statement

We foresee no ethical problems with our work.

Acknowledgements

We would like to thank Chloé Braud for putting us in touch with van der Vliet et al. (2011) and facilitating access to the Dutch RST corpus. Mario Giulianelli was supported by an ETH Zürich Postdoctoral Fellowship.

References

Christoph Aurnhammer and Stefan L. Frank. 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

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. PMID: 15298329.

Matthew Aylett and Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Matthew P. Aylett. 1999. Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 289–292. American Institute of Physics.

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.

Claire Augusta Bergey and Simon DeDeo. 2024. From "um" to "yeah": Producing, predicting, and regulating information flow in human conversation.

Paula C. F. Cardoso, Erick G. Maziero, Mara Luca Castro Jorge, Eloize R. M. Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago A. S. Pardo. 2011. Cstnews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Raoul Cervantes and Glenn Gainer. 1992. The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, 26(4):767–770. Wiley.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2).

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43:651–681.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211. Wiley Online Library.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque.

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

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

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.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Richard Futrell. 2023. Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, 120(39):e2220593120.

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.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain/MIT Press*.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024a. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024b. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669, Miami, Florida, USA. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. 2024c. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. 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.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Eva Huber, Sebastian Sauppe, Arrate Isasi-Isasmendi, Ina Bornkessel-Schlesewsky, Paola Merlo, and Balthasar Bickel. 2024. Surprisal from language models can predict ERPs in processing predicate-argument structures only if enriched by an agent preference principle. *Neurobiology of Language*, 5(1):167–200.

Mikel Iruskieta, María J. Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and

Oier Lopez de la Calle. 2013. The rst basque treebank: An online search interface to check rhetorical relations. In *Proceedings of the 4th Workshop on RST and Discourse Studies*.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Gretchen Kambe Keith Rayner and Susan A. Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, 53(4):1061–1080. PMID: 11131813.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain. Association for Computational Linguistics.

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Edwin Herbert Lewis. 1894. *The history of the English paragraph*. University of Chicago Press.

Jiaxuan Li and Richard Futrell. 2024. An information-theoretic model of shallow and deep language comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Eliot Maës, Philippe Blache, and Leonor Becerra-Bonache. 2022. Shared knowledge in natural conversations: can entropy metrics shed light on information transfers? In *26th Conference on Computational Natural Language Learning*, pages 213–227.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. *Linguistic Data Consortium*, 14.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498, Miami, Florida, USA. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Ting Qian and T. Florian Jaeger. 2011. Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Milena Rabovsky, Steven S Hansen, and James L McClelland. 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Gabriel A. Radvansky and David E. Copeland. 2010. Reading times and the detection of event shift processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1):210–216. Place: US Publisher: American Psychological Association.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a Dutch text corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

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

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92 – 96.

Sophie Slaats, Antje S. Meyer, and Andrea E. Martin. 2024. Lexical surprisal shapes the time course of syntactic structure building. *Neurobiology of Language*, 5(4):942–980.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Nicole K. Speer and Jeffrey M. Zacks. 2005. Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language*, 53(1):125–140.

Manfred Stede. 2004. The Potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128, London, UK. Association for Computational Linguistics.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

N.H. van der Vliet, I. Berzlánovich, G. Bouma, M. Egg, and G. Redeker. 2011. Building a discourse-annotated dutch text corpus. In *Beyond Semantics*, volume 3 of *Bochumer Linguistische Arbeitsberichte*, pages 157 – 171. Ruhr-Universität Bochum. 2011/g.bouma/pub005, 2011/g.redeker/pub001 E-publication only.

Bram Vanroy. 2024. Geitje 7b ultra: A conversational model for dutch.

Vivek Verma, Nicholas Tomlin, and Dan Klein. 2023. Revisiting entropy rate constancy in text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15537–15549, Singapore.

Tim Vieira, Ben LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2024. From language models over tokens to language models over characters.

Sarenne Wallbridge, Peter Bell, and Catherine Lai. 2022. Investigating perception of spoken dialogue acceptability through surprisal. In *Interspeech 2022: The 23rd Annual Conference of the International Speech Communication Association*, pages 4506–4510. International Speech Communication Association.

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

Ethan Gotlieb 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.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721, Singapore. Association for Computational Linguistics.

Yang Xu and David Reitter. 2016. Entropy converges between dialogue participants: Explanations from an information-theoretic perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–546, Berlin, Germany. Association for Computational Linguistics.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. 2024. Detecting subtle differences between human and model languages using spectrum of relative likelihood. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10121, Miami, Florida, USA. Association for Computational Linguistics.

Zuhao Yang, Yingfang Yuan, Yang Xu, Shuo Zhan, Huajun Bai, and Kefan Chen. 2023. Face: Evaluating natural language generation with Fourier analysis of cross-entropy. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 17038–17056.

Jun Sen Yee, Mario Giulianelli, and Arabella J. Sinclair. 2024. Efficiency and effectiveness in task-oriented dialogue: On construction repetition, information rate, and task success. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5562–5577, Torino, Italia. ELRA and ICCL.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Oxford, England.

A Reproducibility

We conduct sentence segmentation on the Spanish RST Discourse Treebank with the `mediacloud`[6] text-to-sentence splitter and on the German Potsdam Commentary Corpus with `wtpsplit`[7] (Minixhofer et al., 2023; Frohmann et al., 2024). We recover text boundaries for the English RST Discourse Treebank using the corresponding texts in the Penn Treebank (Marcus et al., 1999).

To estimate surprisal for the English RST Discourse Treebank and the Spanish RST Discourse Treebank, we follow Tsipidi et al. (2024) and use an RTX 4090 GPU with VRAM 24GB and additional RAM of 64GB for approximately 6 hours; for the Brazilian Portuguese CST-News corpus, we use the same setup for 34 minutes. For the German Potsdam Commentary Corpus, the Dutch corpus, and the RST Basque Treebank, we run inference on an RTX A6000 GPU for circa 3 hours.

Our harmonic regression experiments are implemented with the Statsmodels package[8] (Skipper Seabold and Josef Perktold, 2010). They amount to approximately 37 days of compute time on CPU with 256 GB RAM (without a GPU).

