Surprise! Uniform Information Density Isn't the Whole Story: Predicting Surprisal Contours in Long-form Discourse

Eleftheria Tsipidi **Franz Nowak** **Ryan Cotterell** **Ethan Wilcox**
Mario Giulianelli **Alex Warstadt**
ETH Zürich
{tsipidie, fnowak, rcotterell, ewilcox, mgiulianelli, warstadt}@ethz.ch

Abstract

The Uniform Information Density (UID) hypothesis posits that speakers tend to distribute information evenly across linguistic units to achieve efficient communication. Of course, information rate in texts and discourses is not perfectly uniform. While these fluctuations can be viewed as theoretically uninteresting noise on top of a uniform target, another explanation is that UID is not the only functional pressure regulating information content in a language. Speakers may also seek to maintain interest, adhere to writing conventions, and build compelling arguments. In this paper, we propose one such functional pressure; namely that speakers modulate information rate based on location within a hierarchically-structured model of discourse. We term this the Structured Context Hypothesis and test it by predicting the surprisal contours of naturally occurring discourses extracted from large language models using predictors derived from discourse structure. We find that hierarchical predictors are significant predictors of a discourse's information contour and that deeply nested hierarchical predictors are more predictive than shallow ones. This work takes an initial step beyond UID to propose testable hypotheses for why the information rate fluctuates in predictable ways.

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

1 Introduction

Linguistic communication takes place in a context, a backdrop of both linguistic and non-linguistic content that can determine how utterances' form (Fine et al., 2013) and meaning (Roberts, 2006) are interpreted as well as what words speakers choose to say next (Rohde and Kehler, 2014). We investigate the role of context from an information-theoretic perspective, asking how a **linguistic context**, i.e., what has been said or written previously, shapes the information content of each **linguistic unit**, i.e., a novel word or utterance in that context. One influential hypothesis for the relationship between linguistic units and their context is the **Uniform Information Density (UID)** hypothesis (Fenk and Fenk, 1980; Genzel and Charniak, 2002; Jaeger and Levy, 2006; Meister et al., 2021; Clark et al., 2023), which posits that, subject to the constraints of the grammar, speakers spread out information as evenly as possible across an utterance. If the UID hypothesis is taken to an extreme, i.e., if it is imposed as a hard constraint, then each linguistic unit would add roughly the same amount of information, when the previous context is taken into account.

There is an abundance of empirical support for the UID hypothesis, albeit, in general, for a soft variant of it where there is violable *pressure* towards uniformity. For instance, Clark et al. (2023) gives evidence across a number of languages that word order is optimized for UID. Empirically, however, within a discourse, the information content of individual linguistic units is never observed to be strictly static but rather to fluctuate within a band. We dub this fluctuation in the information content of a discourse its **information contour**; see Fig. 1 for an example. More theoretically, a pressure towards uniformity must naturally be attenuated by other competing functional pressures on linguistic communication. Of course, the grammar constrains word choice, which may make uniformity difficult to achieve (Jaeger and Levy, 2006). Moreover, when an author chooses the next word of a story or a poem, UID might give way to discursive pressures such as a desire for a clean narrative arc or a well-executed rhetorical structure. Indeed, some literary devices, such as rhyme and meter, may even ascribe higher aesthetic value to a non-uniform information rate.

In this article, we propose an elaboration of the UID hypothesis. In addition to a local pressure for uniformity on information modulated by the grammar, we posit that the information contour of a *discourse* itself is a meaningful signal that reflects a richer structured notion of context. The idea that there is a relationship between local information content and hierarchical syntax goes back to Hale (2001) and has been expanded more recently (Jaffe et al., 2020; Oh et al., 2022). However, decades of

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18820–18836
November 12-16, 2024 ©2024 Association for Computational Linguistics

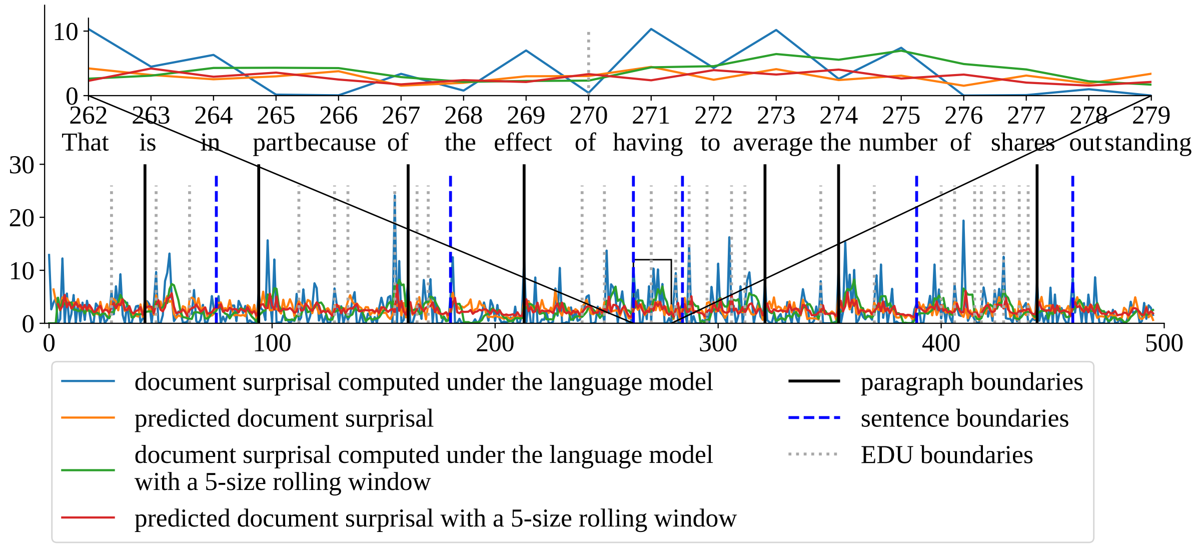


Figure 1: Information contour of the wsj_1111 document from the English RST Discourse Treebank.

previous research have also established that much like sentences are comprised of syntactical constituents, discourses are organized into nested units as well (Mann and Thompson, 1988; Asher and Lascarides, 2003; Prasad et al., 2008). Thus, we hypothesize that, in addition to UID, there is a functional pressure on information contours that respects the hierarchical structure of discourse. We term this the **Structured Context Hypothesis**. In the context of this hypothesis, we put forth the following research questions:

(i) Do structured representations of discourse help explain information contours better than non-structured ones?

(ii) And, if so, what type of structure is best at predicting information rates?

To answer these questions, we use neural language models to estimate the local information content of written English and Spanish texts. We then consider two different representations for the hierarchical discourse structure of a text. The first is the standard prose-writing convention of dividing the document into (shallow) hierarchically nested paragraphs and sentences. The second is based on **Rhetorical Structure Theory** (RST; Mann and Thompson, 1988), which breaks texts into recursively nested spans that are linked by discourse relations. To investigate questions (i) and (ii) above, we apply Bayesian regression analysis to determine whether access to the discourse structure helps us better model information contours. We do find evidence that hierarchical discourse structure helps predict information contours across the board and that RST is more predictive than the shallowly nested paragraph and sentence structures. In sum, this work provides preliminary empirical evidence for the Structured Context Hypothesis and paves the way for a theory explaining how and why information contours may be modulated by discourse structure.

