Is Information Density Uniform when Utterances are Grounded on Perception and Discourse?

Matteo Gay[U,*] **Coleman Haley**[P] **Mario Giulianelli**[D] **Edoardo M. Ponti**[P]
[U]KU Leuven [P] University of Edinburgh [D] University College London
[U]matteo.gay@student.kuleuven.be [P]coleman.haley@ed.ac.uk
[D]m.giulianelli@ucl.ac.uk [P]eponti@ed.ac.uk

Abstract

The Uniform Information Density (UID) hypothesis posits that speakers are subject to a communicative pressure to distribute information evenly within utterances, minimising surprisal variance. While this hypothesis has been tested empirically, prior studies are limited exclusively to text-only inputs, abstracting away from the perceptual context in which utterances are produced. In this work, we present the first computational study of UID in visually grounded settings. We estimate surprisal using multilingual vision-and-language models over image–caption data in 30 languages and visual storytelling data in 13 languages, together spanning 11 families. We find that grounding on perception consistently smooths the distribution of information, increasing both global and local uniformity across typologically diverse languages compared to text-only settings. In visual narratives, grounding in both image and discourse contexts has additional effects, with the strongest surprisal reductions occurring at the onset of discourse units. Overall, this study takes a first step towards modelling the temporal dynamics of information flow in ecologically plausible, multimodal language use, and finds that grounded language exhibits greater information uniformity, supporting a context-sensitive formulation of UID.

1 Introduction

The Uniform Information Density (UID) hypothesis posits that when language users "*have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus)*" (Jaeger, 2010). The preference for avoiding sharp fluctuations in information transmission results in a smoothing effect over the "information contour" of the signal (Tsipidi et al., 2024), a phenomenon thought to be linked to a reduced cognitive load and more efficient processing (Meister et al., 2021; Clark et al., 2023). In practice, this effect is realised through linguistic choices[1] that minimise the variance of surprisal across linguistic units (e.g., words) within a specified contextual window (e.g., a sentence or paragraph).

Information density is commonly measured in terms of *surprisal*, i.e., the bits of information conveyed by a word given its context (Shannon, 1948; Futrell and Hahn, 2022). Traditional computational and psycholinguistic research on UID has predominantly relied on text-only language models to estimate surprisal distribution, focusing on a system-internal notion of UID that abstracts linguistic sequences from extralinguistic context (Genzel and Charniak, 2002; Aylett and Turk, 2004, 2006; Jaeger and Levy, 2006). Moreover, while some studies have examined the influence of broader linguistic contexts (e.g., Doyle and Frank, 2015; Giulianelli et al., 2021), most prior work has treated linguistic units (e.g., sentences) in isolation (e.g., Frank and Jaeger, 2008; Mahowald et al., 2013). In practice, however, linguistic communication often involves the incremental production and interpretation of linguistic material grounded in both perceptual and discourse context. Recent psycholinguistic research finds that this contextual grounding facilitates communication in naturalistic settings (such as face-to-face conversation, Drijvers and Holler, 2023) and in more controlled experimental tasks (such as the Image-Conditioned Maze Task proposed by Pushpita and Levy, 2024).

In this paper, we present—to the best of our knowledge—the first investigation of the UID hypothesis in a grounded setting. Specifically, we estimate linguistic surprisal conditioned on visual stimuli (following Haley et al., 2025) and assess

[*] Research done while visiting the University of Edinburgh.

[1] As in Jaeger (2010), "*the term 'choice' does not imply conscious decision making. It is simply used to refer to the existence of several different ways to encode the intended message into a linguistic utterance.*"

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3825–3853
March 24-29, 2026 ©2026 Association for Computational Linguistics

how such grounding influences the distribution of information density. We adopt a typological perspective, including 33 languages from 11 families and 2 macro-areas[2] as defined in Dryer and Haspelmath (2013). First, we analyse a dataset of image–caption pairs (Haley et al., 2025). We then further extend this analysis by jointly examining the roles of perceptual and discourse context, drawing on visual storytelling datasets (Leong et al., 2022). To measure surprisal, we leverage multilingual vision-and-language models. Specifically, we utilise word-level surprisal estimates for image captions derived from PaliGemma (Beyer et al., 2024) — as provided in GROUND-XM3600 — and we further extract surprisal values for visual narratives using Gemma 3 (Team et al., 2025). This allows us to study, via visual grounding and controlled contextual ablation, the contribution of multimodal information to the dynamics of information flow in texts associated with visual stimuli.

In particular, we study the following research questions: (1) How is the information density of an utterance shaped by grounding on visual perception? (2) How do discourse and perceptual context interact in modulating information density in visual narratives? (3) How does information density evolve over the time course of visually grounded discourse? We find that grounding on images robustly lowers surprisal variance across typologically diverse languages, both locally between adjacent words and globally over entire utterances. In visual narratives, discourse and perceptual context yield additional variance reduction of surprisal, with the strongest smoothing effects concentrated at sentence and paragraph onsets. Uniformity increases over the course of a story, at first abruptly and then slowly, when perception and discourse are considered. Overall, these findings support the view that perceptual and discourse context enables speakers to distribute information more evenly and thereby communicate more efficiently. We release code and data for our experiments at `https://github.com/Imatgay/Grounded-UID`.

2 Background

Information, context, structure. Most computational and psycholinguistic studies of UID have operated in unimodal, text-only regimes. Much of this literature has focused on explaining speakers' syntactic choices—when alternative constructions are available—in terms of a pressure toward UID at the sentence or clause level (e.g., Frank and Jaeger, 2008; Jaeger, 2010; Gamboa et al., 2024). At a broader level of granularity, the UID hypothesis has also been employed as a framework for interpreting the dynamics of information flow in extended discourse (Genzel and Charniak, 2002; Keller, 2004; Giulianelli and Fernández, 2021; Verma et al., 2023). Research on discourse-level information dynamics has frequently intersected with analyses of structural boundaries (e.g., paragraph breaks or conversational turns) and their effect on surprisal distribution, both in monological (Genzel and Charniak, 2003; Tsipidi et al., 2024, 2025) and dialogical contexts (Xu and Reitter, 2018; Giulianelli et al., 2021).

Nevertheless, existing work on UID and surprisal contours has yet to consider the influence of extralinguistic, multimodal context on information distribution. In addition, the empirical scope of these studies has been typologically narrow, focusing primarily on a limited set of Indo-European languages.

Visual grounding. The role of image-mediated context in shaping linguistic expectations has been explored in two recent studies, both operationalising groundedness as a reduction in word-level surprisal under multimodal conditioning via vision-and-language models. Haley et al. (2025) adopt a typological and information-theoretic perspective, treating images as language-agnostic proxies for utterance meaning and defining and measuring *groundedness* across 30 typologically diverse languages. Instead, Pushpita and Levy (2024) approach the problem from an experimental psycholinguistic angle, treating images as visual stimuli and finding that surprisal reductions derived from four VLMs (trained with distinct alignment objectives) predict human reading times in an *Image-Conditioned Maze Task*.

These two perspectives are jointly informative for the present study. On the one hand, the surprisal difference operationalised by Haley et al. (2025) allows us to treat groundedness dynamics as a signal of *sequence-level contentfulness*. On the other hand, Pushpita and Levy (2024) situate visual context within a psycholinguistically plausible model of language processing, validating multimodal surprisal as a cognitively interpretable measure of online comprehension effort. Although VLM-based surprisal is still an emerging psychometric proxy

[2] See Table 3 in Appendix A for the detailed taxonomy.

requiring further validation, together, these frameworks support our use of VLM-predicted surprisal contours not only as a computational lens on information distribution, but also as a tentative approximation of ecologically grounded language use. Unlike the present work, both studies, consistent with their respective aims, abstract away from the *temporal* organisation of surprisal, focusing exclusively on pointwise predictability.

Visual World Paradigm (VWP). Psycholinguistic VWP studies consistently demonstrate that comprehenders exploit visual environments to generate anticipatory eye movements in response to unfolding linguistic input (for a summary, see Huettig et al., 2011). Foundational works show that such visual grounding facilitates syntactic disambiguation (Tanenhaus et al., 1995), enhances sub-lexical prediction (Dahan et al., 2001), and supports the expectation of upcoming referents based on visual scene constraints (Altmann and Kamide, 1999).

More recently, Ankener et al. (2018) show that *multimodal surprisal*—reflecting the integration of visual and linguistic input—predicts cognitive load as measured through pupil dilation and N400 amplitudes, with the surprisal of nouns decreasing under lower visual ambiguity. These findings provide experimental evidence that anticipatory and responsive uncertainty during online processing are not determined solely by the linguistic signal, but reflect dynamically updated expectations conditioned on concurrent multimodal context.

Research gap. Computational psycholinguistics has modelled the temporal dynamics of information flow via surprisal contours, but this work has largely been limited to unimodal, text-only contexts. In contrast, experimental studies have examined language processing in more ecologically plausible, multimodal settings, but they have primarily focused on local facilitation effects at specific lexical targets. *As a result, the broader sequential structure of information in multimodal discourse remains empirically unexplored.* This gap is theoretically significant: if multimodal context facilitates language processing, its effects should extend beyond isolated lexical items to shape the global organisation of information across time.

3 Research Questions and Hypotheses

We combine approaches from two strands of psycholinguistic research, computational and experimental, to study information density across multimodal discourse. Specifically, using multilingual vision-and-language models, we examine how discourse context and visual grounding jointly shape the temporal distribution of surprisal—first in image–caption pairs, and then across extended narrative segments. Our analysis is guided by the following three research questions.

RQ1: ***How is the information density of an utterance shaped by grounding on visual perception?*** Image conditioning makes text more easily predictable: this facilitation effect, quantified as reduction in word surprisal, extends beyond relatively straightforwardly grounded words, consistent with the *comprehensive-grounding hypothesis* (Pushpita and Levy, 2024; Haley et al., 2025). Specifically, visual grounding induces a non-uniform reduction in surprisal across words, generally more pronounced for words traditionally classified as *content* words than for *function* words. Since content words tend to be higher-surprisal in text-only settings, such a decrease should reduce the variance of the surprisal distribution by narrowing its overall spread. As a consequence, we expect that information is more evenly distributed when grounded, yielding a flatter surprisal contour.

RQ2: ***How do discourse and perceptual context interact in modulating the information density in visual narratives?*** We examine how contextual grounding—both textual and visual—modulates information density at the paragraph level in visual storytelling. We hypothesise that discourse context, which shapes the common ground to constrain interpretation and topic, and visual context, which anticipates the referential content of an utterance, independently contribute to reducing surprisal variance. Crucially, we posit that their joint availability—offering complementary cues about both past and upcoming content—yields an additional smoothing effect on information density.

RQ3: ***How does information density evolve over narrative time?*** Building on recent proposals that information flow in discourse exhibits harmonic and hierarchical structuring (Tsipidi et al., 2024, 2025), we examine how the presence of multimodal contextual material modulates surprisal across nested narrative levels (sentences within paragraphs, paragraphs within stories), thereby affecting UID. We hypothesise that: (i) discourse and perceptual context jointly induce a downward drift in surprisal and UID over narrative time. This would be consistent with the accumulation of contextual constraints that increasingly narrow the hypothesis space for upcoming content, thereby reducing predictive uncertainty and flattening the surprisal contour over time. We also hypothesise that (ii) contextual information—whether in the form of discourse history carrying background knowledge, or visual *semantic priming*—primarily attenuates the local discontinuities in surprisal observed at discourse boundaries, as these transitions introduce *surprising* new referents or thematic shifts (Genzel and Charniak, 2002; Tsipidi et al., 2025) that multimodal conditioning can help to anticipate. These constitute one of the main structure-driven sources of deviation from an optimal instantiation of the UID hypothesis.

4 Method

At the most fundamental level, the method for addressing these questions consists of: (i) computing metrics for information uniformity (UID values) for an utterance under different conditions (e.g., in isolation, conditioned on an image, or conditioned on prior discourse); and (ii) comparing these UID values to assess how the conditioning context influences the distribution of information. We adhere to the classical information-theoretic model, quantifying the information content of a linguistic unit with *surprisal*.[3]

Surprisal estimation. We estimate surprisal at the word level with pre-trained autoregressive vision-and-language models and language models. Surprisal is defined as the information content of a word w_t given its preceding context. Concretely, for a word w_t in an utterance $\mathbf{u} = (w_1, \ldots, w_n)$, we compute the surprisal s_t as $-\log p(w_t \mid w_1, \ldots, w_{t-1}, \mathbf{c})$, where $\mathbf{c}$ is some additional context (e.g., perception or discourse). The conditional probability distribution $p(\cdot)$ is provided by the next-word distribution under a given pre-trained model. Since this operates on subword units—in our case tokenised via SentencePiece (Kudo and Richardson, 2018)—we reconstruct word-level surprisal by summing the negative log probabilities of the tokens each word comprises and we adopt the correction for trailing whitespaces introduced by Oh and Schuler (2024) and Pimentel and Meister (2024).[4] This gives us a vector of surprisals $\mathbf{s}$ corresponding to each word in $\mathbf{u}$. Word segmentation—when not available—was determined using Stanza (Qi et al., 2020) for all languages except Bengali, for which we use BNLP (Sarker, 2021).