B Datasets & Models

For English, we use the RST Discourse Treebank (Carlson et al., 2001; Carlson and Marcu, 2001), which consists of 347 documents from the Wall Street Journal. For Spanish, we use the Spanish RST Discourse Treebank (da Cunha et al., 2011), containing 267 documents from a variety of domains, including astrophysics, mathematics, and law.[9] For German, we turn to the Potsdam Commentary Corpus 2.0 (Stede, 2004; Stede and Neumann, 2014), which consists of 176 documents annotated with 30 discourse relations under the RST framework. The Dutch corpus (van der Vliet et al., 2011; Redeker et al., 2012) contains 80 documents from various sources such as science news, encyclopedias, fundraising letters, and commercial advertisements, annotated with 31 relations. We exclude documents with overlapping EDUs, reducing the dataset to 62 documents. For Basque, we process data from the RST Basque TreeBank[10] (Iruskieta et al., 2013), consisting of 88 abstracts from medicine, terminology, and science articles, which are annotated with 31 relations. Finally, the CST-News corpus (Cardoso et al., 2011) includes 140 Brazilian Portuguese news documents annotated with 31 relations. We remove 14 documents that have a mismatch between the raw text and the RSTs, leading to a total of 126 documents.

Tab. 3 presents the number of documents, paragraphs, sentences, and EDUs for each language, while Tab. 4 and Tab. 5 provide the token counts per EDU, sentence, and paragraph. These values vary considerably between languages. For example, English has a median of 11 tokens per EDU, compared to a higher median of 26 tokens in Basque.

Dataset	#Docs	#Pars	#Sents	#EDUs
English	347	3511	7012	19443
Spanish	256	963	2065	3146
German	176	531	2097	3018
Dutch	62	371	1310	1761
Basque	88	198	1413	2509
Brazilian Portuguese	126	927	1815	4847

Table 3: Number of documents, paragraphs, sentences, and EDUs for each dataset.

Additionally, Tab. 6 shows the respective model used to estimate surprisal in each language.

C Other Approaches to Surprisal Contour Modeling

Several methods have been proposed for modeling surprisal contours in discourse, each offering different advantages in analyzing structural patterns. Here, we discuss three common approaches and contrast them with our use of harmonic regression. In particular, we highlight how different approaches handle positional predictors and their role in shaping surprisal curves.

Discrete Fourier Transform. Other studies analyzing the surprisal contours of discourse in the frequency domain usually decompose surprisal sequences using the discrete Fourier transform (DFT). This yields coefficients for the whole frequency band which can be useful, e.g., for comparing the frequency profiles of human and language model-generated text (Yang et al., 2023; Xu et al., 2024). Since we focus on surprisal curves of human discourse, we opted instead for harmonic regression, which easily allows us to identify the most signifi-

[6] https://github.com/mediacloud/sentence-splitter.

[7] https://github.com/segment-any-text/wtpsplit.

[8] https://www.statsmodels.org.

[9] After removing 11 documents with missing nodes, we retain 256 documents.

[10] https://ixa2.si.ehu.eus/diskurtsoa/en/

	English			German			Spanish		
Category	Mean	Variance	Median	Mean	Variance	Median	Mean	Variance	Median
Tokens per EDU	12.87	9.13	11	19.30	11.29	17	30.04	20.75	25
Tokens per sentence	35.67	18.94	33	27.78	17	25	45.76	29.18	40
Tokens per paragraph	71.25	40.76	65	109.70	106.02	91	98.23	104.05	62
Tokens per document	720.89	622.55	523	330.98	29.40	333	369.15	291.28	301
EDUs per sentence	2.77	1.78	2	1.44	0.74	1	1.52	0.94	1
EDUs per paragraph	5.54	3.58	5	5.68	5.07	4	3.27	3.13	2
EDUs per document	56.03	51.41	40	17.15	3.06	17	12.29	9.68	10
Sentences per paragraph	2	1.17	2	3.95	3.32	3	2.15	1.78	1
Sentences per document	20.21	17.93	14	11.91	2.45	12	8.07	5.99	6
Paragraphs per document	10.12	7.65	8	3.02	1	3	3.76	3.26	2

Table 4: Number of tokens, EDUs, sentences, and paragraphs per unit for English, German, and Spanish.

	Basque			Brazilian Portuguese			Dutch		
Category	Mean	Variance	Median	Mean	Variance	Median	Mean	Variance	Median
Tokens per EDU	31.34	22.93	26	16.34	10.41	14	20.08	10.92	18
Tokens per sentence	55.65	34.84	49	43.65	24.66	40	26.99	14.96	25
Tokens per paragraph	397.15	407.25	233	85.46	73.58	74	95.31	77.43	91
Tokens per document	893.59	396.58	903	638.86	301.70	610	570.34	127.63	572
EDUs per sentence	1.78	1.13	1	2.67	1.85	2	1.34	0.63	1
EDUs per paragraph	12.67	12.09	8	5.23	5.16	4	4.75	3.53	4
EDUs per document	28.51	14.37	27	39.09	20.01	35	28.40	5.76	28
Sentences per paragraph	7.14	6.47	5	1.96	1.68	2	3.53	2.45	3
Sentences per document	16.06	7.69	16	14.64	7.93	13	21.13	4.63	21
Paragraphs per document	2.25	2.12	1	7.48	3.90	7	5.98	1.99	6

Table 5: Number of tokens, EDUs, sentences, and paragraphs per unit for Basque, Brazilian Portuguese, and Dutch.

cant frequencies of the signal through significance testing.

Contour Standardization. Before applying the DFT, Xu et al. (2024) additionally standardize the surprisal values (centering values around the mean with unit standard deviation) to facilitate comparisons between humans and different LLMs. We do not apply standardization because our goal is not to compare surprisal curves across models. Moreover, we find that standardization hinders the discovery of significant frequency components.

Linear Modeling. Early studies of surprisal contours model the surprisal of a sentence as a linear function of its absolute position with the help of n-gram models (Genzel and Charniak, 2002; Keller, 2004) or, more recently, transformers (Giulianelli and Fernández, 2021). To account for the effects of sentence length, some studies additionally divide by the average surprisal of all the sentences of equal length (Genzel and Charniak, 2003; Xu and Reitter, 2016, 2018), an adjustment deemed less crucial when dealing with word or subword tokens (Verma et al., 2023). In studies where contextual structures are considered, predictors are either the absolute position of the unit within its containing structure (Giulianelli et al., 2021; Maës et al., 2022), or the unit's relative position (Tsipidi et al., 2024). In the present work, we apply harmonic regression—a variant of linear regression—because it allows us to model surprisal contours as a sum of sinusoidal components.