2 Background

There are myriad information-theoretic theories of language. This section builds up to and introduces the Structured Context Hypothesis while contextualizing it in light of previous proposals.

2.1 Language Models and Surprisal

A **language model** is a probability distribution over Σ^* for a given alphabet Σ. Every language model can be decomposed as a product of conditional next-unit probabilities given the units so far, giving an **autoregressive** language model. Specifically, for any string $\boldsymbol{u} = u_1 \cdots u_T \in \Sigma^*$, we may write

$$p(\boldsymbol{u}) \overset{\text{def}}{=} p(\text{EOS} \mid \boldsymbol{u}) \prod_{t=1}^{T} p(u_t \mid \boldsymbol{u}_{<t}). \tag{1}$$

Here, EOS $\notin \Sigma$ is a special end-of-string symbol.

From an autoregressive factorization of a language model, we can define **Shannon surprisal**. Given a language model p, the Shannon surprisal (Shannon, 1948) of a unit in context is its negative log probability in context, i.e.,

$$s(u_t) \overset{\text{def}}{=} -\log p(u_t \mid \boldsymbol{u}_{<t}), \tag{2}$$

Shannon surprisal (or surprisal for short) is one way to operationalize the notion of a unit's information content under the language model p, though other operationalizations are possible (Giulianelli et al., 2023, 2024a,b). Shannon surprisal as defined above has been hypothesized to correlate with the difficulty of a human reader or listener to process an utterance, a notion known as **surprisal theory** (Levy, 2008) which frames the theory of Hale (2001) in an information-theoretic context. Specifically, surprisal theory states that the surprisal of a unit quantifies the cost of incrementally updating expectations as a result of observing the unit (Levy, 2008). The crucial insight of surprisal theory is that it proposes that, insofar as the language model used to measure the probability of units is a good approximation of the human language model, two distinct properties—information content and processing difficulty—can be quantified with a single metric.

2.2 Uniform Information Density

Linguistic communication can be idealized as the transmission of a linguistic signal through a noisy channel with limited capacity (Shannon, 1948, Part 2). Under this view, a speaker is encouraged to choose a string of linguistic units that contains the most information while not surpassing the channel's capacity. This functional pressure is one motivation of the UID hypothesis (Fenk and Fenk, 1980; Jaeger and Levy, 2006). The reason for this even distribution is as follows: On the one hand, if a speaker's linguistic signal contains more information on average than the channel capacity, the communication would be prone to transmission errors. On the other, if the information content, on average, were lower than the channel capacity, then there could be an alternative, more efficient way of formulating the linguistic signal. The optimal strategy is thus to send information across the channel that is as close to the channel capacity as possible without being too difficult for the comprehender to process.

The UID hypothesis suggests that production choices aim to optimize both the limitations of channel capacity and the need to efficiently convey information. This leads to surprisal being distributed as evenly as possible across a speaker's utterance. By preventing significant fluctuations in surprisal, speakers avoid surpassing or falling below channel capacity, ensuring that processing difficulty remains relatively stable for the listener. The UID hypothesis is supported by empirical studies of language production at the level of syllables (Bell et al., 2003; Aylett and Turk, 2004, 2006), lexical items (Meister et al., 2021; Clark et al., 2023), syntactic structures (Frank and Jaeger, 2008; Jaeger, 2010), and discourse connectives (Torabi Asr and Demberg, 2012, 2015).

2.3 Contextualizing the UID Hypothesis

Most instantiations of the UID hypothesis use the surprisal of a linguistic unit in context as an operationalization of that unit's information content. Despite its empirical success at explaining various linguistic phenomena, the UID hypothesis is limited in several ways, which we detail below.

Empirical shortcomings of UID. Empirical estimates of character surprisal within words (Elman, 1990), estimates of word surprisal within sentences (Levy, 2013; Futrell et al., 2020) and estimates of sentence surprisal within discourse structures (Genzel and Charniak, 2003) demonstrated that the rate of surprisal fluctuates in ways that correspond to linguistic structure. For instance, in the case of character surprisal within words, peaks often correspond to morpheme boundaries (Harris, 1955; Elman, 1990; Pimentel et al., 2021) and the word surprisal within utterances may correspond to constituent boundaries (Jaeger and Levy, 2006). However, less work has studied peaks and troughs in information content throughout an entire *discourse*. We posit that the discourse-level fluctuations are likewise not random and may be due to cognitive and linguistic factors. If information contours fluctuate in a predictable manner, e.g., if they exhibit periodic structure, then this would be evidence against a strong version of the UID hypothesis.

The Constancy Rate Principle. Genzel and Charniak (2002) is one notable example of a study that *does* investigate information contours at the discourse level. The authors propose the **constancy rate principle**, which stipulates that the *expected* surprisal, i.e., the entropy of the next unit distribution given all previously uttered units, stays roughly constant throughout a discourse. Specifically, they posit that while the expected surprisal of the next unit given only its current sentence, i.e., taken out of context, increases throughout the discourse, the information contained in the global context grows, too, so that the expected surprisal given the full context stays the same. As their tools at the time were limited to n-gram language models and probabilistic constituency parsers, Genzel and Charniak (2002) could only empirically verify the former claim, i.e., that the surprisal given the local context increases. More recent studies, however, have exploited Transformer-based models to measure surprisal in the global context. These studies do find weak evidence of the constancy rate principle, especially when considering languages other than English (Verma et al., 2023) or other forms of communication, such as conversation (Giulianelli and Fernández, 2021). However, even in cases where some constancy is observed, it is always subject to fluctuations within a band that are beyond the explanatory power of the constancy rate principle.

Other related work. Besides uniformity pressures, language production and comprehension are also known to be modulated by discourse structure. Previous work has investigated how fluctuations of surprisal rates relate to paragraph boundaries (Genzel and Charniak, 2003), topic shifts in text (Qian and Jaeger, 2011) and open-domain dialog (Xu and Reitter, 2016; Maës et al., 2022), task-determined contextual units in goal-oriented dialog (Giulianelli et al., 2021), as well as extra-linguistic contextual cues (Doyle and Frank, 2015) in multi-party conversations.

Theoretical shortcomings of UID. A second, more theoretical limitation of UID is that it does not inherently take into account language-internal pressures other than grammaticality. Certain linguistic units, regardless of their information profile, might be dispreferred within a linguistic context due to discourse constraints, argumentative considerations, or aesthetic preferences. One good example of how language-internal pressures play out at multiple levels of linguistic structure are **contour principles**, constraints against identical segments (or segments with identical features) occurring consecutively which result in non-uniformity of linguistic units. Although originally developed to explain non-uniformity of phonological features through the Obligatory Contour Principle (Leben, 1973), contour principles have been posited to govern the information content of linguistic units at various degrees of granularity, e.g., words within higher levels organization including paragraphs (Genzel and Charniak, 2003) and discourse topics (Xu and Reitter, 2016). In addition, contour-like principles are often recruited to explain, and teach, good writing (Kharkwal and Muresan, 2014; Snow et al., 2015; Archer and Jockers, 2016). At first blush, it is not clear how to reconcile UID with pressures deriving from such contour principles.