UID computation. Once a surprisal vector $\mathbf{s}$ has been estimated, as described above, we compute UID values using two variance-based metrics introduced by Collins (2014), and later formalised by Meister et al. (2021). The first, given in Equation (1), captures uniformity at the *global* sequence level (e.g., a sentence or a paragraph):

$$\mathrm{UID}_v(\mathbf{s}) = \frac{1}{n}\sum_{t=1}^{n}\big(s_t - \mu\big)^2 \tag{1}$$

where $\mu = \frac{1}{n}\sum_{t=1}^{n} s_t$ is the mean surprisal over the n words in the sequence. This metric quantifies the extent to which the surprisal values of individual units (e.g., words) deviate from the overall sequence average. The second metric, defined in Equation (2), instead focuses on *local* uniformity, quantifying UID as the average change in surprisal between consecutive units within a sequence:

$$\mathrm{UID}_{lv}(\mathbf{s}) = \frac{1}{n-1}\sum_{t=2}^{n}\big(s_t - s_{t-1}\big)^2 \tag{2}$$

In both cases, lower values indicate more uniform surprisal profiles. We take the local version of UID to be more sensitive to fine-grained grammatical and locality-driven effects in language, whereas the global metric is more useful for analysing broader, discourse-level patterns of information transmission.

4.1 Experimental Setup

For an utterance $\mathbf{u}$, we consider these possible contexts: **Utterance** (U), where the utterance is evaluated in isolation, without any preceding linguistic or visual context; **Perception** (P), where the utterance is conditioned on a corresponding image; and **Discourse** (D), where the utterance is conditioned on preceding utterances (when available).

To address our research questions, we compute UID values of textual sequences under four distinct conditions: U, P, D, and P + D. The last condition corresponds to the visual storytelling setting (ViSt; Huang et al., 2016), in which surprisal

[3] A range of alternative information measures may be suitable for such analyses (Rabovsky et al., 2018; Aurnhammer and Frank, 2019; Giulianelli et al., 2023, 2024b, 2026; Meister et al., 2024; Li and Futrell, 2024, *inter alia*); we leave these for future work.

[4] For a broader discussion of these issues and their implications for surprisal estimation, see Giulianelli et al. (2024a).

is computed for each utterance based on the full available context—i.e., all the preceding textual sequences and the associated images.

Models and Datasets. We consider two multimodal datasets for our experiments: the GROUND-XM3600 dataset (Haley et al., 2025)[5] for image–caption pairs (specifically, the CROSSMODAL-3600 split; Thapliyal et al., 2022) and the BLOOMVIST dataset (Leong et al., 2022)[6] for visual storytelling (image-interleaved text). We report the dataset statistics for both in Appendix A.

GROUND-XM3600 provides word-level surprisal values for 213,677 image–caption pairs across 30 typologically diverse languages from 10 families (Table 1). Each word is annotated with two surprisal values, computed under two different conditions: (i) P with the `paligemma-3b-ft-coco35-224` checkpoint of PaliGemma (Beyer et al., 2024) as a vision-and-language model (VLM); and (ii) U starting with the decoder of pre-trained `paligemma-3b-pt-224` used as a language model (LM), which was then fine-tuned on the captions of COCO35L. This ensures exact overlap between the textual data observed by the VLM and its LM counterpart.

BLOOMVIST is a split of the Bloom Library (Leong et al., 2022), containing 11,407 stories structured as interleaved image–paragraph sequences, with a total of 112,080 image–paragraph pairs. An example of a data sample is provided in Figure 5 (Appendix B). Of the original 363 languages represented, we include only the 48 with a number of stories sufficiently large for our analyses (> 20). We truncate stories after the 20th paragraph as the models' surprisal values may become unreliable for exceedingly long sequences. Moreover, to reduce noise and exclude edge cases, we discarded stories containing at least one paragraph with fewer than three words. For this dataset, we used Gemma 3 4B pre-trained (Team et al., 2025)[7] as a unified instantiation of both VLM and LM. This model natively supports both multimodal and unimodal (text-only) input formats, removing the need for additional fine-tuning, and can handle long contexts (up to 128K tokens) with interleaved text-and-image inputs. However, unlike PaliGemma, the composition of Gemma 3's training data mixture remains undisclosed—including the exact languages it covers. Based on empirical heuristics,[8] we ultimately selected the 13 languages (from 7 language families) listed in Table 2.

Metrics. To assess whether visually grounded language exhibits different information contours than text-only language, we examine information uniformity across both GROUND-XM3600 and BLOOMVIST. We compute UID_v and UID_{lv} according to Equations (1) and (2) respectively. For GROUND-XM3600, each caption is treated as a single utterance. For each language, we aggregate UID metrics across all captions, reporting mean UID_v and UID_{lv} under the U and P conditions, and compute their relative difference Δ to quantify the impact of visual grounding. As for BLOOMVIST, word surprisal estimates are obtained using Gemma 3 under U, P, D, and $\text{P} + \text{D}$. UID_v is then computed under all four conditions at the paragraph and sentence level.[9] To make the comparison fair across conditions where the average density differs, we also report the Coefficient of Variation (CV $= \frac{\sigma(\text{s})}{\mu(\text{s})}$), which is a unitless (scale-free) measure of how uneven the densities are.

To investigate how discourse and perceptual context modulate the organisation of information throughout a narrative, we adopted two complementary measures: (i) linear mixed-regression modelling of information-theoretic metrics against relative position within discourse; and (ii) density estimation of surprisal reductions over normalised narrative positions induced by each kind of context. First, adapting the methodology of Giulianelli and Fernández (2021), we fitted linear mixed-effect models for each language and level of granularity (sentence or paragraph), predicting surprisal and UID_v as a function of the relative position of a unit within the immediately higher structural level (in our case, sentence < paragraph < story), under different context conditions (U, P, D, $\text{P} + \text{D}$). The resulting slopes coarsely quantify the drift of information flow under different contexts. Second, we examined the distribution of positive surprisal reductions. For each word, we computed three

[5] https://osf.io/bdhna/

[6] https://huggingface.co/datasets/sil-ai/bloom-vist

[7] https://huggingface.co/google/gemma-3-4b-pt

[8] Character-level perplexity over BLOOMVIST stories was used to contrast our language model with one with randomly initialised embeddings, filtering out languages with moderate and low discrepancies. High perplexity was tolerated in languages with high intrinsic character entropy (e.g., Mandarin Chinese), under the assumption of sufficient representation in the training data.

[9] Paragraphs consist of text interleaved with images according to the predefined alignment of the dataset. Sentences are extracted with Stanza's sentence tokenizer (Qi et al., 2020).

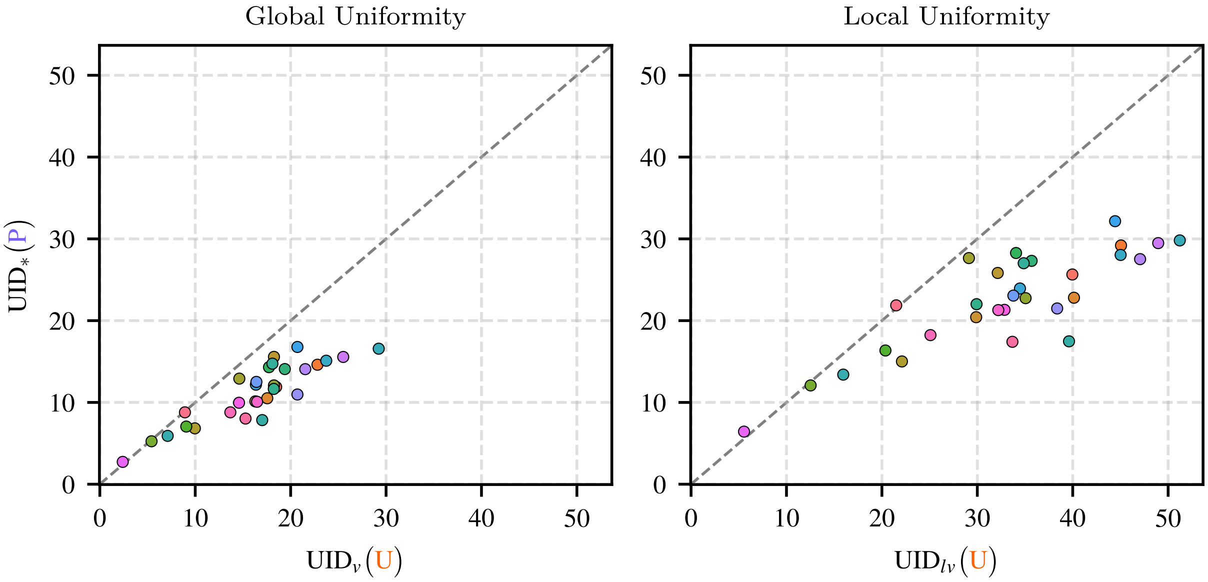

Figure 1: Global and local UID values under the U condition (caption in isolation) and P condition (caption with visual context) across the 30 typologically diverse languages, from 10 language families, in GROUND-XM3600 dataset. Points below the dashed diagonal line indicate increased uniformity (i.e., decreased global or local variability) when the caption's surprisal is conditioned on the image context.

reduction scores (Δ_P, Δ_D, Δ_{P+D}), reflecting the contextual facilitation due to local perceptual, discourse, and joint multimodal context, respectively.

5 Results

5.1 How does Visual Grounding Shape Utterance Information Density?

First, to quantify the impact of visual grounding on textual information distribution, we compared sentence-level UID values—both global (UID_v) and local (UID_{lv})—across 30 languages in the GROUND-XM3600 dataset, as described in Section 4.1. For every caption, we computed UID from word-level surprisal under U and P conditions. We then performed paired Wilcoxon signed-rank tests (Wilcoxon, 1945) per language and UID metric to assess the statistical significance of UID changes due to grounding, while computing standardised effect sizes (paired Cohen's d_z). Results are shown in Figure 1 (for complete data, see Table 4 in Appendix C).

The hypothesis formulated in Section 3 is confirmed in virtually all languages, for both local and global UID metrics: grounding in visual perception reliably reduces UID values, indicating a more uniform information distribution across utterances. Exceptions are limited to Telugu, where both UID_v and UID_{lv} significantly increase when grounded, and Arabic, where only UID_{lv} does. We suspect that these anomalous cases may reflect limitations in the model's vision–language alignment or be influenced by script-specific and morphological properties that affect token segmentation (Park et al., 2021; Petrov et al., 2023; Oh and Schuler, 2025). It should be further noted that UID decreases in all other languages by different degrees, possibly for similar reasons. The shift induced by grounding is further illustrated in Figure 2, which shows the estimated density of UID values, averaged across all sentences and languages. For both metrics, the

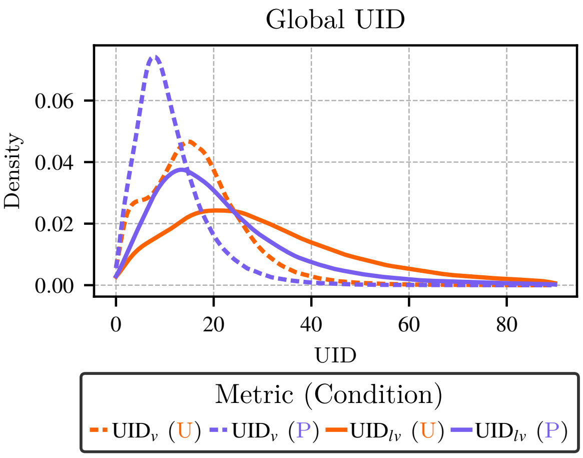

Figure 2: Kernel density estimation of global UID values for captions in GROUND-XM3600, conditioned either on their respective image (P) or without any contextual image (U). Curves are truncated at the 99th percentile to enhance visual clarity.

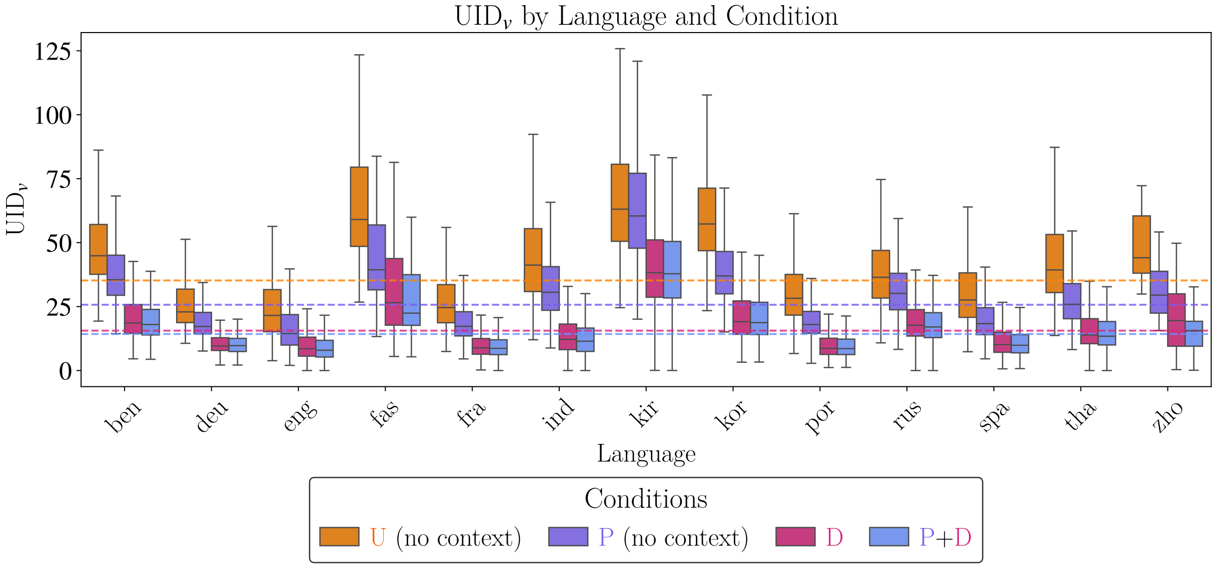

Figure 3: Each box represents the distribution of paragraph-level UID$_v$ values across languages under four conditions: no context (U); paired image as context (P); all preceding paragraphs as context (D); all preceding paragraphs and interleaved images as context (P+D). Dashed horizontal lines indicate the mean UID$_v$ per condition, averaged across all languages.

density curves under visual grounding (P) are consistently shifted leftwards compared to their text-only (U) counterparts, with a more pronounced concentration of low UID values.