D Regularization

To perform feature selection, we use L_1 regularization with a penalty weight[11] of $\lambda = 0.01$, chosen by the lowest mean-squared error (MSE) on one cross-validation fold for each of the six corpora. The Brazilian Portuguese corpus is the exception, with the optimal weight for the cross-validation fold being $\lambda = 0.1$; however, the difference in MSE was minimal (8.988 for $\lambda = 0.1$ and 9.057 for $\lambda = 0.01$) and we opted to train on it with the

[11] The Statsmodels package documentation refers to this penalty weight as α, but we refer to it as λ to prevent confusion with the significance level α.

Language	Model	Source
English	NOUS-YARN-LLAMA-2-7B-64K (Peng et al., 2024)	https://huggingface.co/NousResearch/Yarn-Llama-2-7b-64k
Spanish	LINCE MISTRAL 7B INSTRUCT	https://huggingface.co/clibrain/lince-mistral-7b-it-es
German	LAION LEOLM 7B	https://huggingface.co/LeoLM/leo-hessianai-7b
Basque	LATXA 7B (Etxaniz et al., 2024)	https://huggingface.co/HiTZ/latxa-7b-v1.2
Dutch	GEITJE 7B ULTRA (Vanroy, 2024)	https://huggingface.co/BramVanroy/GEITje-7B-ultra
Brazilian Portuguese	SABIÁ-7B (Pires et al., 2023)	https://huggingface.co/maritaca-ai/sabia-7b

Table 6: Language models used to estimate surprisal for the six languages analyzed in this study.

same λ as the other corpora for consistency.

E Baseline

We set up our baseline as a linear regression model trained on intercept, length of the base unit w_t, i.e., number of characters in the BPE token, surprisal of previous unit $\iota(w_{t-1}; \boldsymbol{w}_{<t-1})$, relative position of w_t in the document, and boolean feature vectors that indicate whether w_t is within windows of 1, 2, and 4 tokens distance from a structural boundary. Against this baseline, we run one-way ANOVA with additional models trained simultaneously on baseline features and each order of harmonic components, setting the significance level $\alpha = 0.001$. We choose a low threshold to ensure we retain features that are highly statistically significant.

E.1 Baseline Coefficients in the Maximal Model

Tab. 7 shows the coefficients (β) of the baseline features in the maximal model. Among these features, the Boolean boundary flags (Boundary $\pm$ 1) consistently have the highest coefficients.

E.2 Baseline Coefficients in the Baseline Model

In Tab. 8, we compare the coefficients (β) of the baseline features in the baseline model. Across languages, the Boolean feature vectors of size 1 (Boundary $\pm$ 1) around the token consistently have the largest coefficient magnitude.

F Significance Testing

In Tab. 9, we report the significance of the MSE reduction over the baseline for each scaling method and language using a paired, one-sided t-test. For each language, we have ten validation folds, hence ten paired ($n = 10$) observations between the baseline and each scaling method. We test whether each scaling method reduces the MSE with a one-sided paired t-test[12] (H_0 : $\mu_{\text{scaled}} \geq \mu_{\text{base}}$, H_1 : $\mu_{\text{scaled}} < \mu_{\text{base}}$). To account for the multiple comparisons, we we the Holm–Bonferroni correc-

English		**Spanish**	
Feature	β	Feature	β
Boundary $\pm$ 1	0.817_{10}	Boundary $\pm$ 1	3.976_{10}
Boundary $\pm$ 2	0.300_{10}	Boundary $\pm$ 2	0.625_{10}
Boundary $\pm$ 4	-	Boundary $\pm$ 4	0.201_{10}
Previous Surprisal	-	Previous Surprisal	0.026_{10}
Relative Position	-0.708_{10}	Relative Position	-1.050_{10}
Token Length	0.390_{10}	Token Length	0.283_{10}
German		**Dutch**	
Feature	β	Feature	β
Boundary $\pm$ 1	1.903_{10}	Boundary $\pm$ 1	2.883_{10}
Boundary $\pm$ 2	-	Boundary $\pm$ 2	-
Boundary $\pm$ 4	-	Boundary $\pm$ 4	-
Previous Surprisal	-	Previous Surprisal	0.028_{10}
Relative Position	-0.350_{10}	Relative Position	-0.287_{10}
Token Length	0.176_{10}	Token Length	0.109_{10}
Basque		**Brazilian Portuguese**	
Feature	β	Feature	β
Boundary $\pm$ 1	3.699_{10}	Boundary $\pm$ 1	0.982_{10}
Boundary $\pm$ 2	-0.639_{10}	Boundary $\pm$ 2	0.468_{10}
Boundary $\pm$ 4	-0.152_{10}	Boundary $\pm$ 4	-
Previous Surprisal	-	Previous Surprisal	-0.015_{10}
Relative Position	-	Relative Position	-
Token Length	0.230_{10}	Token Length	0.298_{10}

Table 7: Mean coefficients β of the baseline predictors in the maximal model (All). Subscripts denote the number of cross-validation folds (out of a total of ten folds) where the sinusoids are significant in the ANOVA. Missing coefficient values (-) indicate features that do not persist through feature selection in all ten cross-validation folds.

[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html.