Underspecificity. The above discussion points to a broader limitation of UID, namely, that it is underspecified. While it postulates that information be spread out as evenly as possible throughout linguistic units, it does not provide a specific formulation of uniformity: Which surprisal rates count as uniform? And, should information be uniform independently of other language-internal or structural pressures discussed above or only after controlling for these? Finally, within which notion of linguistic context should surprisal remain uniform? Different formulations of uniformity have been explored for rates of word (Collins, 2014; Meister et al., 2021) and utterance (Giulianelli and Fernández, 2021) surprisal in discourse, with findings hinting at a global uniformity of surprisal—i.e., surprisal tends to stay close to a discourse-level average throughout—especially when larger communicative units are taken into account.

3 The Structured Context Hypothesis

To harmonize UID with the constraints imposed by contour principles, we propose the Structured Context Hypothesis. In most previous work, context is modeled as an essentially sequential object—a succession of paragraphs, topic episodes, dialogue transactions, or dialogue rounds. In contrast, we rely on a different view, considering context as hierarchical representations made up of sentences within paragraphs or deeply nested discourse trees. We hypothesize that the fluctuations observed in surprisal contours of discourse beyond a baseline uniformity can be at least partially accounted for by considering structured representations of the discourse in question. This means that taking into account hierarchical dependencies beyond the sentence level in our theories should increase their ability to predict the information rate of discourse. We express this view through the following hypothesis:

Hypothesis 1 *The Structured Context Hypothesis: The information contour of a discourse is (partially) determined by the hierarchical structure of its constituent discourse units.*

The objective of our experiments is to empirically test this hypothesis on English and Spanish texts.

In the remainder of this section, we outline two manners to represent documents' hierarchical dis-

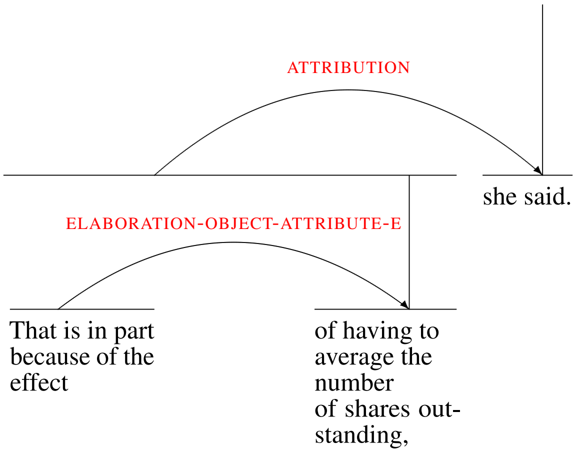

ATTRIBUTION
ELABORATION-OBJECT-ATTRIBUTE-E
That is in part because of the effect
of having to average the number of shares outstanding,
she said.

Figure 2: Discourse sub-tree for a sentence in wsj_1111 from the English RST Discourse Treebank.

course structure: the conventional prose structure of paragraphs and sentences, and the fine-grained Rhetorical Structure Theory.

3.1 Conventional Prose Structure

The first hierarchical discourse structure we consider is the conventional subdivision of documents into paragraphs in which utterances correspond to sentences and the basic linguistic units are individual words. In what follows, we refer to this way of hierarchically structuring a text simply as **prose structure**. Documents structured in this way can be seen as shallow trees with a depth of at most three.

3.2 Rhetorical Structure Theory

Rhetorical Structure Theory (RST; Mann and Thompson, 1988) is a well-known discourse analysis framework that posits a high degree of hierarchical structure in a discourse along with categorizing the relationships between parts of the discourse. The RST representation is a tree structure (Fig. 2); the leaves of the tree correspond to text fragments, usually clauses, which are referred to as **elementary discourse units** (EDUs). Internal nodes of the tree correspond to contiguous spans of non-elementary discourse units called **complex discourse units** (CDUs). While we are primarily concerned with the tree's hierarchical structure, the tree also contains additional information about the text, which may be valuable. A tree node is labeled as a **nucleus** if it provides essential information, and as a **satellite** if its meaning has a more auxiliary function. Tree nodes are also labeled by their **rhetorical relations** to one or more contiguous discourse units, with labels such as CONSEQUENCE or ELABORATION.

4 Methods

The goal of our statistical analyses is to study whether the information contour of a text can be predicted from discourse representations. Our models predict measures of information content (dependent variables) based on a number of predictors (independent variables), some of which we designate as **baseline predictors** while the others designate as **independent predictors** for convenience. For a summary of all variables, see Tab. 1 in App. B.

4.1 Dependent Variables

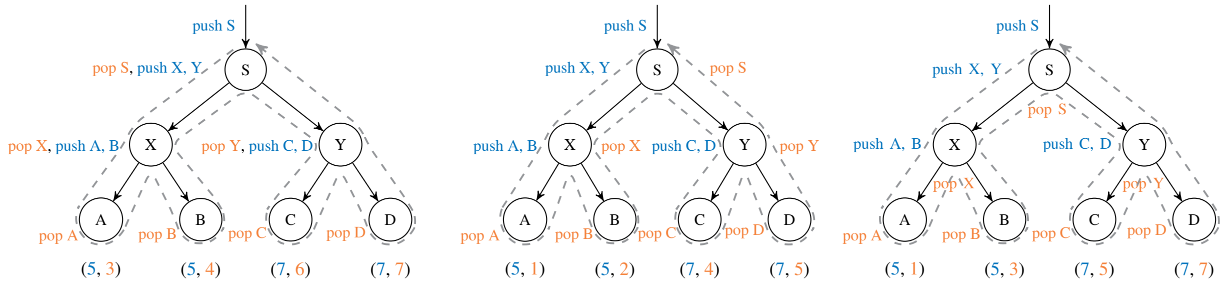
We express information contours in terms of four types of dependent variables (see App. D for formal definitions). The first dependent variable is the global per-unit surprisal, i.e., the surprisal of a unit conditioned on its entire preceding context, starting from the beginning of the document. We also refer to this as **document surprisal**. In addition to global surprisal, our second dependent variable is its **rolling average** of a window of 3, 5, and 7 units (i.e., tokens). The third dependent variable type is the difference between a unit's global surprisal and its surprisal in a local context. This is equivalent to the **pointwise mutual information** (PMI) between the unit and its preceding context conditioned on the local context. Following previous work (Genzel and Charniak, 2002; Giulianelli and Fernández, 2021, *inter alia*), we consider a local context to be the context beginning with the current sentence or current EDU, and the global context to be all material in a document that precedes the current unit. We also compute this PMI conditioned on no local context, which is simply the difference of the global surprisal and the unigram surprisal of the unit. We include these measures to assess how much the particular details of the larger discourse context impact the information content of the current unit.

4.2 Baseline Predictors

Our baseline predictors include the length of the current unit, measured in characters, and the surprisal of the previous unit in all experiments. These are quantities that we expect to be predictive of the current unit's surprisal, but that do not bear directly on the structured context hypothesis.