5.2 How does the Interaction of Discourse and Perceptual Context Modulate Information Density?

Secondly, we extend our analysis to the BLOOMVIST dataset to include discourse context (D). Specifically, for every image–paragraph pair within a story, we estimated UID metrics under the four conditions listed in Section 4.1, as shown in Figure 3. A clear bifurcation emerges between conditions without prior discourse (U, P) and those incorporating it (D, P + D), with the latter group exhibiting consistently lower median UID values. Within each group, visually grounded conditions (P and P + D) yield lower UID values than their text-only counterparts (U and D), indicating a robust contribution of perceptual context across discourse context regimes. Notably, the joint condition P + D on average yields the lowest UID values, and its distribution closely tracks that of D. This indicates that discourse context accounts for a relatively larger reduction in UID, while perceptual context contributes a small but consistent additional smoothing when both contextual modalities are available. To statistically test for the hypothesised monotonic ordering of UID values (U > P > D > P+D), we applied the Page test for ordered alternatives (Page, 1963) independently for each language. The test yielded highly significant results across all languages ($p < 0.001$), providing robust support for the hypothesised hierarchy of contextual effects on UID.

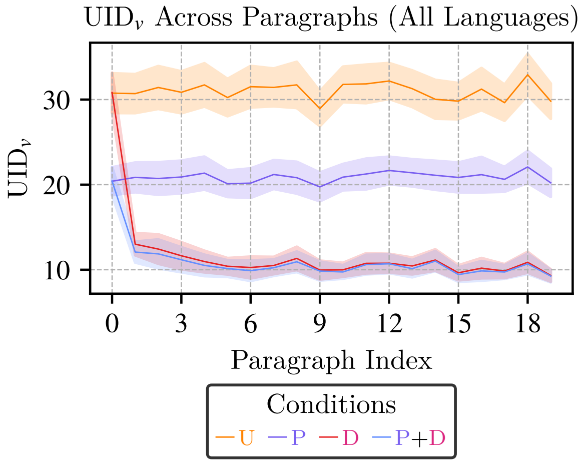

Figure 4: Paragraph-level UID values averaged across BLOOMVIST stories (truncated at 20 paragraphs) under the utterance only (U), perceptual context (P), discourse context (D), and multimodal context (P+D) conditions.

5.3 How do Visually Grounded Information Contours Evolve over Narrative Time?

Figure 4 plots mean UID$_v$ for each paragraph index in visual stories with a maximum of 20 paragraphs, under the four conditions we study. UID trajectories under U and P are relatively flat, whereas discourse-informed conditions (D, P + D) show a marked drop over the early paragraphs and then slowly flatten out. The curves for D and P + D get increasingly closer as more discourse context accumulates. This suggests that the textual history eventually encodes as much information about a paragraph as the image paired with it.

Fitting linear mixed-effects models of UID and mean surprisal against relative (normalised within-unit) position at the paragraph and sentence level yields slope values, which help us determine the evolution of information flow under different conditions. These slopes are extracted from the interaction between position and context condition, with U serving as the baseline. In the U (no context) and P (image context) conditions, surprisal and UID$_v$ tend to drift negatively over narrative time at the sentence level, but show weaker or inconsistent trends at the paragraph level, with most paragraph-level slopes being statistically non-significant (see Tables 7 and 8 in Appendix F). This indicates that without conditioning information estimates on global discourse context, no meaningful discourse-level patterns can be observed.

In contrast, the introduction of discourse (D) and multimodal (P + D) context conditions induces consistent and significantly negative slopes at both sentence and paragraph levels across languages. This suggests that mean surprisal and UID scores of discourse units decrease as constraints accumulate over the course of the narrative. Notably, the negative slope becomes systematically less steep when images are added to discourse (P + D vs D), suggesting that additional contextual information leads to higher global stability throughout the paragraph or story.

This global flattening effect raises the question of *where* within discourse units contextual information exerts its strongest impact. If surprisal is primarily concentrated at unit onsets, as Tsipidi et al. (2025) suggest, we might expect the observed variance reduction to be driven, in large part, by the local smoothing of those unit-initial surprisal peaks. Across all plotted languages and levels (sentence and paragraph), surprisal reductions attributable to perceptual (Δ_{P}), discourse (Δ_{D}), and combined context ($\Delta_{\text{P+D}}$) exhibit a consistent front-loaded distribution (see Appendix G), peaking sharply at the onset of the unit and decaying rapidly thereafter. This pattern indicates that contextual information—whether visual, textual, or multimodal—exerts the strongest disambiguating and constraining effect at the beginning of sentences and paragraphs. The effect is especially pronounced for discourse context (Δ_{D}), which shows the highest density near position zero in nearly all cases. Multimodal context ($\Delta_{\text{P+D}}$) behaves similarly but tends to yield smoother reductions with longer tails, suggesting more sustained facilitation throughout the unit.

6 Further Discussion

6.1 A Scale-free Metric of Uniformity

To account for possible discrepancies in mean surprisal across conditions, which may affect variance, we also computed the Coefficient of Variation as a scale-free measure of dispersion, as described in Section 4.1. The results of this analysis are reported in Tables 5 and 6, Appendix D. In GROUND-XM3600, visual grounding significantly increases CV across all languages ($p < .001$), in contrast with the observed reduction in UID values. This is most likely due to μ decreasing proportionally more than σ in short captions, where images collapse non-onset surprisals. For instance, in *"A polar bear is swimming"*, U surprisals $[10.34, 7.87, 0.08, 0.98, 2.95]$ ($\mu = 4.44, \sigma = 4.46, CV = 1.01$) shift under P to $[10.45, 0.49, 0.01, 1.43, 0.39]$ ($\mu = 2.55, \sigma = 4.44$), yielding $CV = 1.74$. Such behaviour suggests the metric may be less reliable than raw variance when predicting processing effort in short-form referential language conditioned on images. For this reason, we think the interplay between alternative dispersion metrics and standard UID measures warrants further systematic investigation in the future.

Conversely, BLOOMVIST exhibits more coherent dynamics across standard UID and scale-free metrics, with grounding typically preserving or reducing CV. Unlike isolated captions, narrative paragraphs involve complex dynamic discourse structures where visual context disambiguates the semantic content in a more distributed fashion, preventing the collapse of μ.

6.2 Linguistic Analysis

Finally, we propose a simple linguistic analysis of instances of UID reduction failure under perceptual grounding—i.e., cases in which conditioning on perceptual context does not lead to a reduction in UID. By leveraging Part-of-Speech (POS) annotations in GROUND-XM3600, we decompose

UID_v into word-level variance contributions across POS categories. This allows us to quantify how the contribution of each POS category changes when utterances are grounded in visual perception (P) relative to text-only conditions (U). We focus on cases in which perceptual grounding increases sentence-level variance, despite the overall tendency for grounding to reduce UID.

The cross-linguistic heatmap (Figure 6 in Appendix E) visualises the average change in these variance contributions for sentences where $\text{UID}_v(\text{P}) > \text{UID}_v(\text{U})$. Across typologically diverse languages, Proper Nouns (PROPN), Numerals (NUM), and Adjectives (ADJ) emerge as the primary drivers of non-uniformity (that is, words whose surprisal under P deviated strongly from the sentence-level mean). As further shown in Appendix E, higher variance contributions under P relative to U typically coincide with an increase in absolute surprisal for these POS categories.

These findings warrant further investigation and raise questions regarding the role of specific POS categories in UID and the robustness of VLMs as grounded surprisal estimators. For instance, in vision–language models, adjectives are often the main remaining source of uncertainty. In line with Buettner and Kovashka (2024)'s notion of "attribute insensitivity" at the representation level, VLMs may recognise referents in the image better than their attributes.

7 Conclusions and Future Work

We evaluated how grounding utterances on visual perception and discourse affects information contours. Leveraging multilingual vision-and-language models and established information-theoretic metrics, we studied datasets of captioned images (30 languages spanning 10 families) and visual storytelling (13 languages, 7 families).

Our findings indicate that when utterances are conditioned on perceptual context, information is distributed more evenly across the textual sequence than when utterances are considered in isolation, both in image captions and in visually grounded stories. In visual storytelling, discourse context alone induces a substantial smoothing of the surprisal contour, yet the incorporation of visual context to discourse yields a consistent, albeit small, additional reduction in surprisal variance. In light of recent accounts of information dynamics in extended discourse—such as the Structured Context hypothesis (Tsipidi et al., 2024) and the Harmonic Surprisal hypothesis (Tsipidi et al., 2025)—we further suggest that surprisal reduction from grounding exhibits systematic structure across scales: contextual information exerts its strongest effect at unit onsets, both at the sentence and paragraph level. This strongly front-loaded profile indicates that discourse and perceptual grounding primarily enhance UID acting at structural boundaries, which we identify as key loci for contextual integration in the regulation of information flow.

Moreover, building on Haley et al. (2025)'s interpretation of images as "*language-agnostic representations of meaning*", we extended their line of inquiry beyond static image–caption pairs—where captions are typically tightly and referentially linked to their associated images—to the looser, more anticipatory image–paragraph pairings found in visual storytelling. In this latter setting, the relationship between image and text is not strictly referential but often reflects broader narrative structure and event dynamics. Within this framework, we find suggestive evidence that images can convey higher-level discourse-semantic content, thereby modulating the global information distribution of subsequent utterances. This points to a more abstract form of visual grounding, in which perceptual context contributes not only to local lexical predictability but also to the shaping of information flow at the level of discourse structure.

A natural extension of the present work is to move beyond static images toward dynamic perceptual streams, such as video, where the temporal structure of the input can more tightly constrain linguistic predictions. Unlike static images, dynamic stimuli unfold over time and encode event-level structure (e.g., causality, agency, temporal progression), potentially offering a richer basis for modelling how perceptual context informs discourse organisation. Additionally, these findings suggest the need for formal decomposition into redundant and synergistic contributions (e.g., via Partial Information Decomposition; Luppi et al., 2024), enabling a more precise characterisation of how modalities combine. Finally, inspired by Pushpita and Levy (2024), it would be valuable to complement our computational modelling analysis with a tailored behavioural experiment designed to probe the cognitive correlates of the effects we observe.

Limitations

This study is subject to three main limitations. First, while we adopt a typologically informed approach, the intersection between available datasets and model capabilities effectively constrains our choice of languages. In particular, although BLOOMVIST provides coverage of several under-resourced languages, most of these fall outside the training distribution of Gemma 3, resulting in a final language set heavily biased towards Indo-European languages and limited to two WALS macro-areas (Eurasia and Africa, per Dryer and Haspelmath, 2013). Second, our analysis is computational, without psycholinguistic validation: while model-based surprisal estimates are supported by recent studies (Pushpita and Levy, 2024), no experimental data yet links these predictions to actual human processing patterns in multimodal discourse; this underscores the need for complementary human studies to validate our findings. Third, the current availability of open-weight multilingual VLMs with long-context processing capabilities remains very limited, as does the availability of multilingual visual storytelling datasets. This restricts the ability to test the robustness of our findings across different models and data sources.

Acknowledgements

Matteo Gay acknowledges the School of Advanced Studies IUSS Pavia for funding his research visit to the University of Edinburgh, where he conducted the research presented in this paper.

References

Gerry T.M Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Christine S. Ankener, Heiner Dernhaus, Matthew W. Crocker, and Maria Staudte. 2018. Multimodal Surprisal in the N400 and the Index of Cognitive Activity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

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. *Journal of the Acoustical Society of America*, 119(5):3048–58.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. PaliGemma: A versatile 3B VLM for transfer. *Preprint*, arXiv:2407.07726.

Kyle Buettner and Adriana Kovashka. 2024. Investigating the Role of Attribute Context in Vision-Language Models for Object Recognition and Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5474–5484.

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.

Michael X. Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681.

Delphine Dahan, James S. Magnuson, and Michael K. Tanenhaus. 2001. Time Course of Frequency Effects in Spoken-Word Recognition: Evidence from Eye Movements. *Cognitive Psychology*, 42(4):317–367.

Gabriel Doyle and Michael Frank. 2015. Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 19–28, Denver, Colorado. Association for Computational Linguistics.

Linda Drijvers and Judith Holler. 2023. The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, 30(2):792–801.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Austin F. Frank and Tim F. 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.