English		Spanish	
Feature	β	Feature	β
Boundary $\pm$ 1	0.899_9	Boundary $\pm$ 1	2.713_{10}
Boundary $\pm$ 2	-	Boundary $\pm$ 2	0.378_{10}
Boundary $\pm$ 4	-	Boundary $\pm$ 4	-0.254_{10}
Previous Surprisal	-	Previous Surprisal	0.047_{10}
Relative Position	-0.415_9	Relative Position	1.356_{10}
Token length	0.401_8	Token length	0.315_{10}
German		**Dutch**	
Feature	β	Feature	β
Boundary $\pm$ 1	1.680_9	Boundary $\pm$ 1	2.082_{10}
Boundary $\pm$ 2	-	Boundary $\pm$ 2	-
Boundary $\pm$ 4	-0.189_9	Boundary $\pm$ 4	-
Previous Surprisal	0.079_8	Previous Surprisal	0.083_9
Relative Position	-0.980_9	Relative Position	-0.688_{10}
Token length	0.248_8	Token length	0.140_8
Basque		**Brazilian Portuguese**	
Feature	β	Feature	β
Boundary $\pm$ 1	2.677_{10}	Boundary $\pm$ 1	1.003_{10}
Boundary $\pm$ 2	-0.686_{10}	Boundary $\pm$ 2	0.196_{10}
Boundary $\pm$ 4	-0.190_7	Boundary $\pm$ 4	-
Previous Surprisal	0.054_9	Previous Surprisal	-
Relative Position	-0.365_{10}	Relative Position	-0.301_{10}
Token length	0.267_{10}	Token length	0.318_{10}

Table 8: Mean coefficients (β) of the baseline features in the baseline model. Subscripts denote the number of cross-validation folds (out of a total of ten folds) where the sinusoids are significant in the ANOVA. Missing coefficient values (-) indicate features that do not persist through feature selection in all ten cross-validation folds.

tion[13] (Holm, 1979) and report both original and corrected p-values in Tab. 9.

G Permuted Surprisal

We replicate the experiments from §5, this time with randomly permuted surprisal values. As shown in Tab. 10, MSE values are generally higher compared to the results in Tab. 1. Additionally, we observe no notable difference between different scaling methods.

H Visualizations

Here, we present additional visualizations showing the contribution of individual sinusoids in the maximal model and how well our models recover the observed surprisal curves under different scaling methods.

H.1 Sinusoid Visualizations

In Fig. 4 and Fig. 5, we present visualizations of surprisal contours, unit boundaries, and the three most dominant sinusoids for individual documents in all languages. For sinusoids, higher amplitudes correspond to a greater effect on the shape of the surprisal contour. Note that the contribution of individual sinusoids is relatively small. For combined predictions across different settings, see App. H.2.

H.2 Prediction Visualizations

Similar to the predicted curves based on EDU-scaled sinusoids in Fig. 2, we visualize the predicted curves for each scaling method in Fig. 6.

I Sinusoid Amplitudes

In Tab. 11, we report the total number of sinusoids and the number that remain in all folds after regularization. Sinusoids that persist across all folds exhibit higher mean amplitudes (A_k) than those excluded in some folds. Additionally, in Tab. 13, we show the twenty-five sinusoids with the highest amplitudes that persist in all folds after L_1 regularization. Notably, among all settings, only the EDU-scaled sinusoids are significant across all folds in the ANOVA.

J Surprisal at Boundaries

We report the mean and standard deviation of token surprisal immediately before and after paragraph, sentence, and EDU boundaries. As shown in Tab. 12, surprisal tends to be lower before and higher after each type of boundary relative to tokens located farther from any boundary. Here, a window size of 1 means we include only the single token immediately before or after the boundary. Non-boundary tokens exclude any tokens within one position on either side of a boundary.

[13] https://www.statsmodels.org/dev/generated/statsmodels.stats.multitest.multipletests.html.

Language	Setting	ΔMSE ↓	p_{raw}	p_{Holm}
English	All	0.54*	5.61e-12	1.63e-10
English	Document-scaled	0.00	0.727	1
English	EDU-scaled	0.45*	2.77e-11	7.47e-10
English	Paragraph-scaled	0.19*	8.04e-10	1.85e-08
English	Sentence-scaled	0.36*	1.61e-11	4.5e-10
Spanish	All	1.54*	1.01e-07	1.72e-06
Spanish	Document-scaled	1.11*	1.92e-06	2.76e-05
Spanish	EDU-scaled	0.80*	2.2e-09	4.83e-08
Spanish	Paragraph-scaled	0.23*	1.1e-05	0.000131
Spanish	Sentence-scaled	0.46*	1.17e-08	2.35e-07
German	All	1.05*	4.05e-10	9.71e-09
German	Document-scaled	0.14*	1.81e-05	0.0002
German	EDU-scaled	1.11*	3.36e-12	1.01e-10
German	Paragraph-scaled	0.19*	1.84e-06	2.76e-05
German	Sentence-scaled	0.86*	7.35e-11	1.84e-09
Dutch	All	0.10	0.0271	0.217
Dutch	Document-scaled	-0.28	1	1
Dutch	EDU-scaled	0.59*	1.91e-08	3.63e-07
Dutch	Paragraph-scaled	0.01	0.319	1
Dutch	Sentence-scaled	0.40*	1.99e-06	2.76e-05
Basque	All	-0.09	0.975	1
Basque	Document-scaled	-0.17	1	1
Basque	EDU-scaled	0.33*	2.31e-09	4.85e-08
Basque	Paragraph-scaled	-0.24	1	1
Basque	Sentence-scaled	0.15*	5.62e-07	8.99e-06
Brazilian Portuguese	All	0.21*	9.79e-05	0.000979
Brazilian Portuguese	Document-scaled	-0.17	1	1
Brazilian Portuguese	EDU-scaled	0.56*	3.25e-11	8.44e-10
Brazilian Portuguese	Paragraph-scaled	0.08	0.00641	0.0577
Brazilian Portuguese	Sentence-scaled	0.31*	9.47e-08	1.7e-06

Table 9: One-sided paired t-test ("greater") comparing each scaling method to the baseline. Holm-adjusted p values control the family-wise error rate. Asterisks (*) indicate $p < .001$.

	English	Spanish	German	Dutch	Basque	Brazilian Portuguese
Document-scaled	$10.81 \pm .45$	$15.60 \pm .50$	$12.91 \pm .28$	$9.88 \pm .85$	$9.36 \pm .55$	$10.02 \pm .84$
EDU-scaled	$10.80 \pm .45$	$15.51 \pm .48$	$12.86 \pm .27$	$9.78 \pm .85$	$9.32 \pm .56$	$9.98 \pm .84$
Sentence-scaled	$10.81 \pm .45$	$15.51 \pm .48$	$12.86 \pm .27$	$9.79 \pm .84$	$9.33 \pm .56$	$9.98 \pm .84$
Paragraph-scaled	$10.81 \pm .45$	$15.53 \pm .48$	$12.89 \pm .27$	$9.84 \pm .85$	$9.36 \pm .57$	$10.00 \pm .83$
All	$10.81 \pm .45$	$15.65 \pm .49$	$12.96 \pm .27$	$9.97 \pm .84$	$9.41 \pm .56$	$10.06 \pm .84$

Table 10: Mean and standard deviation for 10-fold validation MSEs across scaling settings and languages for permuted surprisal values. We observe no notable differences between different settings.