4.3 Independent Predictors

Beyond the baseline predictors, our analyses are based on two main sets of independent predictors: those derived from prose structure trees and those



(a) top-down parsing (b) bottom-up parsing (c) left-corner parsing

Figure 3: Illustration of the PUSHes and POPs from different parsing strategies, with top-down parsing (left) popping nodes in preorder, and bottom-up parsing (right) popping nodes in postorder. Note that the pushing of rules during the depth-first-search is equal in both cases.

derived from RST trees. Independent predictors for both types of discourse trees are of four main types.

Relative position. These predictors encode the distance of our most granular-level unit[1] from the beginning of a higher-level structural unit, normalized by the higher-level unit's length; for example, the distance between a token and the start of a paragraph in which it is located, normalized by the number of tokens in the paragraph.

Nearest boundary. These predictors encode the distance of a granular-level unit from the closest boundary—left or right—of a higher-level structural unit, normalized by the higher-level unit's length. Nearest boundary predictors allow us to test for non-monotonic relationships between surprisal and a unit's position in its parent and ancestor units.

Hierarchical position. These predictors encode the relative position of a unit within its parent in the hierarchical structure, such as the relative position of the unit in a sentence, or of a paragraph (that contains the unit) in the document. These predictors allow us to assess the level of a hierarchical context structure that most affects surprisal values.

Transition predictors. These predictors encode parsing information on RST and prose structure trees. We define integer-valued predictors from the discourse trees yielded by the RST and prose structure annotations of our data. We do this by traversing binarized versions of the various trees using common parsing strategies (top-down, bottom-up, and left-corner) for context-free grammars and recording corresponding PUSH and POP actions between the leaves of the trees; we illustrate this in Fig. 3. For more details, see App. C.

4.4 Predictive Modeling Framework

To assess the predictive power of different discourse representations, we compare the goodness of fit of a Bayesian linear regressor (Clyde et al., 2022) that includes independent and baseline predictors (the **target model**) to one that uses only the baseline predictors (the **baseline model**) to predict information contours. Dependent variables and their predictors are described above in §§ 4.1 to 4.3; App. B provides a summary. For each set of predictors, we perform 5-fold cross-validation, estimating a posterior on four folds of the data at a time. We fit the Bayesian linear regressor using the using the Pyro framework (Bingham et al., 2019) with its implementation of stochastic variational inference (Hoffman et al., 2013). We use an AutoNormal autoguide, the Adam optimizer (Kingma and Ba, 2015), a learning rate of 0.03, and the evidence lower bound (Kingma and Welling, 2014) as our objective. Then, we compute the expected mean-squared error (MSE) under the Bayesian posterior on the held-out fold. We aggregate the expected MSEs across the held-out folds to approximate the expected MSE across the entire dataset. The predictive power of a set of predictors is calculated as the difference in expected MSE between the baseline model and the target model. We refer to this metric as ΔMSE. To assess the statistical significance of a predictor group's ΔMSE, we run paired permutation tests with the cross-validation results.

5 Data

We conduct experiments on the English RST Discourse Treebank (Carlson et al., 2001; Carlson and

[1] At the most granular level, our units are tokens, obtained by running the tokenizers of the language models we use to estimate our ground truth surprisal values.

Marcu, 2001) and the Spanish RST Discourse Treebank (da Cunha et al., 2011). For the English Treebank, we consider only the train set, containing 347 documents from the Wall Street Journal. The Spanish Treebank contains 267 specialist-authored documents in 9 domains, e.g., astrophysics, mathematics, and law; we discard 11 documents due to missing nodes in the RST trees.

Data preprocessing and RST annotations. We preprocess the data following Braud et al. (2017), e.g., we skip document titles which are not part of the RST trees themselves. We also use their code[2] to perform right-binarizarization of the RST trees, but do not perform label harmonization (Braud et al., 2017, §4.2) because we do not make use of any rhetorical relation labels in our experiments.

Prose structure annotations. To recover prose structure boundaries, i.e., paragraph and sentence boundaries, we match English documents to the corresponding plaintexts provided in the Penn Treebank (Marcus et al., 1999). The Spanish Discourse Treebank directly provides paragraph boundaries, and we recover sentence structure with a text-to-sentence splitter,[3] with manual corrections where necessary. We also perform right-binarization using the NLTK library (Bird et al., 2009) to make the prose structure trees consistent with the RST trees.

Surprisal estimation. On the English RST Discourse Treebank, we compute the next-unit surprisal with the NOUSRESEARCH/YARN-LLAMA-2-7B-64K language model (Peng et al., 2024). We selected NOUSRESEARCH/YARN-LLAMA-2-7B-64K because it is trained with a long context window while still being lightweight enough to run on our compute budget. We compute surprisals on the Spanish RST Discourse Treebanks with the LINCE Mistral 7B Instruct language model.[4]

6 Empirical Findings

We overview our empirical results in this section, structuring our presentation in terms of five research questions relating to the Structured Context Hypothesis In Q1–4, we move from shallower to deeper discourse structure representations, focusing on RST-based predictors. In Q5, we compare RST to conventional prose structure.

Q1: Are information contours predictable from the relative position within a discourse unit? To answer this question, in Fig. 4 we visualize the ΔMSE (§4.4) of models trained on RST relative position information. We find that including these RST predictors into the model leads to lower ΔMSE on the held-out data compaared to the baseline ($p < 0.001$) indicating that structured contexts help to predict the information contours of a text. Relative position is the best-performing RST-based predictor group for English across dependent variables (Fig. 4a; $p < 0.001$ against all other predictor groups) and second best for Spanish (Fig. 4b; $p < 0.001$ against all but hierarchical position).

Q2: Is the effect of relative position within a discourse unit non-monotonic? To account for possible non-monotonicity, we trained models on predictors including relative distance to nearest boundaries within a discourse unit. These predictors can account for increases in information content close to the end of the unit after a decrease in the middle of the unit or, vice versa, for lower rates of information content closer to the unit's boundaries. However, the resulting ΔMSE for both English and Spanish shows less improvement over the baseline compared to the relative position predictors ($p < 0.001$), indicating the effect of a unit's position within a discourse unit is better modeled as monotonic.