Richard Futrell and Michael Hahn. 2022. Information Theory as a Bridge Between Language Function and Language Form. *Frontiers in Communication*, Volume 7 - 2022.

John C. B. Gamboa, Leigh B. Fernandez, and Shanley E. M. Allen. 2024. Investigating the Uniform Information Density hypothesis with complex nominal compounds. *Applied Psycholinguistics*, 45(2):322–367.

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.

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. 2026. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Journal of Memory and Language*.

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.

Coleman Haley, Sharon Goldwater, and Edoardo Ponti. 2025. A Grounded Typology of Word Classes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10380–10399, Albuquerque, New Mexico. Association for Computational Linguistics.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Falk Huettig, Joost Rommers, and Antje S. Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2):151–171. Visual search and visual world: Interactions among visual attention, language, and working memory.

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

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

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Andrea I. Luppi, Fernando E. Rosas, Pedro A. M. Mediano, David K. Menon, and Emmanuel A. Stamatakis. 2024. Information decomposition and the informational architecture of the brain. *Trends in Cognitive Sciences*, 28(4):352–368. Epub 2024 Jan 9.

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. Epub 2012 Oct 30.

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.

Byung-Doh Oh and William Schuler. 2024. Leading Whitespaces of Language Models' Subword Vocabulary Poses a Confound for Calculating Word Probabilities. *Preprint*, arXiv:2406.10851.

Byung-Doh Oh and William Schuler. 2025. The Impact of Token Granularity on the Predictive Power of Language Model Surprisal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.

Ellis Batten Page. 1963. Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, 58(301):216–230.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tiago Pimentel and Clara Meister. 2024. How to Compute the Probability of a Word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Subha Nawer Pushpita and Roger P. Levy. 2024. Image-conditioned human language comprehension and psychometric benchmarking of visual language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 447–457, Miami, FL, USA. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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.

Sagor Sarker. 2021. BNLP: Natural language processing toolkit for Bengali language. *Preprint*, arXiv:2102.00405.

Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423.

Michael K. Tanenhaus, Monica J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634. Erratum in: Science. 2005 Feb 11;307(5711):851.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eleftheria Tsipidi, Samuel Kiegeland, Franz Nowak, Tianyang Xu, Ethan Wilcox, Alex Warstadt, Ryan Cotterell, and Mario Giulianelli. 2025. The Harmonic Structure of Information Contours. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31636–31659, Vienna, Austria. 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.

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.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

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

Appendix

A Dataset Statistics

Dataset statistics for Ground-XM3600 (Table 1) and BloomVIST (Table 2). Table 3 provides the full list of the 33 languages categorised by WALS family and macro-area.

Lang	Caps	Words	Mean	Std
ARB	6928	50225	7.2	3.3
CES	5867	38299	6.5	5.0
DAN	6696	58845	8.8	4.4
DEU	8615	93753	10.9	3.8
ELL	6040	48281	8.0	4.0
ENG	7179	67956	9.5	3.9
FAS	7142	85867	12.0	6.3
FIN	6726	51444	7.6	4.2
FRA	8520	107899	12.7	4.6
HEB	6572	74965	11.4	8.2
HIN	8501	114194	13.4	5.2
HRV	6555	59564	9.1	4.9
HUN	6711	60706	9.0	5.5
IND	7115	95210	13.4	5.5
ITA	8448	100852	11.9	4.0
JPN	7067	109036	15.4	8.7
KOR	7522	68462	9.1	4.2
NLD	7494	59313	7.9	3.7
NOR	7014	62969	9.0	4.2
POL	6999	64241	9.2	4.7
POR	7023	71809	10.2	5.8
RON	7048	112008	15.9	10.3
RUS	7136	75043	10.5	5.2
SPA	8597	77948	9.1	3.1
SWE	6514	54298	8.3	3.9
TEL	7195	53807	7.5	2.0
TUR	6849	63026	9.2	6.5
UKR	7061	70935	10.0	5.9
VIE	6699	100610	15.0	6.9
ZHO	5834	84401	14.5	10.3

Table 1: Summary statistics of caption lengths across languages in the Ground-XM3600 dataset. *Caps*: number of captions; *Words*: total word count; *Mean* and *Std*: average and standard deviation of caption lengths (in words).

Lang	Stories	Pars	Words	W/P	W/S
BEN	217	1919	58198	30.33	268.19
DEU	21	220	10827	49.21	515.57
ENG	1997	19602	766901	39.12	384.03
FAS	125	599	8128	13.57	65.02
FRA	294	3444	136148	39.53	463.09
IND	225	1774	30659	17.28	136.26
KIR	265	2679	109642	40.93	413.74
KOR	129	2114	44233	20.92	342.89
POR	143	2270	68755	30.29	480.80
RUS	262	2818	142974	50.74	545.70
SPA	426	4275	134219	31.40	315.07
THA	259	2762	89090	32.26	343.98
ZHO	34	221	2796	12.65	82.24

Table 2: BloomVist data: number of stories, paragraphs (Pars), words, mean words per paragraph (W/P), and words per story (W/S).

Macro-area	WALS Family	ISO
Africa (+ Eurasia)	Afro-Asiatic	ARB
Eurasia	Afro-Asiatic	HEB
	Indo-European	BEN, CES, DAN, DEU, ELL, ENG, FAS, FRA, HIN, HRV, ITA, NLD, NOR, POL, POR, RON, RUS, SPA, SWE, UKR
	Uralic	FIN, HUN
	Turkic	TUR, KIR
	Sino-Tibetan	ZHO
	Austro-Asiatic	VIE
	Austronesian	IND
	Tai-Kadai	THA
	Dravidian	TEL
	Japanese	JPN
	Korean	KOR

Table 3: Classification of the 33 analysed languages by macro-area and family, as defined in WALS (Dryer and Haspelmath, 2013).

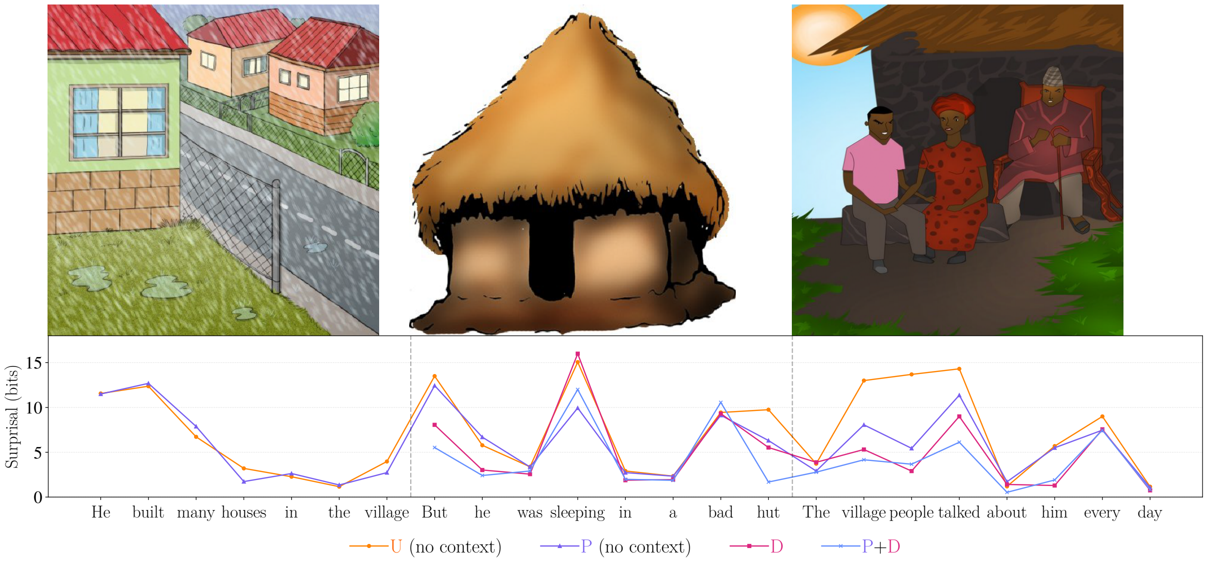

Figure 5: Surprisal contours for the first three paragraphs of a BLOOMVIST story (ID: dd0b0ef6-3889-47c7-b5ca-93e18e76aba8) across all four experimental conditions. Dashed vertical lines denote paragraph boundaries.

B BLOOMVIST Example

Figure 5 provides an example illustration of the surprisal dynamics in a visual narrative from the BLOOMVIST dataset. As a qualitative illustration, observe that concrete referential terms such as *hut* exhibit substantial surprisal reductions when conditioned on visual context. Conversely, abstract or non-referential descriptors like *sleeping* (where the action is not explicitly depicted) and *bad* (a qualification not strictly entailed *a priori* by the image) retain relatively high surprisal.

C Grounding Effect in GROUND-XM3600

Table 4 reports the full language-level results for RQ1, detailing UID_v and UID_{lv} values in the GROUND-XM3600 dataset under text-only and visually grounded conditions.

D Coefficient of Variation

In Table 5 (GROUND-XM3600) and Table 6 (BLOOMVIST), we report the results for the scale-free dispersion metric (Coefficient of Variation, CV) discussed in Section 6.1.

E Part-of-Speech Analysis

The heatmap in Figure 6 visualises the average change in variance contribution ($\overline{\Delta C}_{\text{POS}}$, defined in Equation (4)), across POS tags and languages, in GROUND-XM3600 examples. This metric quantifies the extent to which each POS category's contribution to sentence-level variance changes when utterances are grounded on visual perception (P) compared to text-only conditions (U). The analysis considers only sentences where visual grounding *increased* the global UID value (i.e., where $\text{UID}_v(P) > \text{UID}_v(U)$). We aggregate results for the 10 most common POS categories in Universal Dependencies (VERB, NOUN, ADJ, ADV, PROPN, ADP, DET, PUNCT, NUM, PRON). A POS-language pair was filtered out if fewer than 50 POS instances per language were present within the identified failure sentences. We define the word-level change in variance contribution as $\Delta C_w = C_{P,w} - C_{U,w}$ where $C_{X,w}$ is the contribution of word w to the total sentence variance (UID_v) under condition $X \in \{U, P\}$, i.e.:

$$C_{X,w} = \frac{(s_{X,w} - \mu_X)^2}{n} \tag{3}$$

where s is the word surprisal and μ_X is the mean surprisal over the n words in the sentence under condition X. We then compute the mean of these word-level shifts for each POS group within each language:

$$\overline{\Delta C}_{\text{POS}} = \frac{1}{N_{\text{POS}}} \sum_{w \in \text{POS}} \Delta C_w \tag{4}$$

where N_{POS} is the total number of words belonging to a POS category across the identified failure sentences for a given language.

Because this contribution is agnostic to the direction of change in surprisal, we further calculate

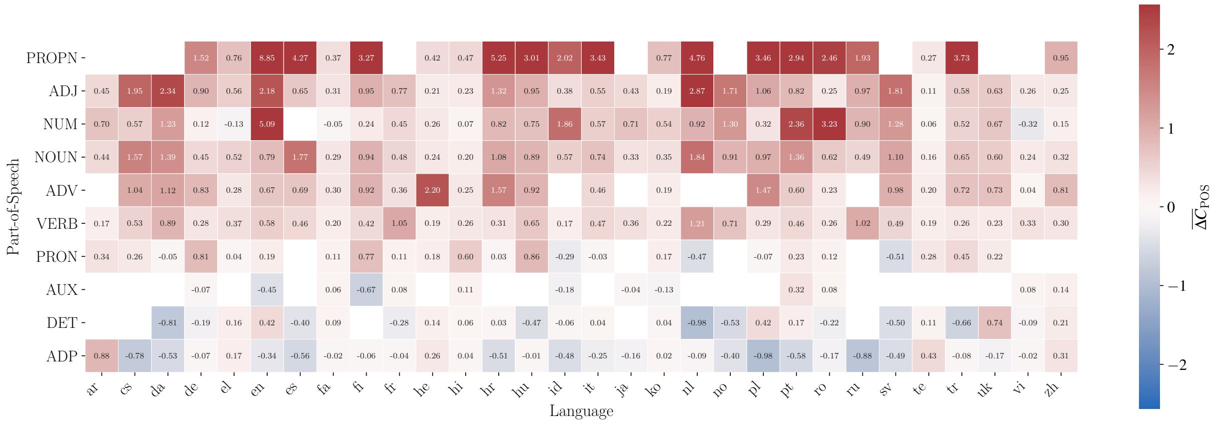

Figure 6: Average change in variance contribution ($\overline{\Delta C}_{\text{POS}}$) by POS across languages for UID reduction failure cases in GROUND-XM3600. Rows are ordered, from top to bottom, by descending cross-linguistic mean. Intuitively, red cells denote an increased divergence from the sequence mean under visual grounding (P) relative to text-only (U) conditions.

the percentage of words within each POS category that exhibit an absolute increase ($s_{\mathcal{P},w} > s_{\mathcal{U},w}$) or decrease ($s_{\mathcal{P},w} < s_{\mathcal{U},w}$) in surprisal. Results are plotted in Figure 7.