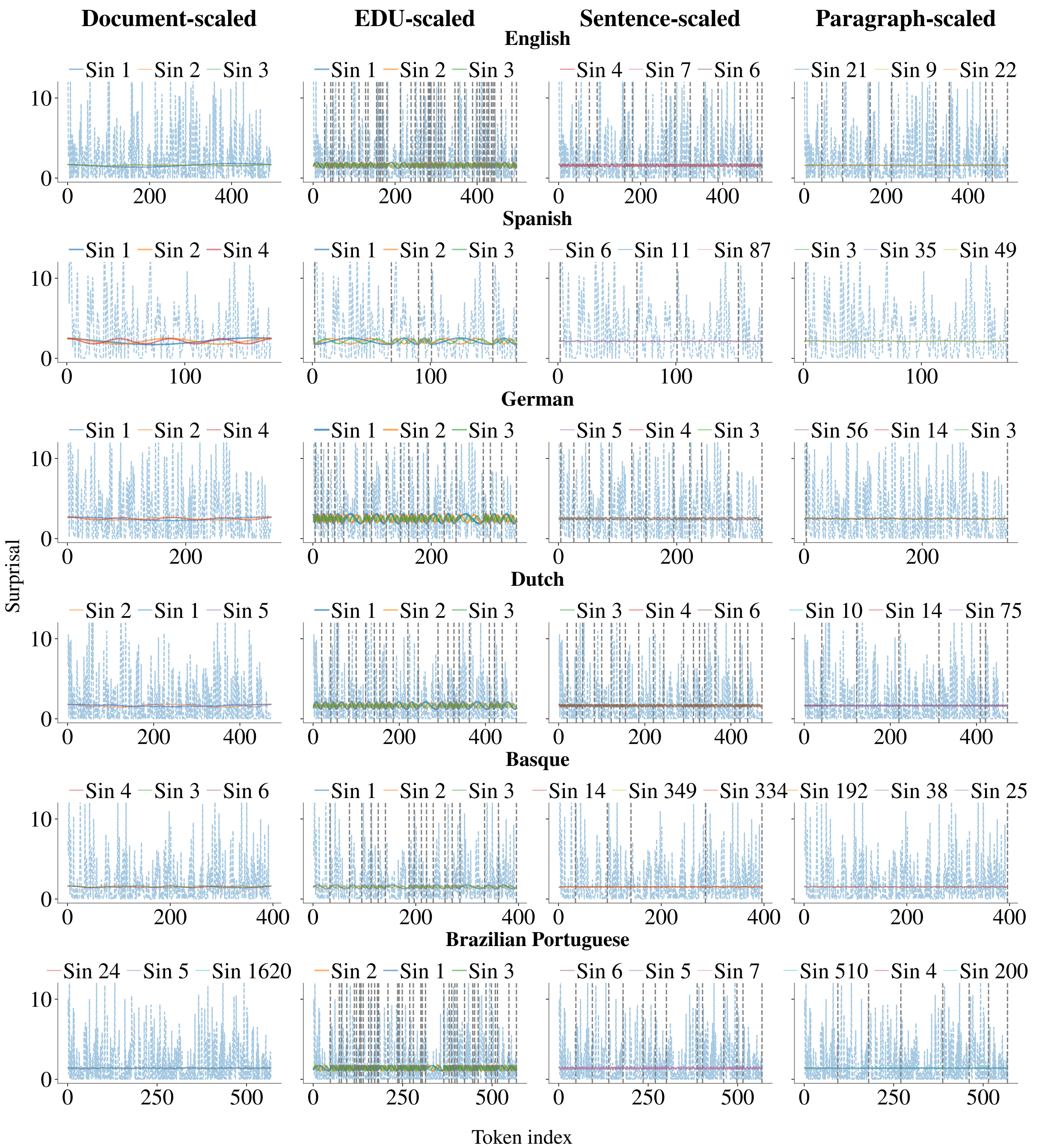

Figure 4: Top three most dominant sinusoids for English (doc `wsj_1111`), Spanish (doc `ec00002`), German (doc `maz-11280`), Dutch (doc `FL13_Unicef`), Basque (doc `TERM29-GS`), Brazilian Portuguese (doc `D2_C38_Estadao`). Amplitudes signify the contribution to the overall variation, with higher amplitudes indicating a larger effect.