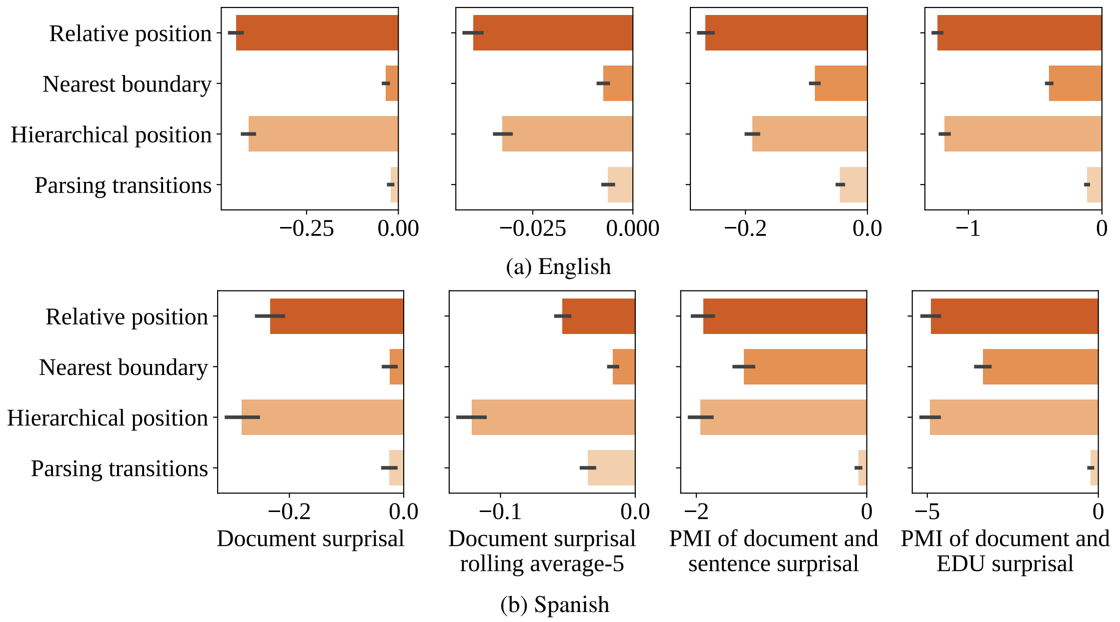
Q3: Does relative position in higher-order structures predict information contours? To assess the explanatory power of hierarchical discourse structures for information contours, we use models that include as predictors the relative position of a unit within its parent in the hierarchical structure. We find hierarchical position is a significant predictor of all dependent variables analyzed, and either the best or the second-best out of all predictor groups tested. In the English data, it is moderately less predictive than relative position ($p < 0.001$; see Fig. 4a). In the Spanish data, it is the strongest predictor of document surprisal and its rolling average ($p < 0.001$) against all other predictors), and on par with the relative position ($p > 0.001$) for the PMI dependent variables.

Q4: Does hierarchical structure encoded by discourse parsing transitions help predict information contours? As an alternative way to represent the hierarchical structure of the text, we consider predictors obtained by deriving the RST

[2] https://bitbucket.org/chloebt/discourse

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

[4] https://huggingface.co/NousResearch/Yarn-Llama-2-7b-64k; https://huggingface.co/clibrain/lince-mistral-7b-it-es



(a) English

(b) Spanish

Figure 4: ΔMSE comparison of models trained on four RST-based predictor groups. Note that the scale for surprisal with a rolling window of 5 is smaller, as rolling average dependent variables exhibit less variance. All these results are statistically significant against the baseline ($p < 0.001$).

tree structure via constituency parsing algorithms (App. C). Although the ΔMSE is negative in all cases, indicating an increase in predictive power over a baseline model, transition predictors are significantly worse predictors of information contours than relative and hierarchical position ($p < 0.001$), for both English and Spanish.

Q5: What representation of discourse structure, RST or Prose Structure, best explains information contours? Fig. 5 presents a comparison of all individual RST predictors analyzed so far and their prose structure analogs in terms of their ΔMSE scores across dependent variables on the English data. We consider two models (referred to as RST all and PS all in Fig. 5) that include all predictors derived from either representation of discourse structure. Results for the Spanish data are shown in Fig. 6 (App. E). Our findings are consistent across the two languages. For document surprisal and surprisal with a rolling average of 5 units, RST predictors are better than PS predictors ($p < 0.001$). We observe similar trends for PMI of document and unit surprisal, and rolling averages of 3 and 7 (see Fig. 7 in App. E). Furthermore, when considering the locally conditioned PMI variables, we find a correspondence between the strongest family of predictors and the local context over which the PMI is conditioned: The predictive power of RST predictors is higher for EDU-conditioned PMI ($p < 0.001$) while PS predictors are better for sentence-conditioned PMI ($p < 0.001$).

Summary. Taken together, our results indicate information contours extracted from language models do exhibit discourse-structural dependencies. These dependencies are determined both by structural units of conventional prose writing and by more hierarchical discourse units. However, explanatory power is higher for the finer-grained and higher-order structures determined by rhetorical relations between discourse units.

7 Future Work

We hypothesized that violations of the communicative pressure to communicate at a uniform rate might be predictable, and that part of their predictability is linked to how production choices depend on discourse structure (the Structured Context Hypothesis). While we could not determine violations of the UID hypothesis precisely due to its inherent underspecificity, our predictive modeling framework captured deviations from a constant information rate by design, with the intercept representing the baseline rate and predictors capturing deviations. As Fig. 1 shows, the structured context helps predict oscillations around the base rate, though we only account for a small portion of these deviations. There are, however, additional intuitive explanations for violations of uniformity which our predictors do not capture, or only do so partially.

Maintaining interest. High surprisal content may help to maintain a listener's attention. In the

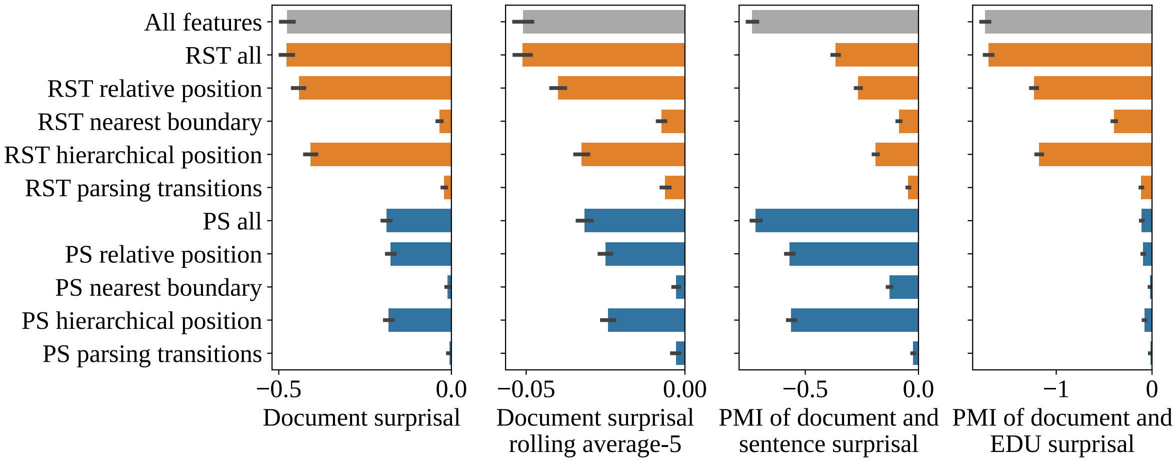


Figure 5: ΔMSE across dependent variables of all RST and Prose Structure (PS) predictors on the English data.

domain of music synthesis, it has already been proposed that modulating surprisal can affect listener engagement (Kothinti et al., 2020; Bjare et al., 2024). Extending this idea to language, Venkatraman et al. (2023) found that when controlling for total surprisal, non-uniformity of information density correlates with text quality.

Improving comprehension. Overly information-dense content may hinder comprehension. In such cases, low-surprisal content such as repetitions, reiterations, and summaries at strategic points in the discourse structure can intuitively help to reinforce new information and reduce confusion. Indeed, redundancy is an important feature in error-correcting codes (Hamming, 1950), and repetitions are important for comprehension in noisy-channel situations such as conversations between second-language learners (Cervantes and Gainer, 1992).

