Analysing Human Strategies of Information Transmission as a Function of Discourse Context

Mario Giulianelli and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
{m.giulianelli|raquel.fernandez}@uva.nl

Abstract

Speakers are thought to use rational information transmission strategies for efficient communication (Genzel and Charniak, 2002; Aylett and Turk, 2004; Jaeger and Levy, 2007). Previous work analysing these strategies in sentence production has failed to take into account how the information content of sentences varies as a function of the available discourse context. In this study, we estimate sentence information content *within discourse context*. We find that speakers transmit information at a stable rate—i.e., rationally—in English newspaper articles but that this rate decreases in spoken open domain and written task-oriented dialogues. We also observe that speakers' choices are not oriented towards local uniformity of information, which is another hypothesised rational strategy. We suggest that a more faithful model of communication should explicitly include production costs and goal-oriented rewards.

1 Introduction

Linguistic communication can be understood as information exchange through a noisy channel. Speakers are sensitive to the properties of the channel in two ways (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). On the one hand, they try to reduce the processing effort of the addressee. For example, in the absence of established discourse context, speakers can produce utterances that are easier to process in order to minimise the chance of transmission error. On the other hand, speakers try to reduce their own production effort. For example, given a fixed amount of information that they intend to transmit, speakers can take the risk of producing more concise utterances that are less costly from the production point of view, and expect the addressee to exploit the utterance context for interpretation. Effective and efficient information exchange under these two competing pressures can be modelled using the tools of Information Theory (Shannon, 1948). Indeed, information-theoretic models have successfully accounted for surprisal in speech perception (Jelinek et al., 1975; Clayards et al., 2008), reading (Keller, 2004; Demberg and Keller, 2008; Levy et al., 2009), sentence interpretation (Levy, 2008; Gibson et al., 2013), and overlap in turn taking (Dethlefs et al., 2016).

The information content of a sentence $H(S)$—i.e., its entropy or the effort it takes to process it out of context—and the informativeness of its discourse context $I(S;C)$ are hypothesised to be related. According to the principle of Entropy Rate Constancy (ERC; Genzel and Charniak, 2002), as discourse develops, these two quantities increase at a similar rate; thus, the difference between them—i.e., the effort that it takes to process a sentence *in context*—remains constant over the course of a discourse: $H(S|C) \equiv H(S) - I(S;C)$. A slight relaxation of this prediction is that the information content of a sentence in context remains uniform, rather than constant. This second prediction follows from the principle of Uniform Information Density (UID; Jaeger and Levy, 2007; Jaeger, 2010), according to which speakers make rational linguistic choices that avoid peaks in the density of the information transmitted. Evidence in favour of these principles has been found in texts (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011) and, under certain conditions, in conversations (Vega and Ward, 2009; Doyle and Frank, 2015a,b; Xu and Reitter, 2018; Giulianelli et al., 2021). However, these studies base their conclusions only on estimates of the decontextualised entropy $H(S)$: under the assumption that a larger context is always more informative, an increase in $H(S)$ suffices as an indication that the ERC and UID principles hold.

In this work,[1] we dispose of the assumption that context informativeness increases constantly within a discourse, and we test whether the ERC and UID

[1] Code and statistical analysis are available at https://github.com/dmg-illc/uid-dialogue.

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 647–660
November 10–11, 2021. ©2021 Association for Computational Linguistics

principles hold using, for the first time, direct estimates of the contextualised entropy $H(S|C)$ of an utterance and thus of the informativity of its linguistic context $I(S;C)$. Using a pre-trained Transformer-based language model, which allows us to obtain more accurate probability estimates than the n-gram models used in previous studies and to condition the estimates on discourse context, we replicate Genzel and Charniak's (2002; 2003) seminal experiments on newspaper articles, and in addition apply the analysis to open domain spoken dialogues and to written task-oriented dialogues to test the principles in interactive settings. The proposed operationalisation allows us to test whether the increase in decontextualised entropy observed in earlier work corresponds to an increase in context informativeness or whether speakers simply change their transmission rate over time. Furthermore, this approach allows us to differentiate, for the first time, between the ERC and the UID predictions at the level of discourse.

By studying language production using information-theoretic tools, this paper directly informs the development of computational models of utterance generation. Our findings suggest that architectures and training objectives that enforce a uniform organisation of information density (Meister et al., 2020; Wei et al., 2021) are better suited for reproducing human strategies of information exchange.

2 Related Work

If speakers were to use rational strategies in language production, they would optimise successful communication by transmitting information at a stable rate, close to the capacity of the communication channel (Shannon, 1948). This rational strategy of information exchange (Genzel and Charniak, 2002; Aylett and Turk, 2004; Jaeger and Levy, 2007) is employed at many levels of language production. For example, speakers tend to reduce the duration of more predictable sounds (Aylett and Turk, 2004, 2006; Bell et al., 2003; Demberg et al., 2012), they tend to drop sentential material within more predictable scenarios (Jaeger and Levy, 2007; Jaeger, 2010; Frank and Jaeger, 2008), and they are more likely to overlap at turn transitions that are less information dense (Dethlefs et al., 2016).