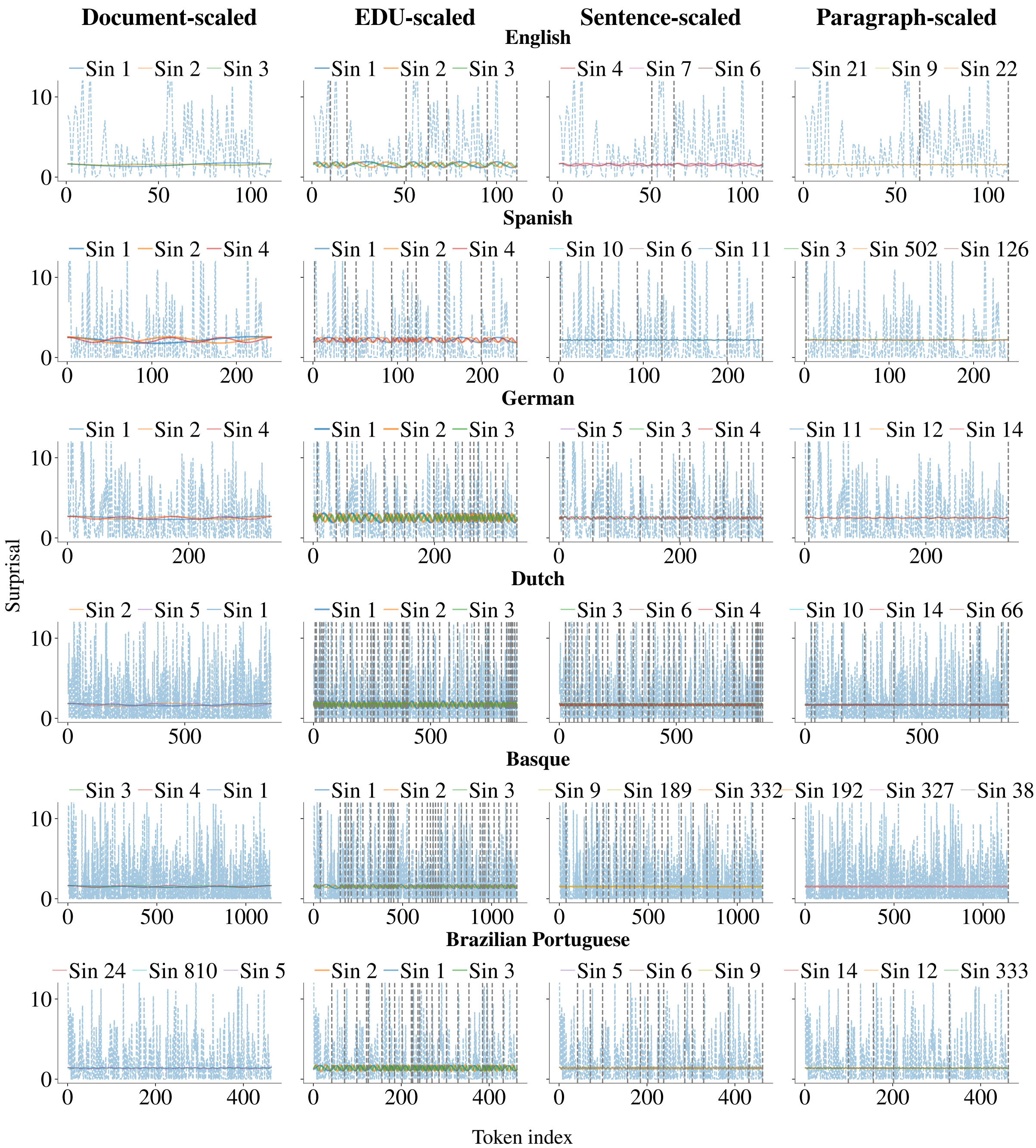

Figure 5: Top three most dominant sinusoids for English (doc `wsj_0605`), Spanish (doc `ec00007`), German (doc `maz-11507`), Dutch (doc `AD02_Atkins`), Basque (doc `GMB0201-GS`), Brazilian Portuguese (doc `D3_C11_OGlobo`). Amplitudes signify the contribution to the overall variation, with higher amplitudes indicating a larger effect.

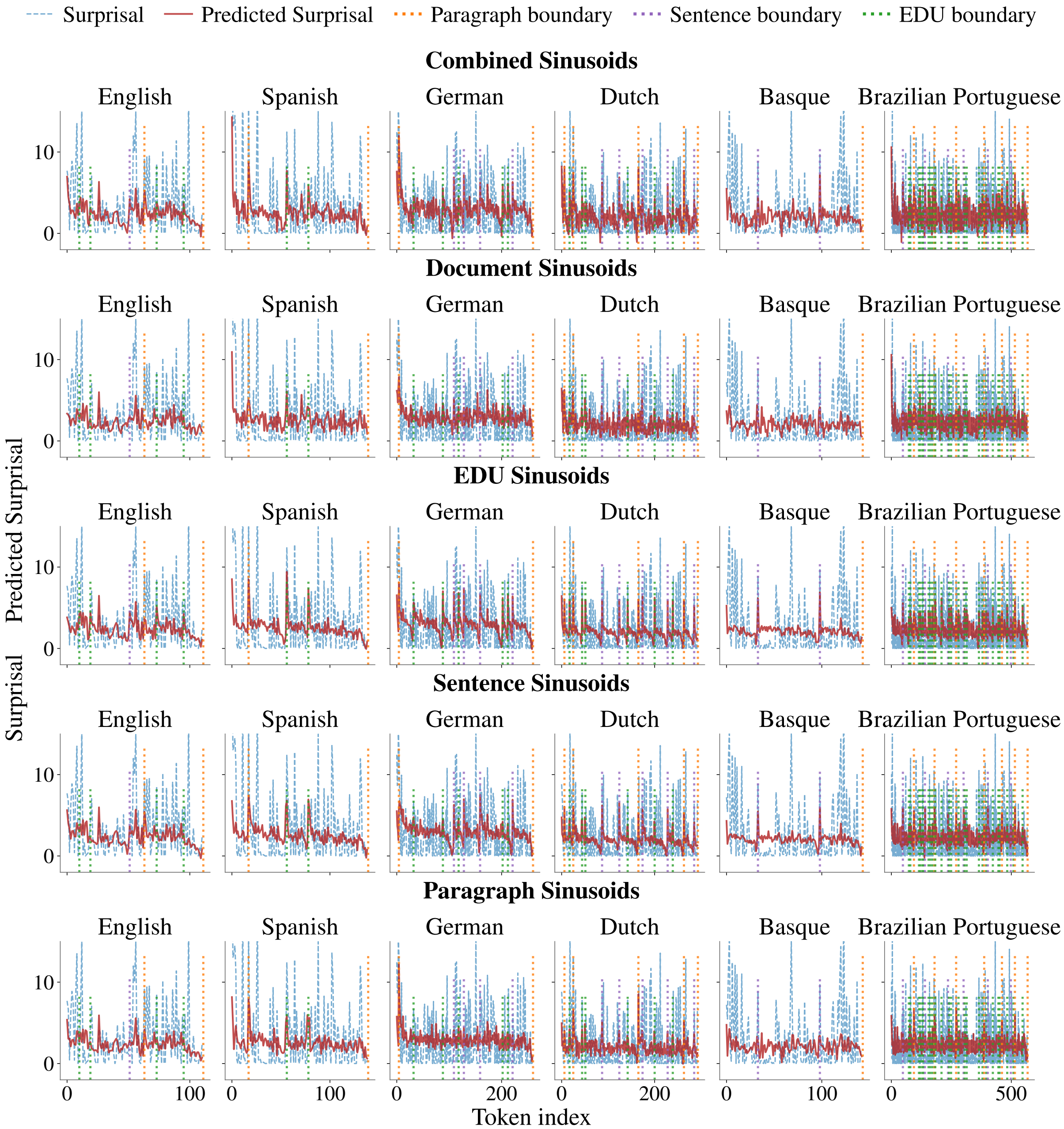

Figure 6: Predicted curves for all languages and all time-scaled sinusoids. Languages correspond to the following documents: English (`wsj_0605`), Spanish (`as00007`), German (maz-1818), Dutch (`AD14_CarpeDiem`), Basque (GMB0002-GS), Brazilian Portuguese (`D2_C38_Estadao`).