Production constraints. Peaks and troughs need not be only out of concern for a listener. Speakers have limited effort to expend on formulating utterances, and so they may use repetition to maintain the flow of conversation (Giulianelli et al., 2022) or hold the floor while formulating a new, more informative utterance (Bergey and DeDeo, 2024).

Aesthetic conventions. General aesthetic principles or specific stylistic conventions may intentionally manipulate surprisal. Repetition is common to many rhetorical devices (Harris, 2017), and poetic devices such as rhyme and meter increase predictability. At the level of an entire narrative, emotional arcs have been argued to be conventionalized or to cluster into one of several archetypes (Reagan et al., 2016; Brown and Tu, 2020), though this idea has not yet, as far as we know, been extended to information content.

Each of these explanations draws on intuitions that we believe to be widespread and compelling, yet are out-of-scope for the UID hypothesis. More importantly, they make empirical predictions that can be tested by investigating surprisal contours in texts and discourses from different genres that have been annotated for features such as interest and comprehensibility. Moving forward, it will become necessary to look at the surprisal contour as just one of many possible types of time-series data that can be associated with a discourse, and which may be related to each other in meaningful ways.

8 Conclusion

We conclude by briefly highlighting the theoretical and empirical contributions of this paper. Theoretically, we have enumerated the limitations of the UID hypothesis and have provided an initial hypothesis, the Structured Context Hypothesis, to predict how information fluctuates during a discourse, namely discourse trees based on prose conventions and RST. Empirically, we have found support for this hypothesis by evaluating two structured representations of discourse in English and Spanish. We view this work as one step in developing theories that can explain the vast variation in discourses, texts, and writing genres observed across human cultures..

Limitations

One major limitation of the present work is that it is conducted only in English and Spanish. In order to expand to a greater number of languages we have already identified RST-annotated corpora in Basque (Iruskieta et al., 2013, 2015), Brazilian Portuguese (Maziero et al., 2015; Cardoso et al., 2011; Collovini et al., 2007; Pardo and Seno, 2005; Pardo and Nunes, 2003, 2004), Dutch (Van Der Vliet et al., 2011; Redeker et al., 2012) and German (Stede, 2004; Stede and Neumann, 2014; Bourgonje and Stede, 2020). These corpora should additionally be tested as a possible next step. One other limitation of this work is that we have only used linear models. Although we do investigate whether the relationship between discourse boundaries and surprisal is monotonic, it may be the case that the relationship is non-linear. Finally, while our theoretical discussion of (non-)uniformity applies to linguistic units of any size, in practice we only measure and predict the surprisal of tokens under the language model (roughly words). Our conclusions might change if the surprisals of characters, phonemes, sentences, intonation phrases, or any number of other units are considered.

Ethics Statement

We foresee no ethical problems with our work.

References

Jodie Archer and Matthew L. Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel*. St. Martin's Press, Inc., USA.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

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.

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.

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.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Mathias Rose Bjare, Stefan Lattner, and Gerhard Widmer. 2024. Controlling surprisal in music generation via information content curve matching. ArXiv:2408.06022 [cs, eess].

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Steven Brown and Carmen Tu. 2020. The shapes of stories: A "resonator" model of plot structure. *Frontiers of Narrative Studies*, 6(2):259–288.

Paula C. F. Cardoso, Erick G. Maziero, Mara Luca Castro Jorge, Eloize M. R. Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas 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, USC 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.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for

Uniform Information Density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Merlise Clyde, Mine Çetinkaya Rundel, Colin Rundel, David Banks, Christine Chai, and Lizzy Huang. 2022. *An Introduction to Bayesian Thinking*, 1 edition. Academic Press.

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

Sandra Collovini, Thiago I. Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informaçoes discursivas visandoa sumarizaçao automática. *Proceedings of TIL*.

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.

Hal Daume III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 449–456, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gabriel Doyle and Michael Frank. 2015. Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1587–1596, Denver, Colorado. Association for Computational Linguistics.

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

R. M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communication*. MIT Press Classics. MIT Press.

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

Alex B. Fine, T. Florian Jaeger, Thomas A. Farmer, and Ting Qian. 2013. Rapid expectation adaptation during syntactic comprehension. *PLOS One*, 8(10):e77661.

Austin F. Frank and T. Florian 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*.

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.

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.

Dale Gerdemann. 1994. Parsing as tree traversal. In *The 15th International Conference on Computational Linguistics*, Kyoto, Japan.

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, Andreas Opedal, and Ryan Cotterell. 2024a. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 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, Arabella Sinclair, and Raquel Fernández. 2022. Construction repetition reduces information rate in dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 665–682, Online only. Association for Computational Linguistics.

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

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.

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

R. W. Hamming. 1950. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Robert A. Harris. 2017. *Writing With Clarity and Style: A Guide to Rhetorical Devices for Contemporary Writers*. Routledge.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*.

John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2001. *Introduction to Automata Theory, Languages, and Computation*, 3 edition. Pearson.

Mikel Iruskieta, Marıa J Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *Proceedings of the 4th Workshop RST and Discourse Studies*, pages 40–49.

Mikel Iruskieta, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2015. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 11(2):303–334.

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

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

Evan Jaffe, Cory Shain, and William Schuler. 2020. Coreference information guides human expectations during natural reading. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4587–4599, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915, Jeju Island, Korea. Association for Computational Linguistics.

Gaurav Kharkwal and Smaranda Muresan. 2014. Surprisal as a predictor of essay quality. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–60, Baltimore, Maryland. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, San Diego, CA, USA.

Diederik P. Kingma and Max Welling. 2014. Autoencoding variational Bayes. In *2nd International Conference on Learning Representations*.

Sandeep Kothinti, Benjamin Skerritt-Davis, Aditya Nair, and Mounya Elhilali. 2020. Synthesizing engaging music using dynamic models of statistical surprisal. In *International Conference on Acoustics, Speech and Signal Processing*, pages 761–765, Barcelona, Spain. IEEE.

William Ronald Leben. 1973. *Suprasegmental Phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

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

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In *Sentence Processing*, pages 78–114. Psychology Press.

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.

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.

Erick G. Maziero, Graeme Hirst, and Thiago A.S. Pardo. 2015. Adaptation of discourse parsing models for the Portuguese language. In *2015 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 140–145.

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.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Opedal, Eleftheria Tsipidi, Tiago Pimentel, Ryan Cotterell, and Tim Vieira. 2023. An exploration of left-corner transformations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13393–13427, Singapore. Association for Computational Linguistics.

Thiago Alexandre Salgueiro Pardo and Maria das Graças Volpe Nunes. 2003. A construção de um corpus de textos científicos em português do brasil e sua marcação retórica. Technical report, Universidade de São Paulo.

Thiago Alexandre Salgueiro Pardo and Maria das Graças Volpe Nunes. 2004. Relações retóricas e seus marcadores superficiais:: análise de um corpus de textos científicos em português do brasil. *Relatório Técnico NILC*.