F Regression Model for RQ3

Design. To model the evolution of surprisal and information density over narrative time across different conditions, we implemented a set of linear mixed-effects regression models with the following structure:

$$y \sim \text{position} \times \text{condition} + \log(\text{length}) \\ + (1 + \text{position} \mid \text{story}) \tag{5}$$

Here, the dependent variable y corresponds to either the mean word-level surprisal or the global UID score (UID_v) computed over a given unit—either a sentence or a paragraph. The fixed-effect structure includes the unit's relative position within the next-higher structural level (i.e., sentence within paragraph or paragraph within story), a categorical variable for context condition (U, P, D, P + D), and their interaction. This interaction term allows the slope of surprisal or UID over narrative time to vary across conditions, with the baseline (U) serving as the reference level. A log-transformed control for unit length (in words) is included to account for length-related variance in surprisal and information distribution.

The model also includes random intercepts by story to account for between-narrative variability, and random slopes over position by story. This structure tests whether the trajectory of surprisal or UID over discourse position changes systematically as a function of contextual grounding.

Implementation. Models were fit independently for each language and unit type (sentence or paragraph). Sentence-level analyses required that each paragraph contain at least three sentences.

The response variable was regressed on relative position, context, their interaction, and unit length. After fitting, we extracted the fixed-effect slopes per context and dependent variable. Each slope represents the estimated rate of change in surprisal or UID over the course of the narrative, conditional on the availability of contextual information. These results are reported in Table 7 and Table 8.

G Surprisal Reduction Densities

Figures 8–20 display the distribution of positive surprisal reductions across the relative position of words within sentences and paragraphs, for each language in BLOOMVIST. The densities are estimated by computing histograms over normalized positions and then smoothed via Gaussian filters. The three curves in each plot correspond to:

- Δ_{P}: reduction due to local visual context (only previous image) vs. uncontextualised text, (i.e., Surprisal(U − P)).
- Δ_{D}: reduction due to discourse-level context vs. sentence-level context (i.e., Surprisal(U − D)).
- $\Delta_{\text{P+D}}$: reduction due to global visual context vs. text-only discourse level (i.e., Surprisal(D − [P + D])).

H Surprisal Discontinuities at Discourse Boundaries

The tables below report differences around sentence (Table 9) and paragraph (Table 10) boundaries for each language in BloomVIST, quantifying information spikes at discourse unit onsets for varying window sizes.

I Compute Budget

Surprisal values from BloomVIST were extracted using the 4B-parameter Gemma 3 model, loaded with HuggingFace's Transformers library.[10] The full extraction required approximately 24 hours of compute on a single NVIDIA A100-80GB GPU.

[10] https://pypi.org/project/transformers/

Lang	UID$_v$ (U)	UID$_v$ (P)	UID$_{lv}$ (U)	UID$_{lv}$ (P)	Δ_v (%) (d)	Δ_{lv} (%) (d)
arb	8.92	8.79	21.50	21.87	-1.44 * (-0.02)	**1.75** *** (0.02)
ces	18.48	11.89	39.96	25.64	-35.67 *** (-0.72)	-35.83 *** (-0.54)
dan	22.81	14.61	45.06	29.19	-35.93 *** (-0.93)	-35.22 *** (-0.64)
deu	17.55	10.51	40.13	22.80	-40.09 *** (-0.95)	-43.18 *** (-0.87)
ell	16.30	10.13	29.88	20.42	-37.89 *** (-0.79)	-31.66 *** (-0.53)
eng	18.25	15.57	32.14	25.83	-14.70 *** (-0.35)	-19.63 *** (-0.31)
fas	9.96	6.83	22.10	15.01	-31.45 *** (-0.59)	-32.11 *** (-0.49)
fin	14.62	12.91	29.12	27.65	-11.64 *** (-0.21)	-5.05 *** (-0.07)
fra	18.25	12.08	35.06	22.75	-33.82 *** (-0.92)	-35.11 *** (-0.72)
heb	5.42	5.25	12.53	12.08	-3.27 n.s. (-0.04)	-3.59 n.s. (-0.04)
hin	9.06	7.05	20.37	16.35	-22.21 *** (-0.48)	-19.72 *** (-0.43)
hrv	17.73	14.31	34.06	28.27	-19.26 *** (-0.42)	-17.02 *** (-0.25)
hun	19.39	14.08	35.69	27.31	-27.40 *** (-0.61)	-23.47 *** (-0.38)
ind	18.23	11.65	29.92	22.01	-36.10 *** (-1.05)	-26.44 *** (-0.53)
ita	18.09	14.74	34.86	27.02	-18.53 *** (-0.54)	-22.49 *** (-0.48)
jpn	17.02	7.84	39.61	17.47	-53.93 *** (-1.37)	-55.88 *** (-1.17)
kor	7.11	5.91	15.95	13.40	-16.97 *** (-0.23)	-16.00 *** (-0.21)
nld	29.23	16.57	51.22	29.81	-43.31 *** (-1.28)	-41.80 *** (-0.84)
nor	23.73	15.10	45.02	28.03	-36.38 *** (-1.00)	-37.74 *** (-0.70)
pol	16.36	12.17	34.47	23.93	-25.65 *** (-0.58)	-30.59 *** (-0.52)
por	20.71	16.78	44.44	32.16	-18.97 *** (-0.49)	-27.64 *** (-0.54)
ron	16.41	12.51	33.78	23.06	-23.79 *** (-0.61)	-31.72 *** (-0.65)
rus	20.71	10.97	38.37	21.49	-47.06 *** (-1.36)	-44.00 *** (-0.90)
spa	21.53	14.07	47.06	27.52	-34.64 *** (-0.92)	-41.51 *** (-0.83)
swe	25.51	15.56	48.96	29.48	-38.99 *** (-1.00)	-39.79 *** (-0.70)
tel	2.40	2.73	5.55	6.43	**13.86** *** (0.17)	**15.89** *** (0.17)
tur	14.59	9.96	32.85	21.32	-31.73 *** (-0.59)	-35.08 *** (-0.52)
ukr	16.48	10.08	32.20	21.29	-38.84 *** (-0.69)	-33.87 *** (-0.34)
vie	13.68	8.79	25.09	18.23	-35.73 *** (-0.90)	-27.35 *** (-0.54)
zho	15.27	8.03	33.68	17.41	-47.43 *** (-0.98)	-48.29 *** (-0.80)

Table 4: UID values of **Ground-XM3600** across languages under U and P conditions. The Δ columns report the relative change in UID from U to P, computed as $\frac{(\mathrm{P}-\mathrm{U})}{\mathrm{U}}$ and Cohen's d effect size in parentheses. Bold values mark languages where UID is higher under P, indicating lower uniformity in textual information when grounded in perception. Statistical significance is based on paired Wilcoxon signed-rank tests across sentences, with Benjamini–Hochberg FDR correction across languages. Significance thresholds: * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$, n.s. (not significant).

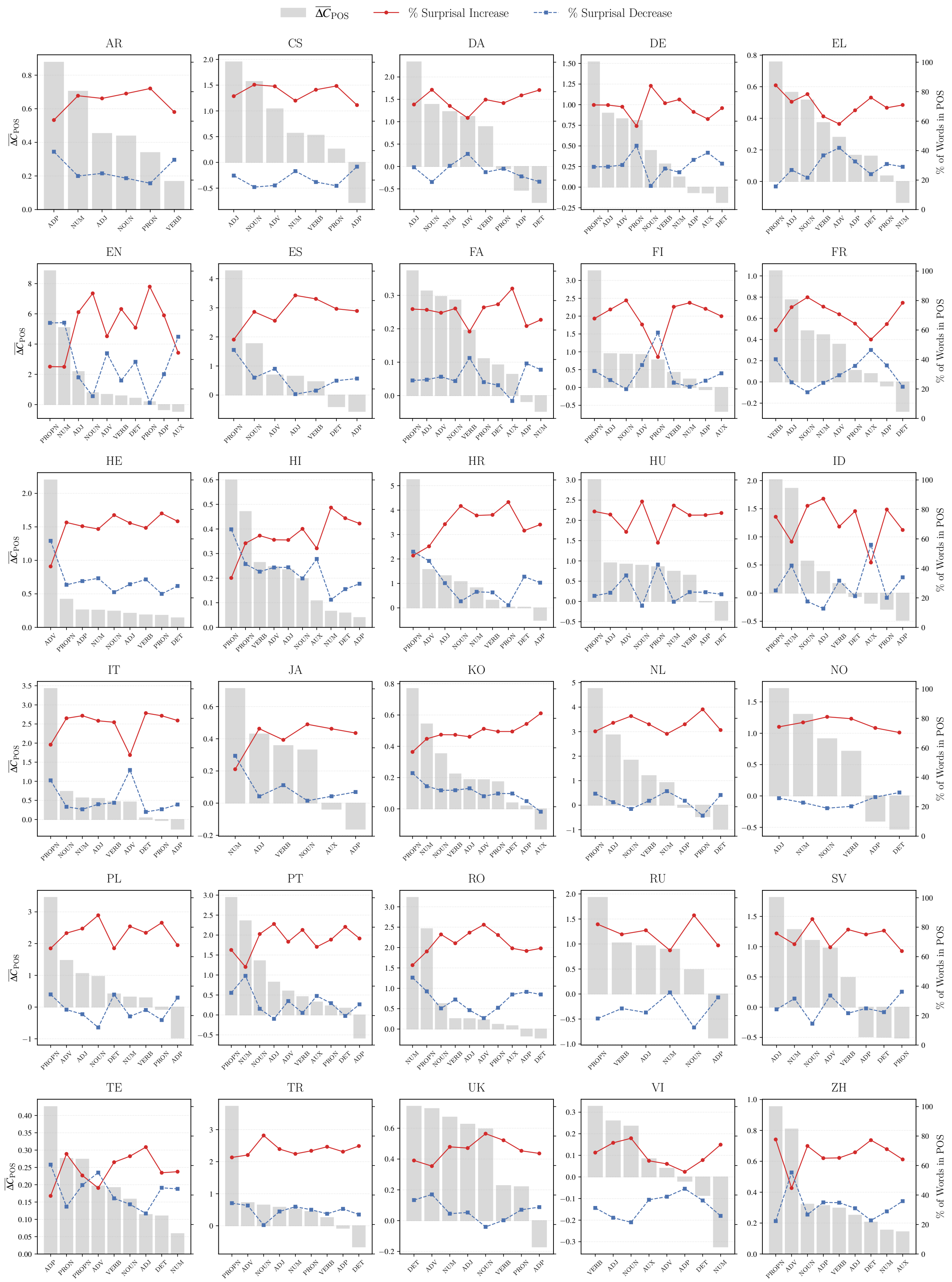

Figure 7: Cross-linguistic dynamics of UID reduction failure across 30 languages in GROUND-XM3600. The bar charts represent the average change in variance contribution ($\overline{\Delta C}_{\text{POS}}$) per Part-of-Speech in sentences where visual grounding increases global variance ($\text{UID}_v(\text{P}) > \text{UID}_v(\text{U})$) (see also Figure 6). Line plots illustrate the percentage of words per POS category exhibiting surprisal increases (red circles) versus decreases (blue squares).

Lang	Global CV			Local CV		
	U	P	Δ%	U	P	Δ%
arb	0.52	0.75	**43.45*****	0.73	1.05	**45.16*****
ces	0.72	0.89	**23.68*****	0.91	1.13	**24.53*****
dan	0.80	0.98	**22.49*****	1.00	1.23	**23.38*****
deu	0.71	0.86	**21.73*****	1.00	1.19	**18.50*****
ell	0.71	0.81	**12.91*****	0.87	1.03	**18.64*****
eng	0.82	1.10	**34.71*****	0.98	1.28	**30.24*****
fas	0.64	0.80	**24.23*****	0.89	1.10	**23.50*****
fin	0.58	0.77	**33.67*****	0.72	1.01	**39.16*****
fra	0.80	1.03	**27.79*****	1.04	1.31	**26.60*****
heb	0.52	0.77	**48.65*****	0.72	1.07	**48.67*****
hin	0.68	0.79	**15.98*****	0.97	1.13	**15.87*****
hrv	0.68	0.89	**30.28*****	0.84	1.12	**32.81*****
hun	0.65	0.83	**27.12*****	0.79	1.03	**30.62*****
ind	0.74	0.89	**20.71*****	0.89	1.15	**29.89*****
ita	0.72	0.99	**36.03*****	0.93	1.24	**33.08*****
jpn	0.90	1.00	**11.46*****	1.29	1.41	**9.50*****
kor	0.63	0.82	**30.89*****	0.87	1.14	**30.61*****
nld	0.91	1.06	**16.91*****	1.08	1.27	**17.86*****
nor	0.85	1.01	**18.13*****	1.06	1.24	**17.38*****
pol	0.64	0.91	**42.78*****	0.84	1.16	**37.43*****
por	0.78	1.04	**32.35*****	1.03	1.29	**25.22*****
ron	0.73	0.91	**24.23*****	0.98	1.15	**17.66*****
rus	0.71	0.92	**30.22*****	0.88	1.19	**34.12*****
spa	0.84	1.03	**22.05*****	1.12	1.30	**16.39*****
swe	0.84	1.02	**22.13*****	1.03	1.25	**21.94*****
tel	0.45	0.65	**42.06*****	0.64	0.90	**42.25*****
tur	0.61	0.76	**26.18*****	0.82	1.01	**23.58*****
ukr	0.66	0.79	**18.69*****	0.84	1.05	**24.02*****
vie	0.75	0.93	**24.81*****	0.95	1.26	**32.73*****
zho	0.77	0.84	**9.61*****	1.05	1.15	**9.45*****

Table 5: Mean Global CV (Coefficient of Variation) and Local CV across languages and conditions for captions in GROUND-XM3600.