At the level of discourse, it has been shown that the decontextualised information content $H(S)$ of sentences increases in written texts with the amount of relevant discourse (Genzel and Charniak, 2002). This has been confirmed across document types and languages (Genzel and Charniak, 2003; Qian and Jaeger, 2011), and it has been hypothesised that an increase in $H(S)$ results in a constant level of information content once the informativeness of discourse context $I(S;C)$ is taken into account. This hypothesis has never been tested directly.

In interactive settings, it is as yet empirically unclear whether speakers use this strategy of information transmission. Vega and Ward (2009) and Xu and Reitter (2018) investigate this in spoken dialogue and show that $H(S)$ grows throughout dialogues, too, and that the contribution in information content of interlocutors shows converging patterns (Xu and Reitter, 2018). In contrast, Doyle and Frank (2015b) fail to find UID effects in Twitter dialogues and multi-party conversations. Again, these studies focus on out-of-context utterance information content and do not measure the contribution of the dialogue context. Doyle and Frank (2015a), in another study with Twitter conversations, are the first to take the informativeness of context into account. They focus on non-linguistic contextual cues (i.e. information about hashtagged events) and show that they have an effect on the overall word-level information transmission profiles of conversations. In this study, we model the effect of *linguistic* context on the information content of sentences in written monologue as well as in written and spoken dialogue.

Recent work has extended this view of human communication to computational models of language generation. Meister et al. (2020) show that the successfulness of many common generation algorithms is related to their tendency to discard lexical choices that make the information content of the words in a sentence less uniform. Wei et al. (2021) build on this finding and augment the training objective of language models with various uniform information density regularisers, thus consistently improving the models' perplexity and generating more lexical diverse texts. This suggests that more faithful modelling of information transmission strategies in humans, to which we contribute in this paper, can inform the development of better computational models of language generation.

3 Data

We analyse trends of information content of written and spoken English, in texts and in dialogue.

Pos.	Sentence	$H(S)$	$H(S\|C)$
1	Stanislav Ovcharenko, who represents the Soviet airline Aeroflot here, has some visions that are wild even by the current standards of perestroika.	5.44	5.44
2	In his office overlooking the runway of Shannon Airport, Mr. Ovcharenko enthusiastically throws out what he calls "just ideas":	6.53	5.61
3	First, he suggests, GPA Group Ltd., the international aircraft leasing company based in Ireland, could lease some of its Boeing jetliners to the Soviet airline.	6.10	5.82

Pos.	Id.	Utterance	$H(S)$	$H(S\|C)$
1	B	Hi. Two women with bagels?	5.61	5.61
2	A	nope	4.18	4.22
3	A	guy with a beard and big pizza	4.95	4.77
4	B	No. A woman and child in dimly lit room	5.24	5.02
5	A	yep she has a green jacket on	5.46	5.21
6	A	a wood table with empty beer bottles on it	4.42	4.55
7	B	Yes.	4.86	4.91
8	A	ok ready	7.13	7.64
9	B	Done	11.85	11.30
10	A	k go	10.32	10.57

Table 1: The first three paragraphs of a Penn Treebank article (document id: 36) and the first round of a PhotoBook dialogue (dialogue id: 2037), annotated with sentence or utterance positions (Pos.), speaker identifier (Id.), and information content estimates.

Excerpts from our corpus of written texts and from our written dialogue corpus are shown in Table 1; further excerpts from all our corpora can be found in Appendix A.

Penn Treebank The Penn Treebank corpus[2] (Mitchell et al., 1999) contains 2,499 English newspaper articles from the Wall Street Journal. We follow the data splits used by Genzel and Charniak (2002, 2003) and divide the corpus into a training set (sections 0–20) and a test set (sections 21–24).

PhotoBook The PhotoBook corpus[3] (Haber et al., 2019) contains 2,500 English task-oriented dialogues between two participants who interact via written chat. The task is set up as game with 5 rounds. In each round, each dialogue participant is shown a set of six images which partially overlap with the set shown to their partner. The goal of the game is to discover which images are common to both participants. The images change in each round, but a subset reappears, which elicits re-descriptions of images that have already been referred to in the dialogue. We split these dialogues into a 70% training set (games 0-1751) and a 30% test set (games 1752-2501).

Spoken BNC The Spoken British National Corpus[4] (Love et al., 2017) contains 1,251 samples of contemporary British English open-domain dialogues, collected in a range of real-life contexts. To be consistent with PhotoBook and previous work (Vega and Ward, 2009; Xu and Reitter, 2018), we select the dialogues that feature only two speakers. We then randomly split these 622 dialogues into a 70% training set and a 30% test set.

4 Method

In this section, we present how we obtain estimates of the information content of sentences (or utterances), both when considered out of context and within their discourse context. We define our main information theoretic measures and describe the computational models that produce the estimates.