Language	Setting	# Sinusoids	# All folds	Mean A_k (All folds)	Mean A_k (Excluded)
English	Document-scaled	474	1	0.235	0.038
English	EDU-scaled	60	6	0.222	0.058
English	Sentence-scaled	63	11	0.099	0.059
English	Paragraph-scaled	112	1	0.037	0.031
Spanish	Document-scaled	1470	80	0.149	0.042
Spanish	EDU-scaled	116	13	0.195	0.050
Spanish	Sentence-scaled	144	0	-	0.041
Spanish	Paragraph-scaled	495	5	0.050	0.040
German	Document-scaled	457	13	0.087	0.048
German	EDU-scaled	59	12	0.230	0.072
German	Sentence-scaled	98	5	0.064	0.061
German	Paragraph-scaled	344	6	0.066	0.046
Dutch	Document-scaled	817	27	0.074	0.044
Dutch	EDU-scaled	64	11	0.180	0.080
Dutch	Sentence-scaled	106	2	0.170	0.056
Dutch	Paragraph-scaled	390	20	0.065	0.047
Basque	Document-scaled	1263	15	0.062	0.038
Basque	EDU-scaled	135	6	0.132	0.060
Basque	Sentence-scaled	248	2	0.048	0.040
Basque	Paragraph-scaled	1233	3	0.054	0.037
Brazilian Portuguese	Document-scaled	1225	13	0.054	0.040
Brazilian Portuguese	EDU-scaled	67	10	0.188	0.087
Brazilian Portuguese	Sentence-scaled	278	6	0.073	0.042
Brazilian Portuguese	Paragraph-scaled	635	2	0.043	0.038

Table 11: Total number of sinusoids and number of sinusoids that remain in all folds after L_1 regularization. Sinusoids that remain in all folds exhibit higher mean amplitudes (A_k) than those excluded in some folds.

Language	Window size	Paragraph Boundary	Sentence Boundary	EDU Boundary	Non-boundary
Before Boundaries					
English	1	1.03 ± 1.44	1.08 ± 1.34	1.79 ± 2.33	2.54 ± 3.24
German	1	1.47 ± 2.69	1.07 ± 1.77	1.30 ± 1.87	2.88 ± 3.53
Spanish	1	1.24 ± 2.10	1.35 ± 1.74	1.48 ± 1.99	2.44 ± 3.52
Basque	1	1.27 ± 2.08	1.32 ± 1.60	1.43 ± 1.68	2.00 ± 3.00
Dutch	1	1.60 ± 2.32	1.63 ± 2.01	1.53 ± 2.02	1.88 ± 2.99
Br. Port.	1	1.27 ± 1.64	1.37 ± 1.52	1.34 ± 1.69	2.21 ± 3.14
English	2	1.26 ± 1.94	1.27 ± 1.88	1.86 ± 2.57	2.47 ± 3.19
German	2	1.88 ± 3.18	1.02 ± 2.05	1.13 ± 2.09	2.89 ± 3.51
Spanish	2	1.70 ± 4.07	1.31 ± 3.03	1.35 ± 2.90	2.41 ± 3.47
Basque	2	0.99 ± 1.78	0.92 ± 1.52	1.02 ± 1.64	2.05 ± 3.04
Dutch	2	1.24 ± 2.51	1.12 ± 2.11	1.10 ± 2.09	1.87 ± 2.97
Brazilian Portuguese	2	1.04 ± 1.88	1.02 ± 1.68	1.19 ± 2.01	2.17 ± 3.09
After Boundaries					
English	1	6.34 ± 4.12	6.05 ± 3.80	4.67 ± 3.78	2.54 ± 3.24
German	1	7.75 ± 5.31	7.06 ± 3.73	6.39 ± 3.79	2.88 ± 3.53
Spanish	1	12.39 ± 12.75	9.06 ± 10.09	7.88 ± 8.77	2.44 ± 3.52
Basque	1	5.41 ± 3.69	5.89 ± 3.11	5.68 ± 3.24	2.00 ± 3.00
Dutch	1	7.80 ± 4.77	6.52 ± 3.84	5.76 ± 3.89	1.88 ± 2.99
Br. Port.	1	5.84 ± 3.86	5.64 ± 3.54	4.80 ± 3.45	2.21 ± 3.14
English	2	5.30 ± 4.27	5.00 ± 4.00	4.29 ± 3.78	2.47 ± 3.19
German	2	7.26 ± 4.84	6.04 ± 3.97	5.61 ± 3.90	2.89 ± 3.51
Spanish	2	9.79 ± 10.58	7.24 ± 8.24	6.39 ± 7.21	2.41 ± 3.47
Basque	2	4.64 ± 3.37	4.09 ± 3.40	3.84 ± 3.49	2.05 ± 3.04
Dutch	2	6.01 ± 4.94	4.88 ± 4.08	4.54 ± 3.95	1.87 ± 2.97
Brazilian Portuguese	2	5.46 ± 3.95	5.02 ± 3.74	4.39 ± 3.60	2.17 ± 3.09

Table 12: Mean and standard deviation of token surprisal before and after paragraph, sentence, and EDU boundaries. Surprisal is lower before and higher after all boundary types compared to tokens distant from any boundary.