Lang	U	P	D	[P + D]
ben	1.05	0.99	0.94	0.93
deu	1.11	1.07	1.08	1.08
eng	1.06	1.01	1.13	1.13
fas	0.87	0.83	0.93	0.91
fra	1.04	1.01	1.07	1.08
ind	0.91	0.88	0.99	1.01
kir	0.83	0.83	0.87	0.88
kor	0.88	0.81	0.80	0.80
por	1.01	0.96	0.99	0.99
rus	0.97	0.95	0.99	1.00
spa	1.00	0.95	1.00	1.01
tha	0.91	0.83	0.86	0.86
zho	0.94	0.88	1.12	1.09

Table 6: Mean Global CV across languages and conditions for paragraph in BLOOM-VIST.

Lang	Unit	U	P	D	[P + D]
ben	paragraph	0.04 n.s	0.01 n.s	-1.37 ***	-1.19 ***
	sentence	-4.13 ***	-3.27 ***	-0.81 ***	-0.56 ***
deu	paragraph	-0.12 n.s	-0.08 n.s	-0.98 ***	-0.71 **
	sentence	-3.25 ***	-2.09 ***	-0.51 ***	-0.31 ***
eng	paragraph	0.24 ***	0.27 n.s	-1.00 ***	-0.67 ***
	sentence	-2.93 ***	-1.92 ***	-0.64 ***	-0.39 ***
fas	paragraph	-0.23 n.s	-0.26 n.s	-3.36 **	-2.84 **
	sentence	-5.83 ***	-3.77 *	-1.13 ***	-0.91 ***
fra	paragraph	0.15 *	0.17 n.s	-1.00 ***	-0.72 ***
	sentence	-3.48 ***	-2.39 ***	-0.62 ***	-0.41 ***
ind	paragraph	0.19 n.s	0.35 n.s	-2.15 ***	-1.61 ***
	sentence	-5.18 ***	-4.05 ***	-1.27 ***	-0.90 ***
kir	paragraph	0.25 n.s	0.28 n.s	-1.76 ***	-1.61 ***
	sentence	-5.11 ***	-4.68 ***	-0.96 ***	-0.76 ***
kor	paragraph	-0.17 n.s	-0.11 n.s	-1.62 ***	-1.14 ***
	sentence	-6.15 ***	-3.77 ***	-0.10 ***	0.13 ***
por	paragraph	0.07 n.s	0.13 n.s	-0.98 ***	-0.63 ***
	sentence	-4.71 ***	-2.72 ***	-0.08 ***	0.13 ***
rus	paragraph	0.08 n.s	0.12 n.s	-1.27 ***	-1.01 ***
	sentence	-3.75 ***	-2.74 ***	-0.79 ***	-0.50 ***
spa	paragraph	0.11 n.s	0.13 n.s	-1.08 ***	-0.75 ***
	sentence	-3.31 ***	-2.06 ***	-0.36 ***	-0.07 ***
tha	paragraph	0.19 *	0.35 n.s	-1.53 ***	-1.07 ***
	sentence	-3.42 ***	-2.54 ***	-0.86 ***	-0.62 ***
zho	paragraph	2.23 **	1.85 n.s	-1.52 ***	-1.10 **

Table 7: Fixed-effect slope estimates over relative position for **mean surprisal**, from mixed-effects models (with random effects by story), computed separately per language and discourse unit (sentence, paragraph). Slopes reflect condition-specific trajectories via interaction terms, with U as the baseline. Chinese was excluded from sentence-level analysis due to insufficient longer sentences for reliable estimation. Significance thresholds: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. (not significant).

Lang	Unit	U	P	D	[P + D]
ben	paragraph	0.74 n.s	-0.46 n.s	-17.31 ***	-11.90 ***
	sentence	-66.38 ***	-44.07 ***	-9.45 ***	-6.01 ***
deu	paragraph	-1.48 n.s	-0.56 n.s	-9.49 ***	-5.66 *
	sentence	-31.17 ***	-16.52 ***	-3.22 ***	-1.27 ***
eng	paragraph	1.97 ***	2.18 n.s	-8.09 ***	-3.99 ***
	sentence	-25.15 ***	-12.32 ***	-4.40 ***	-1.93 ***
fas	paragraph	9.40 n.s	6.48 n.s	-25.22 *	-14.72 n.s
	sentence	-54.23 n.s	-19.38 n.s	2.37 **	-0.79 *
fra	paragraph	-0.20 n.s	0.40 n.s	-7.70 ***	-4.46 ***
	sentence	-28.39 ***	-15.45 ***	-4.18 ***	-2.58 ***
ind	paragraph	-2.69 n.s	-0.95 n.s	-19.95 ***	-12.42 ***
	sentence	-47.16 ***	-32.37 ***	-9.67 ***	-6.76 ***
kir	paragraph	1.97 n.s	1.59 n.s	-19.30 ***	-17.86 ***
	sentence	-55.03 ***	-48.58 ***	-8.16 ***	-6.96 ***
kor	paragraph	-5.33 *	-4.52 n.s	-19.99 ***	-10.69 *
	sentence	-88.33 ***	-39.77 ***	-3.17 ***	-0.98 ***
por	paragraph	-1.17 n.s	-0.48 n.s	-8.18 ***	-4.10 **
	sentence	-46.08 ***	-22.87 ***	-1.31 ***	-0.18 ***
rus	paragraph	-0.74 n.s	-0.39 n.s	-11.81 ***	-8.38 ***
	sentence	-35.96 ***	-23.21 ***	-7.11 ***	-4.54 ***
spa	paragraph	0.72 n.s	0.99 n.s	-9.03 ***	-4.75 ***
	sentence	-31.38 ***	-14.77 ***	-2.81 ***	-0.27 ***
tha	paragraph	-0.72 n.s	0.49 n.s	-15.20 ***	-7.15 ***
	sentence	-38.14 ***	-20.21 ***	-7.07 ***	-4.45 ***
zho	paragraph	15.42 n.s	13.83 n.s	-27.87 **	-16.64 n.s

Table 8: Fixed-effect slope estimates over relative position for **UID**$_v$, from mixed-effects models (with random effects by story), computed separately per language and discourse unit (sentence, paragraph). Slopes reflect condition-specific trajectories via interaction terms, with U as the baseline. Chinese was excluded from sentence-level analysis due to insufficient longer sentences for reliable estimation. Significance thresholds: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. (not significant).

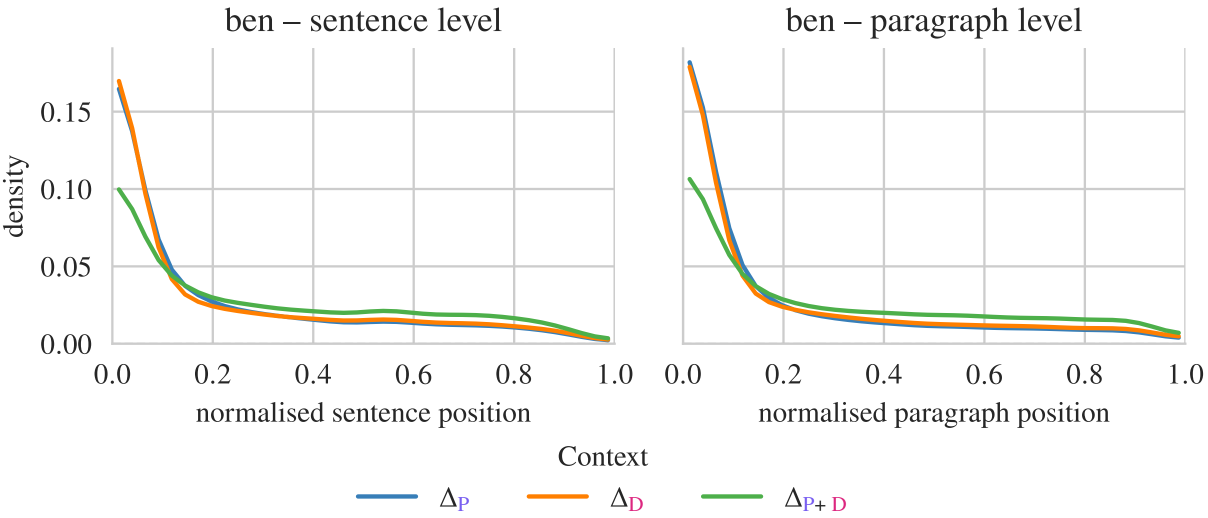

Figure 8: Surprisal reduction density for ben.

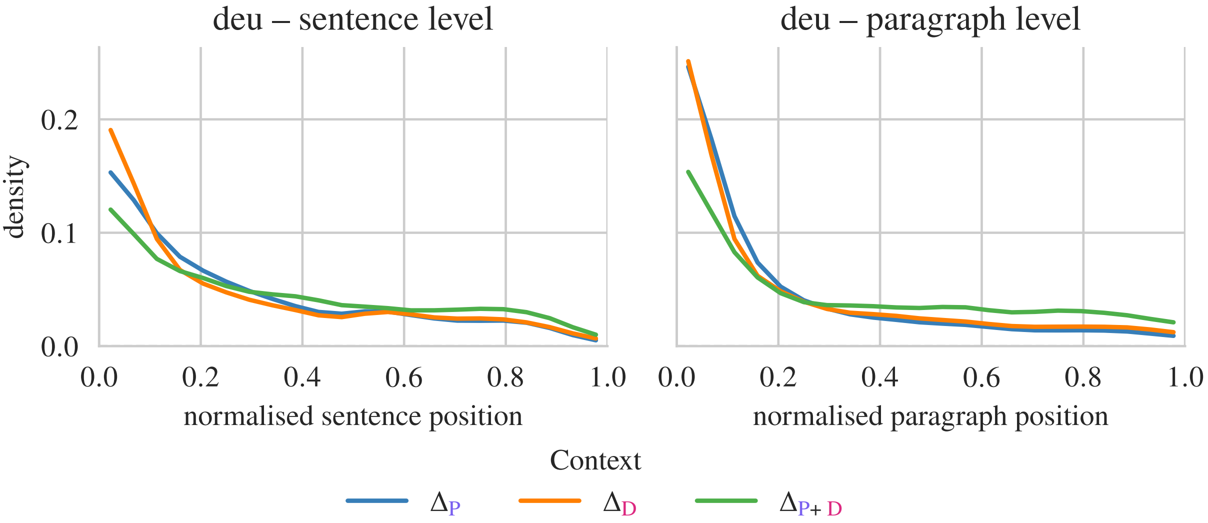

Figure 9: Surprisal reduction density for deu.

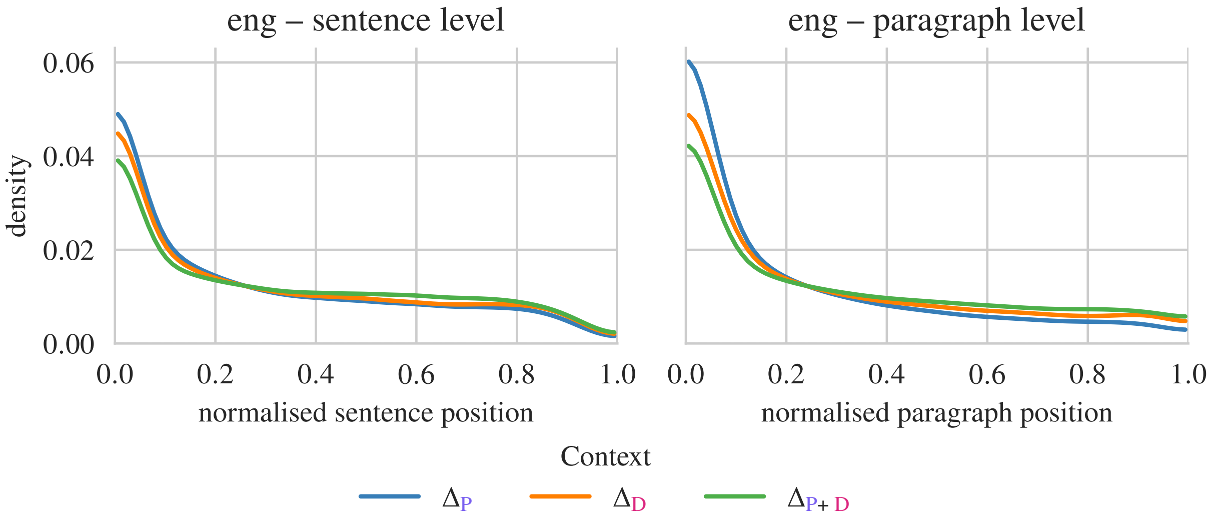

Figure 10: Surprisal reduction density for eng.