Thiago Alexandre Salgueiro Pardo and Eloize Rossi Marques Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. *Proceedings of Encontro de Corpora*.

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*.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(31).

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*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association.

Craige Roberts. 2006. Context in dynamic interpretation. *The Handbook of Pragmatics*, pages 197–220.

Hannah Rohde and Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

D. J. Rosenkrantz and P. M. Lewis. 1970. Deterministic left corner parsing. In *11th Annual Symposium on Switching and Automata Theory (swat 1970)*, pages 139–152.

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

Erica L. Snow, Laura K. Allen, Matthew E. Jacovina, Cecile A. Perret, and Danielle S. McNamara. 2015. You've got style: Detecting writing flexibility across time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, page 194–202, New York, NY, USA. Association for Computing Machinery.

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*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2669–2684, Mumbai, India.

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.

Nynke Van Der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated Dutch text corpus. *S. Dipper and H. Zinsmeister (Eds.), Beyond Semantics, Bochumer Linguistische Arbeitsbericht*, 3:157–171.

Saranya Venkatraman, He He, and David Reitter. 2023. How do decoding algorithms distribute information in dialogue responses? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 953–962, Dubrovnik, Croatia. Association for Computational Linguistics.

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. 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.

A Reproducibility

We extracted the true surprisal values using an RTX 4090 GPU with VRAM 24GB and additional RAM of 64GB for 6 hours. Our predictive modeling experiments required a total of 70 hours on an RTX 3080 GPU with a 10 GB VRAM and 32GB RAM. For details on autoguides, we refer to Pyro documentation[5].

B Overview of Variables

Tab. 1 provides an overview of the independent and dependent variables used in our experiments.

C Transition Predictors

RST transition predictors. To incorporate the hierarchical structure information of the RST annotations, we extract several integer variables from the RST trees corresponding to their tree structure. In line with related work (Daume III and Marcu, 2002; Joty et al., 2012, *inter alia*), we assume that RST annotations correspond to parse trees of a context-free grammar (CFG). Most CFG constituency parsers are of one of three variants that determine in which order the nodes of the parse tree are constructed: Top-down (TD), bottom-up (BU), and left-corner (LC) (Rosenkrantz and Lewis, 1970; Opedal et al., 2023). As Gerdemann (1994) notes, each of the three parsing variants follows a specific depth-first-search tree traversal strategy. Specifically, TD, BU, and LC correspond to pre-order, post-order, and in-order traversal for a given parse tree, respectively.

We first preprocess the RST trees by right-binarizing them. Then, for each of the parsing strategies, we assign integer values to the leaves of the RST trees using the following steps: 1) Traverse the RST tree in depth-first order; 2) When adding a CFG rule to the set of rules to be evaluated later, we increment a PUSH counter; 3) When evaluating a rule at a node, we increment a POP counter; 4) When reaching a leaf node, we assign that node the value of the PUSH and POP counters. In other words, each RST terminal node gets assigned a tuple containing the number of PUSHes and POPs that happened before evaluating the leaf node's rule under TD, BU, and LC parsing. Note that this is related to how pushdown automata parse context-free grammars (Hopcroft et al., 2001, Ch. 6). Since in bottom-up parsing, not all POPs happen before the last EDU of a document, we report the actions twice for each EDU, recording both the **previous** actions up to the given EDU and the **next** actions which are the same values but shifted by one position to the left. See Fig. 3 for an illustration.

Prose structure Transition predictors. We also extract transition predictors from the prose structure of our data using the same method. The main difference is that the structural units of prose structure are sentences and paragraphs rather than RST EDUs. To perform constituency parsing on the (flat) prose structure trees, we first right-binarize them, as we did for the RST trees. The transition predictors can then be extracted using the same rules as described above since the parsing strategies work for arbitrary binary trees.

D Dependent Variables

The goal of our analyses is to test whether the information rate of text can be predicted from discourse trees. We express information rate in terms of four types of dependent variables. We consider a document to be made up of hierarchically arranged units, where each higher-level unit contains the units below it in the structure. We use the following notation: u is a unit drawn from an alphabet Σ, and $\boldsymbol{u}$ is a string of units, i.e., an element of Σ^*. Note that we consider the alphabet Σ of units to correspond to a *specific* level of the discourse tree; e.g., characters, words, sentences, etc. When looking at such a string of same-level units in a hierarchical document, each individual unit can be contextualized as a tuple $(u, \boldsymbol{\ell}, \boldsymbol{g})$, where $\boldsymbol{g}$ is the **global context**, i.e., all the units that linearly preceded u in the document, and $\boldsymbol{\ell}$ is the **local context**, i.e., all the units that preceded u in the document with the additional restriction that they share the same parent in the hierarchical structure. Note the global context subsumes the local context. When we use prose structure to compute the dependent variables, the units are tokens and the parent units are sentences, meaning the global context $\boldsymbol{g}$ of a token u contains all the tokens before u in the document, while the local context $\boldsymbol{\ell}$ contains all the tokens from the start of the sentence. When we use RST trees, the local context of a unit is all the preceding units in the given EDU.

Global surprisal. This is the per-unit surprisal conditioned on the entire preceding context, start-

[5] https://docs.pyro.ai/en/dev/infer.autoguide.html

Variable Family	**Variable Type**	**Description**
Document surprisal	Dependent	Surprisal of unit u with global context $\boldsymbol{g}$
Rolling average (n)	Dependent	Rolling average of document surprisal with a window $n \in \{3, 5, 7\}$
PMI	Dependent	Pointwise mutual information of: (i) u with global context $\boldsymbol{g}$ and u without context (unigram) (ii) u with global context $\boldsymbol{g}$ and u with local context $\boldsymbol{\ell}$ (i.e., the containing sentence in prose structure, or the containing EDU in RST)
Relative position	Independent	Relative position of unit u within higher-level unit
Boundary distance	Independent	Relative distance of u from the nearest boundary (start or end) of higher level unit
Hierarchical position	Independent	Relative position of discourse unit v (where v is or contains u) within higher-level unit w normalized by the total number of discourse units nested directly under w
Parsing transitions	Independent	{previous, next} × {PUSHes, POPs} × {bottom-up, top-down, left corner} number of transitions of either type directly preceding or following u according to different parsing strategies
Unit length	Baseline	length of u in terms of lower-level units
Previous unit surprisal	Baseline	Surprisal of unit preceding u

Table 1: Summary of all the variables (dependent variables, independent predictors, and baseline predictors) used in our regression analysis. All variables are associated with a single unit u.

ing from the beginning of the document:

$$s_g(u) \overset{\text{def}}{=} -\log p(u \mid \boldsymbol{g}), \tag{3}$$

where p is the probability produced by a language model. We will also refer to global surprisal as **document surprisal** in experiments. Eq. (3) is identical to Eq. (2).

Rolling average of global surprisal. We compute the rolling average of document information contours over windows of size $n \in \{3, 5, 7\}$. Thus, the highly local peaks and throughs of the original information contour are smoothened out in the resulting contours.