English

Document k	Document A_k	EDU k	EDU A_k	Sentence k	Sentence A_k	Paragraph k	Paragraph A_k
1	0.235_{10}	1	0.370_{10}	4	0.171_{10}	9	0.037_{10}
		2	0.330_{10}	5	0.151_{10}		
		3	0.241_{10}	10	0.144_{10}		
		4	0.180_{10}	3	0.134_{10}		
		5	0.123_{10}	2	0.112_{10}		
		7	0.086_{10}	14	0.092_{10}		
				1	0.066_{10}		
				15	0.060_{10}		
				19	0.059_{10}		
				16	0.056_{10}		
				20	0.046_{10}		

Spanish

Document k	Document A_k	EDU k	EDU A_k	Sentence k	Sentence A_k	Paragraph k	Paragraph A_k
1	0.422_{10}	1	0.364_{10}			126	0.058_{7}
4	0.323_{10}	2	0.313_{10}			535	0.054_{10}
5	0.293_{10}	4	0.264_{10}			150	0.053_{8}
7	0.287_{10}	3	0.261_{10}			381	0.047_{7}
6	0.284_{10}	5	0.235_{10}			173	0.039_{10}
8	0.284_{10}	6	0.205_{10}				
10	0.276_{10}	7	0.205_{10}				
9	0.253_{10}	8	0.159_{10}				
13	0.246_{10}	10	0.140_{10}				
15	0.244_{10}	9	0.139_{10}				
11	0.236_{10}	11	0.107_{10}				
17	0.223_{10}	13	0.107_{10}				
21	0.218_{9}	37	0.036_{10}				
12	0.210_{9}						
14	0.207_{9}						
24	0.206_{10}						
20	0.205_{8}						
16	0.203_{9}						
25	0.197_{10}						
28	0.197_{8}						
19	0.195_{9}						
23	0.194_{10}						
18	0.194_{9}						
46	0.191_{10}						
22	0.190_{8}						

German

Document k	Document A_k	EDU k	EDU A_k	Sentence k	Sentence A_k	Paragraph k	Paragraph A_k
4	0.165_{10}	1	0.599_{10}	10	0.101_{10}	11	0.087_{10}
5	0.140_{10}	2	0.515_{10}	12	0.063_{10}	56	0.077_{10}
6	0.137_{10}	3	0.380_{10}	13	0.059_{10}	345	0.066_{9}
7	0.112_{10}	4	0.256_{10}	33	0.049_{10}	57	0.062_{0}
10	0.091_{10}	5	0.206_{10}	35	0.049_{10}	110	0.057_{8}
8	0.091_{10}	7	0.181_{10}			167	0.045_{10}
11	0.085_{10}	8	0.150_{10}				
22	0.073_{10}	9	0.135_{10}				
285	0.056_{6}	11	0.118_{10}				
111	0.053_{7}	14	0.088_{10}				
55	0.047_{1}	16	0.073_{10}				
86	0.045_{8}	19	0.053_{10}				
63	0.039_{8}						

Dutch

Document k	Document A_k	EDU k	EDU A_k	Sentence k	Sentence A_k	Paragraph k	Paragraph A_k
5	0.153_{10}	1	0.470_{10}	3	0.198_{10}	10	0.095_{10}
6	0.135_{10}	2	0.308_{10}	1	0.141_{10}	66	0.089_{10}
7	0.103_{6}	3	0.247_{10}			8	0.074_{10}
3	0.102_{2}	7	0.165_{10}			43	0.072_{9}
4	0.082_{2}	5	0.159_{10}			32	0.071_{10}
218	0.082_{2}	4	0.153_{10}			175	0.068_{10}
453	0.080_{0}	11	0.133_{10}			23	0.068_{0}
8	0.078_{0}	16	0.121_{10}			4	0.067_{10}
13	0.076_{0}	18	0.084_{10}			7	0.064_{3}
140	0.076_{4}	6	0.074_{10}			247	0.063_{1}
656	0.076_{0}	47	0.060_{10}			5	0.063_{8}
9	0.075_{0}					13	0.061_{9}
227	0.074_{0}					357	0.059_{3}
845	0.067_{0}					47	0.058_{0}
365	0.065_{0}					35	0.057_{9}
296	0.064_{0}					202	0.055_{9}
199	0.062_{0}					38	0.054_{10}
812	0.060_{0}					305	0.053_{0}
15	0.059_{0}					29	0.051_{10}
11	0.059_{0}					290	0.048_{8}
540	0.059_{0}						
245	0.053_{0}						
130	0.053_{0}						
132	0.052_{0}						
235	0.052_{0}						

Basque

Document k	Document A_k	EDU k	EDU A_k	Sentence k	Sentence A_k	Paragraph k	Paragraph A_k
7	0.099_{10}	1	0.260_{10}	189	0.053_{0}	25	0.066_{10}
6	0.098_{10}	2	0.196_{10}	112	0.043_{10}	651	0.054_{9}
8	0.093_{10}	5	0.122_{10}			119	0.041_{7}
10	0.084_{10}	13	0.080_{10}				
26	0.064_{10}	11	0.071_{10}				
12	0.061_{7}	21	0.067_{10}				
27	0.054_{10}						
56	0.052_{8}						
36	0.051_{7}						
443	0.049_{0}						
33	0.047_{2}						
1223	0.047_{0}						
1655	0.045_{2}						
1716	0.045_{0}						
505	0.038_{1}						

Brazilian Portuguese

Document k	Document A_k	EDU k	EDU A_k	Sentence k	Sentence A_k	Paragraph k	Paragraph A_k
24	0.091_{10}	2	0.389_{10}	5	0.129_{10}	27	0.046_{10}
14	0.065_{4}	3	0.312_{10}	4	0.084_{10}	34	0.041_{10}
20	0.059_{3}	4	0.274_{10}	3	0.082_{10}		
22	0.053_{4}	5	0.254_{10}	19	0.058_{10}		
63	0.052_{4}	6	0.197_{10}	2	0.052_{10}		
345	0.051_{3}	8	0.114_{10}	34	0.036_{10}		
791	0.051_{3}	7	0.098_{10}				
1436	0.051_{3}	10	0.090_{10}				
173	0.051_{3}	14	0.088_{10}				
817	0.049_{2}	9	0.060_{10}				
1228	0.048_{3}						
898	0.040_{1}						
730	0.039_{2}						

Table 13: Mean amplitudes (A_k) of the twenty-five most dominant sinusoids persistent across all folds after L_1 regularization. Subscripts denote the number of cross-validation folds (out of ten folds) where the sinusoids are significant.