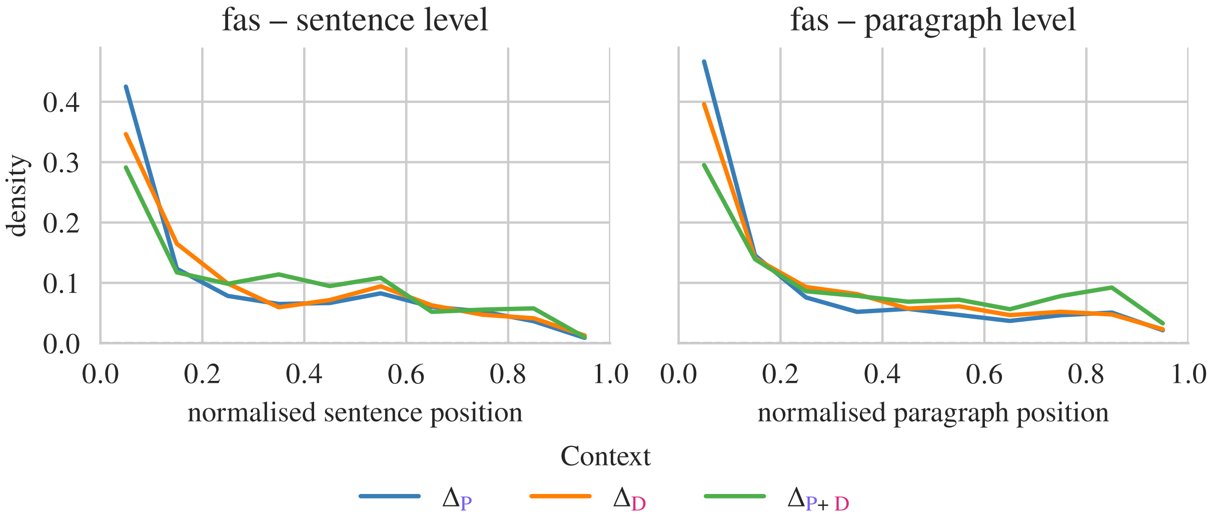

Figure 11: Surprisal reduction density for fas.

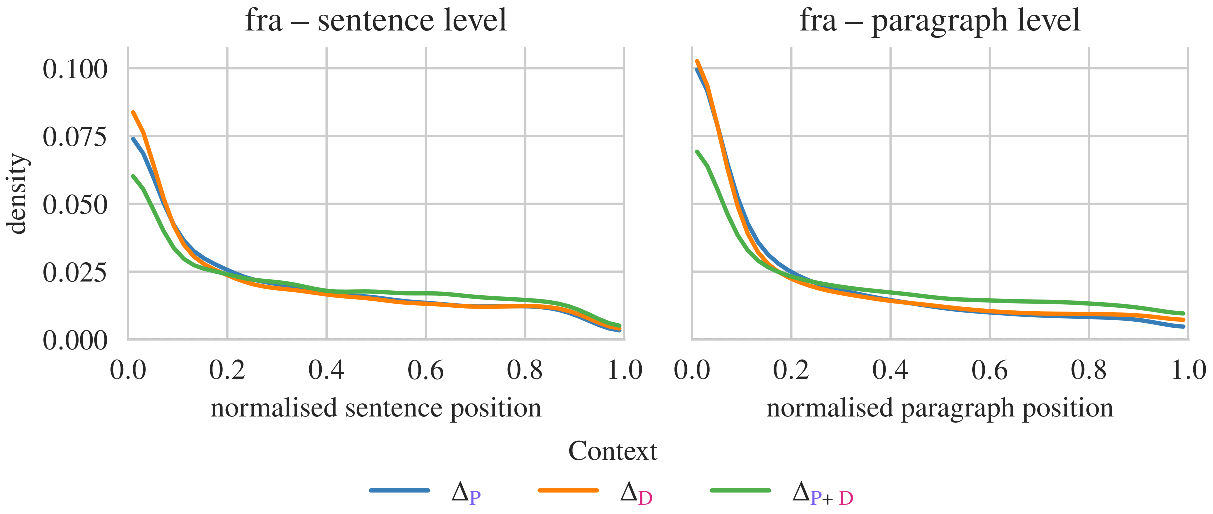

Figure 12: Surprisal reduction density for fra.

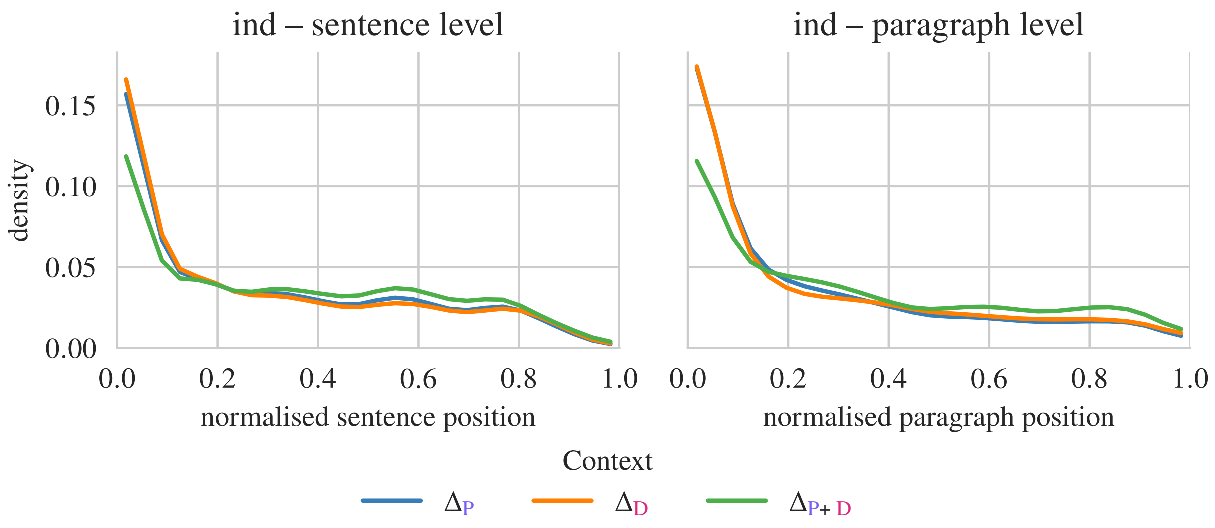

Figure 13: Surprisal reduction density for `ind`.

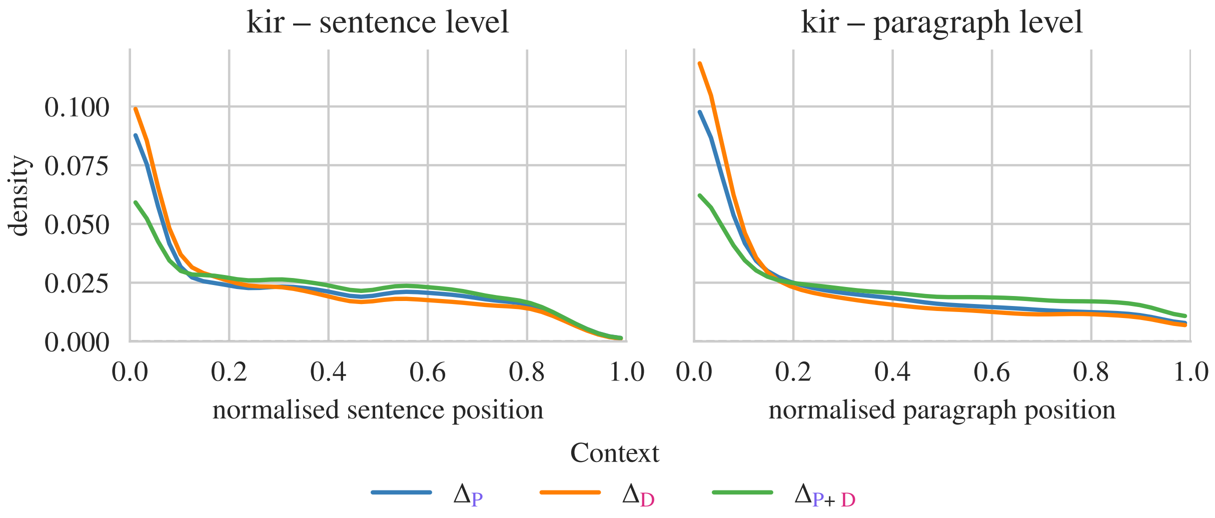

Figure 14: Surprisal reduction density for `kir`.

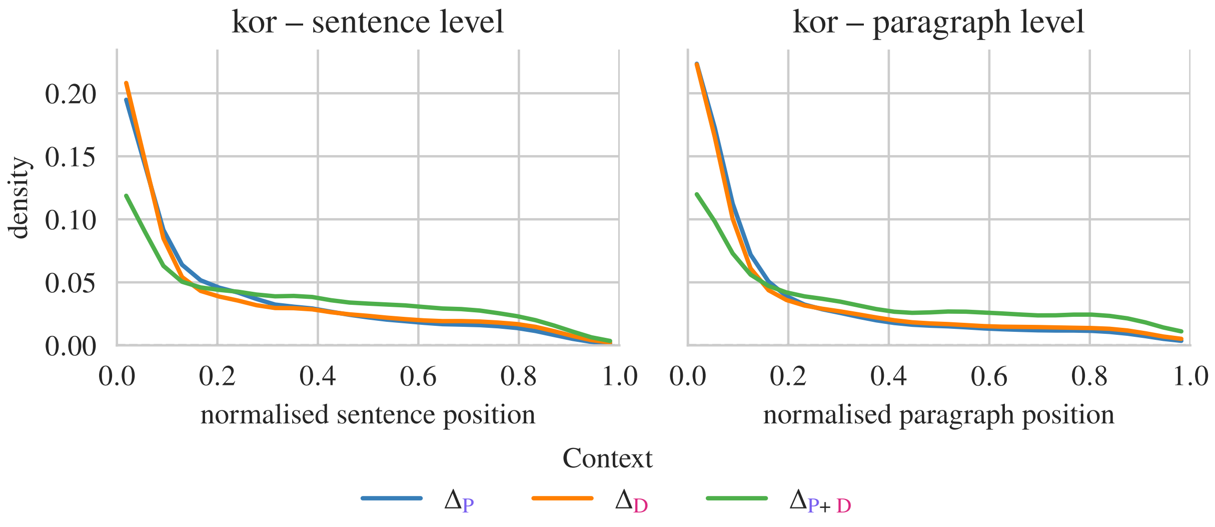

Figure 15: Surprisal reduction density for `kor`.

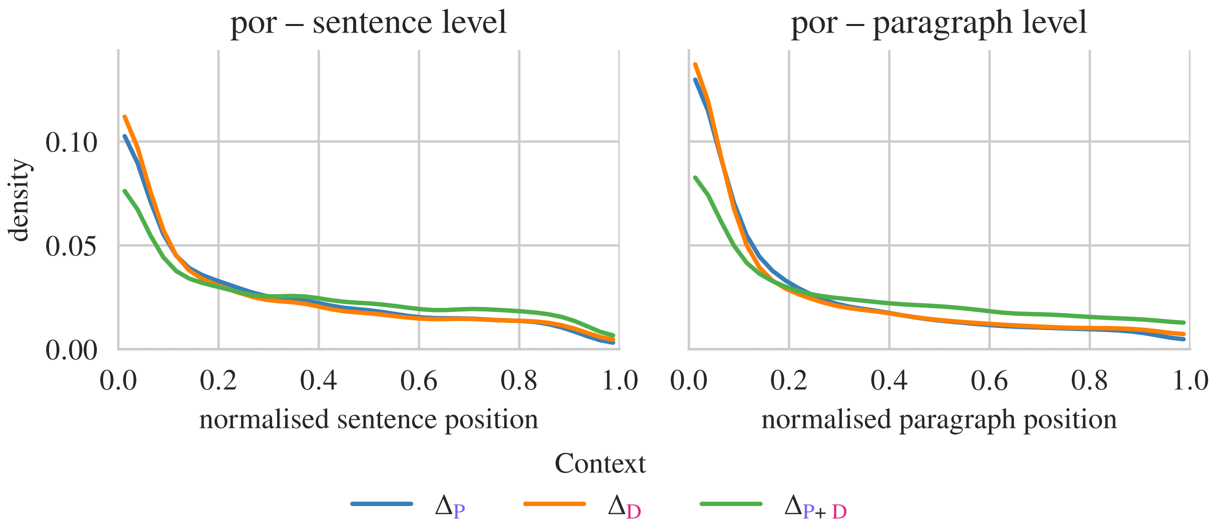

Figure 16: Surprisal reduction density for `por`.

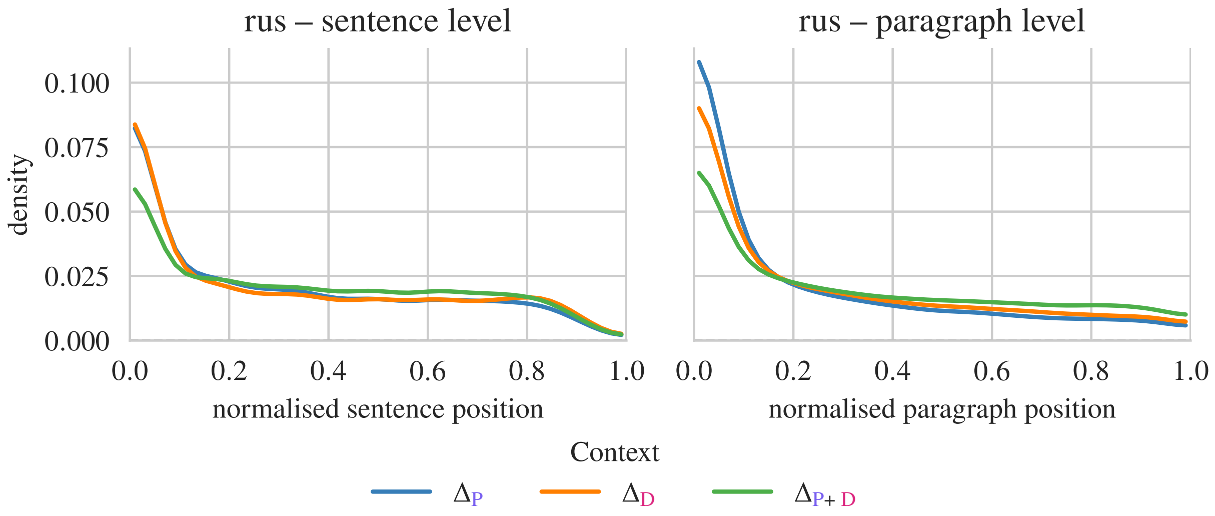

Figure 17: Surprisal reduction density for `rus`.