PMI: Unit and global context. We also measure the difference between a unit's unigram probability and its global surprisal under our language model. This difference is the pointwise mutual information (PMI; Fano, 1961) between the unit and its global context:

$$\text{PMI}(u; \boldsymbol{g}) = \log p_{\text{uni}}(u) - \log p(u \mid \boldsymbol{g}). \tag{4}$$

where p_{uni} is u's unigram probability (Opedal et al., 2024, Eq. 10b). PMI is a common measure in NLP (Church and Hanks, 1990) and measures the degree of association, or mutual dependence, between the two variables.

PMI: Unit and global context conditioned on local context. We also measure the PMI between a unit and its global context, when the local context is taken into account:

$$\begin{aligned}&\text{PMI}(u; \boldsymbol{g} \mid \boldsymbol{\ell})\\ &\quad = \log p(u \mid \boldsymbol{\ell}) - \log p(u \mid \boldsymbol{g}, \boldsymbol{\ell}).\end{aligned} \tag{5}$$

This is a measure of how much larger discourse context impacts the information of the current unit, even when local information is taken into account. Specifically, we take units to be tokens and compute two versions of this value, one where the local context is the containing sentence and one where it is the containing EDU.

E Further Experimental Results

In Fig. 6, we show the Spanish results corresponding to the English ones in Fig. 5. We also provide the results on the remaining dependent variables in Fig. 7a for English and Fig. 7b for Spanish.

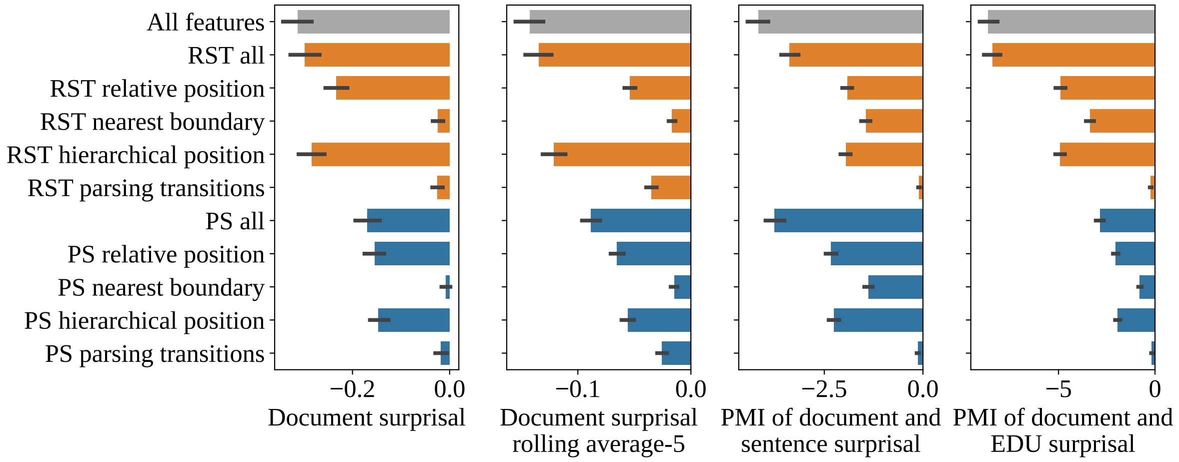
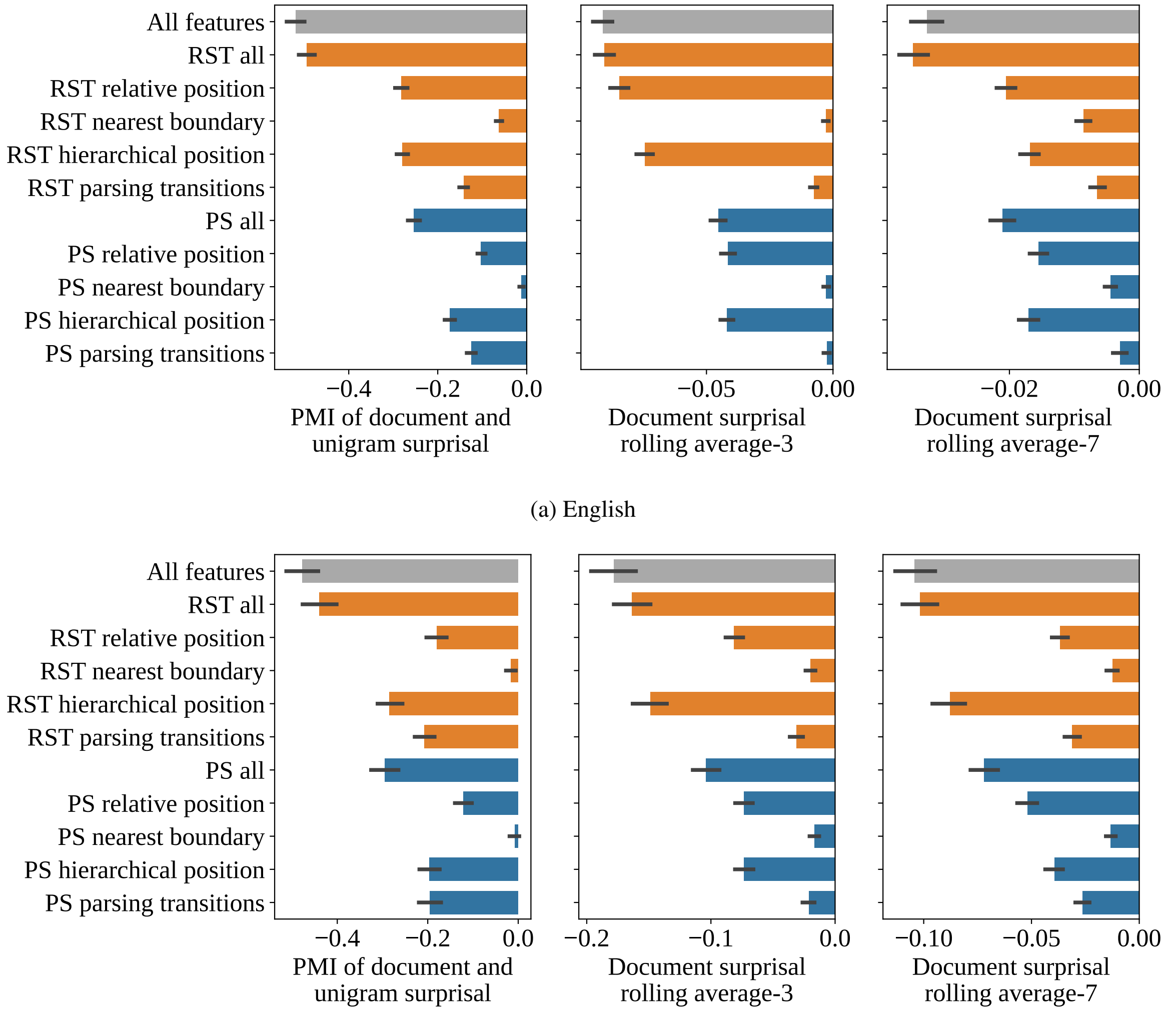


Figure 6: ΔMSE of RST and Prose Structure (PS) across the same dependent variables as Fig. 5 on Spanish data.



(a) English

(b) Spanish

Figure 7: ΔMSE of RST and Prose Structure (PS) across the remaining dependent variables.