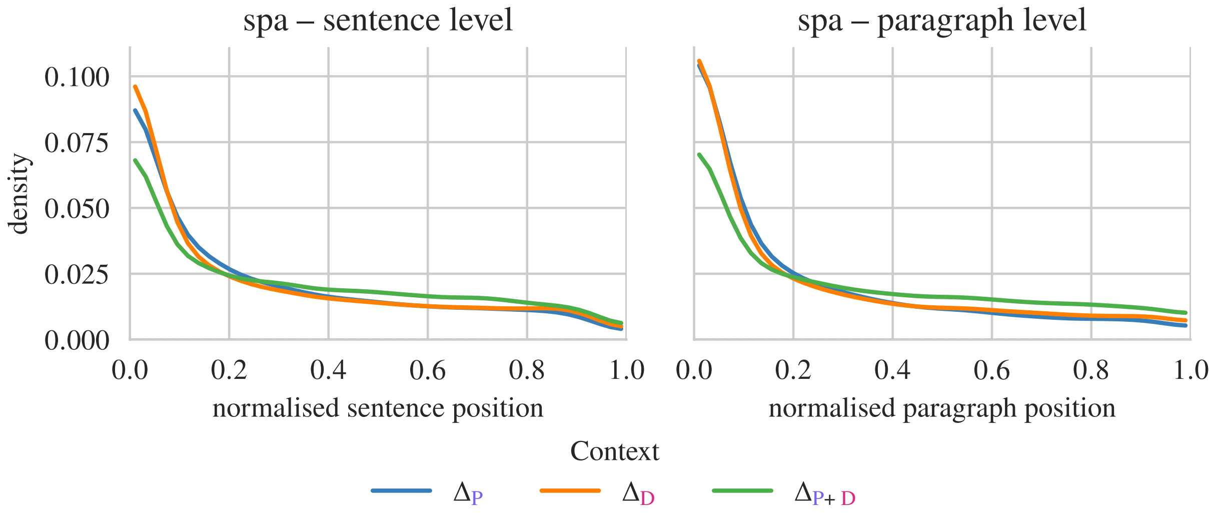

Figure 18: Surprisal reduction density for `spa`.

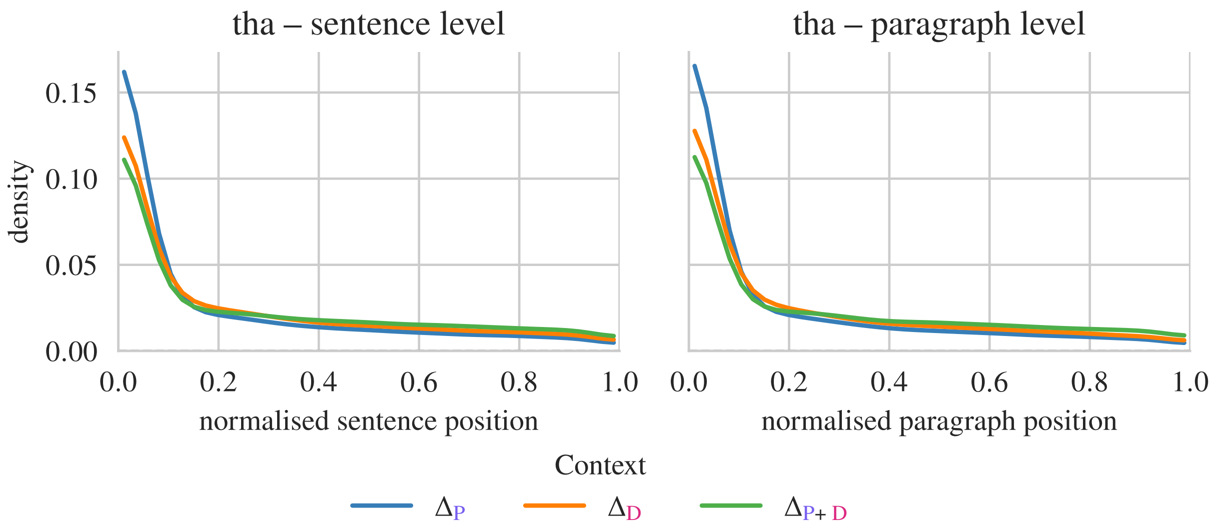

Figure 19: Surprisal reduction density for tha.

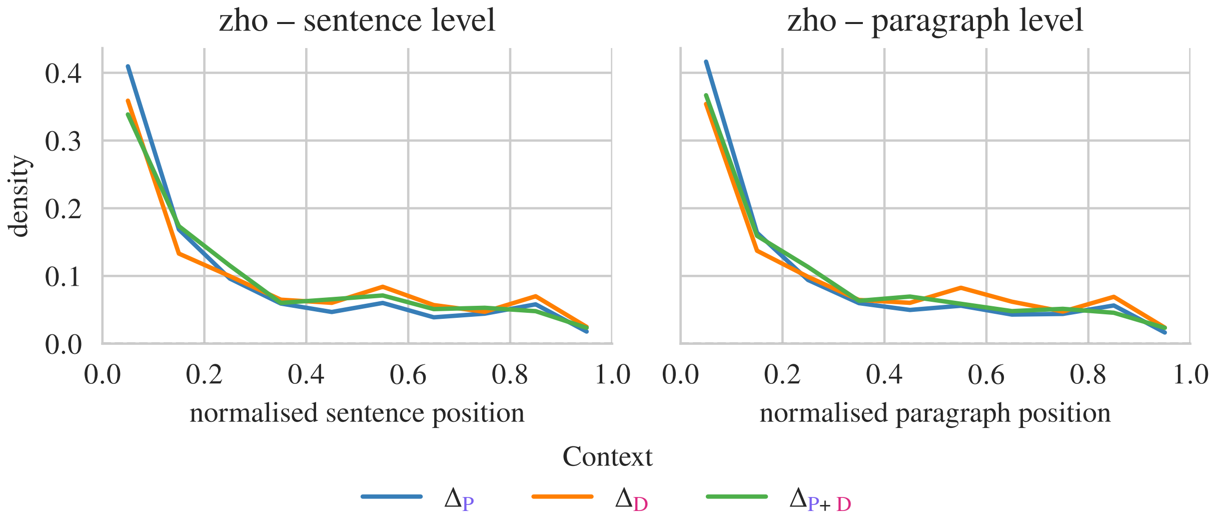

Figure 20: Surprisal reduction density for zho.

Lang	Δ_w	U	P	D	[P + D]
ben	Δ_1	-33.798	-27.867	-7.777	-7.036
	Δ_2	-21.496	-17.803	-5.861	-5.360
	Δ_3	-15.853	-13.089	-4.450	-3.980
deu	Δ_1	-18.408	-14.799	-4.418	-4.002
	Δ_2	-15.674	-12.237	-3.351	-3.141
	Δ_3	-12.439	-9.234	-2.398	-2.246
eng	Δ_1	-13.444	-9.272	-4.644	-3.685
	Δ_2	-10.643	-7.339	-3.264	-2.682
	Δ_3	-8.773	-6.126	-2.540	-2.100
fas	Δ_1	-28.405	-18.189	-6.006	-5.242
	Δ_2	-18.923	-13.385	-3.971	-3.856
	Δ_3	-14.230	-10.333	-3.066	-2.908
fra	Δ_1	-17.818	-13.930	-4.165	-3.513
	Δ_2	-13.930	-10.357	-2.882	-2.447
	Δ_3	-11.239	-8.360	-2.382	-2.051
ind	Δ_1	-21.821	-17.403	-4.656	-3.873
	Δ_2	-15.581	-12.719	-3.133	-2.731
	Δ_3	-11.727	-9.542	-2.233	-1.909
kir	Δ_1	-28.612	-25.267	-7.203	-6.643
	Δ_2	-20.486	-18.859	-4.737	-4.297
	Δ_3	-15.758	-14.631	-3.302	-2.943
kor	Δ_1	-31.875	-23.544	-6.223	-5.794
	Δ_2	-20.749	-14.585	-3.998	-3.802
	Δ_3	-15.638	-10.827	-2.874	-2.706
por	Δ_1	-17.604	-12.163	-3.229	-2.869
	Δ_2	-14.695	-10.185	-2.405	-2.222
	Δ_3	-11.453	-7.802	-1.827	-1.693
rus	Δ_1	-21.067	-15.896	-5.535	-4.513
	Δ_2	-15.306	-11.684	-4.033	-3.392
	Δ_3	-11.230	-8.438	-2.552	-2.086
spa	Δ_1	-17.153	-12.912	-4.245	-3.693
	Δ_2	-12.860	-9.129	-2.745	-2.349
	Δ_3	-10.326	-7.234	-2.128	-1.808
tha	Δ_1	-25.718	-16.269	-4.808	-2.878
	Δ_2	-15.599	-10.085	-2.472	-1.338
	Δ_3	-11.468	-7.542	-1.626	-0.809
zho	Δ_1	-29.135	-20.430	-6.761	-4.975
	Δ_2	-18.282	-12.742	-4.687	-3.503
	Δ_3	-13.019	-9.277	-3.033	-2.332

Table 9: Mean surprisal difference across **sentence** boundaries (Δ_w) for window sizes $w = 1, 2, 3$, reported per language. For each sentence transition, we subtract the average surprisal of the final w words of the preceding sentence from that of the first w words of the following one. More negative values indicate sharper surprisal spikes at sentence onsets, reflecting stronger deviations from uniform information flow across sentence boundaries.

Lang	Δ_w	U	P	D	[P + D]
ben	Δ_1	-33.927	-27.903	-7.909	-7.095
	Δ_2	-21.433	-17.691	-5.919	-5.350
	Δ_3	-15.800	-12.983	-4.470	-3.947
deu	Δ_1	-18.890	-15.328	-4.508	-4.064
	Δ_2	-15.885	-12.417	-3.393	-3.175
	Δ_3	-12.516	-9.246	-2.368	-2.196
eng	Δ_1	-12.443	-8.348	-3.913	-3.196
	Δ_2	-9.996	-6.667	-2.747	-2.319
	Δ_3	-8.264	-5.576	-2.108	-1.784
fas	Δ_1	-27.202	-16.985	-5.372	-4.548
	Δ_2	-17.732	-12.085	-3.205	-3.021
	Δ_3	-13.460	-9.521	-2.597	-2.377
fra	Δ_1	-17.932	-14.049	-4.070	-3.502
	Δ_2	-13.919	-10.280	-2.755	-2.402
	Δ_3	-11.134	-8.186	-2.225	-1.964
ind	Δ_1	-22.326	-17.783	-4.632	-3.841
	Δ_2	-15.722	-12.798	-3.028	-2.622
	Δ_3	-11.914	-9.668	-2.218	-1.885
kir	Δ_1	-28.969	-25.674	-7.307	-6.724
	Δ_2	-20.228	-18.679	-4.742	-4.294
	Δ_3	-15.590	-14.501	-3.328	-2.948
kor	Δ_1	-31.906	-23.686	-6.366	-5.906
	Δ_2	-20.678	-14.559	-4.066	-3.850
	Δ_3	-15.589	-10.820	-2.957	-2.768
por	Δ_1	-17.340	-12.061	-3.323	-2.979
	Δ_2	-14.404	-10.039	-2.436	-2.259
	Δ_3	-11.305	-7.743	-1.894	-1.764
rus	Δ_1	-21.450	-16.293	-5.547	-4.662
	Δ_2	-15.427	-11.786	-3.981	-3.434
	Δ_3	-11.436	-8.606	-2.620	-2.210
spa	Δ_1	-17.171	-12.787	-4.072	-3.586
	Δ_2	-12.954	-9.054	-2.680	-2.324
	Δ_3	-10.404	-7.173	-2.106	-1.820
tha	Δ_1	-25.826	-16.347	-4.758	-2.835
	Δ_2	-15.534	-10.017	-2.423	-1.281
	Δ_3	-11.407	-7.479	-1.616	-0.785
zho	Δ_1	-28.585	-20.354	-6.126	-4.386
	Δ_2	-17.784	-12.421	-4.135	-2.950
	Δ_3	-12.591	-8.992	-2.589	-1.904

Table 10: Mean surprisal difference across **paragraph** boundaries (Δ_w) for window sizes $w = 1, 2, 3$, reported per language. For each transition, we subtract the average surprisal of the final w words of a paragraph from that of the first w words of the following one. More negative values indicate larger spikes in surprisal at paragraph onsets, reflecting greater non-uniformity in information flow across discourse boundaries.