4.1 Measuring information content

Our goal is to identify and describe patterns of information transmission throughout a discourse. Following prior work (Genzel and Charniak, 2002, 2003; Doyle and Frank, 2015a,b; Qian and Jaeger, 2011; Xu and Reitter, 2018), we start by taking the sentence (or utterance) S as the basic unit of information transmission,[5] and estimate its information density as the Shannon information content:

$$H(S) = -\log_2 P(S) \qquad [1]$$

In this formulation, the information content of a sentence measures how surprising, or unpredictable, the sentence is if taken out of context.

[2] https://catalog.ldc.upenn.edu/LDC99T42

[3] https://dmg-photobook.github.io

[4] http://www.natcorp.ox.ac.uk

[5] For convenience, throughout the paper, we may use the term ‘sentence’ to refer to both sentences in text and utterances in dialogue.

However, because sentences appear in a discourse, their true information content is always modulated by the informativeness of their context. The availability of contextual cues (e.g., the topic of the text, references to the main entities in the discourse, the writing style) alters the expectations of the listener and, in most cases, it makes new sentences less surprising and less effortful to process.

The contextualised information content of a sentence can also be estimated as the Shannon information content, but using the negative log conditional probability of the sentence given its context C:

$$H(S|C) = -\log_2 P(S|C) \qquad [2]$$

Following the classic information-theoretic model of communication, Genzel and Charniak (2002) put forward the principle of Entropy Rate Constancy (ERC): they hypothesised that the contextualised information content of sentences remains constant throughout a discourse. Drawing from similar ideas, others have hypothesised that speakers make linguistic choices that reduce peaks of comprehension processing effort, leading to the formulation of the principle of Uniform Information Density (UID; Jaeger and Levy, 2007; Jaeger, 2010), which predicts that the contextualised information content of sentences remains uniform throughout a discourse.

Although both principles generate predictions about contextualised information content, previous studies have tried to confirm or disprove them by relying only on estimates of decontextualised information content, due to the lack of suitable computational models. As we have alluded to earlier, to make this simplification, they have relied on the assumption that an increase in the available context always corresponds to an increase in context informativeness (Genzel and Charniak, 2002, 2003): three sentences are more informative to predict the fourth sentence in a text than two sentences are to predict the third sentence. The operationalisation of this assumption requires rewriting the contextualised information content of a sentence as the difference between the decontextualised information content and the mutual information between the next sentence and the context:

$$H(S|C) \equiv H(S) - I(S;C) \qquad [3]$$

Again, as the relevant context is built up, $I(S;C)$ is assumed to increase. So for the ERC and UID principles to hold—i.e., for $H(S|C)$ to remain constant or uniform in Eq. 3—the decontextualised information content $H(S)$ must increase. In prior work, an increase in $H(S)$ was therefore considered sufficient evidence in favour of the principles.

In this paper, we do not assume an increase in $I(S;C)$ and estimate both the decontextualised and the contextualised information content of a sentence. This allows us to directly test the ERC and UID principles and to measure the true informativeness of context as the reduction in sentence surprisal contributed by the context.

4.2 Definitions

The *decontextualised information content* of a sentence is computed by averaging over the negative logarithms of all word probabilities, conditioned only on the preceding words:

$$H(S) = -\frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, ..., w_{i-1}) \qquad [4]$$

The *contextualised information content* of a sentence is computed as the average per-word negative probability, conditioned on the preceding words in the sentence as well as on the entire relevant discourse context:

$$H(S|C) = -\frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, ..., w_{i-1}, C) \qquad [5]$$

Context informativeness is computed as the difference between the previous two quantities:

$$I(S;C) \equiv H(S) - H(S|C) \qquad [6]$$

4.3 Modelling

We compute the log probabilities in Eq. 4 and 5 using GPT-2 (Radford et al., 2019), a pre-trained autoregressive Transformer language model, which allows us to obtain more accurate probability estimates than the n-gram models used in previous work (Genzel and Charniak, 2002, 2003; Doyle and Frank, 2015a,b; Qian and Jaeger, 2011; Xu and Reitter, 2018) and to include discourse context in the computation. We rely on HuggingFace's implementation of GPT-2 with default tokenizers and parameters (Wolf et al., 2020) and to adapt the language model to the idiosyncrasies of different types of language use, we finetune it separately on a 70% split of each target corpus. As shown in Table 2, finetuning yields a substantial reduction in the model's perplexity. More information on model parameters and the finetuning procedure can be found in Appendix B. We use the finetuned language models to estimate decontextualised and contextualised information content (Eq. 4 and 5) of the 30% held-out portion of each corpus.[6]

Fixed context window We use the language model's context window up to its maximum size (1024 tokens). This means that once sentence position in a document is relatively high, the entire window is filled and earlier portions of the context are systematically tossed out. Therefore, the language model cannot exploit long-distance relations involving information present in earlier portions of the discourse that fall outside this window. To ensure that the $H(S|C)$ estimates are not biased for high sentence positions, we determine, for each corpus c, the first sentence position pos^c_{1024} where the sum of context length's average and standard deviation across documents is 1024. Our experiments are then executed on all sentences with position smaller or equal to pos^c_{1024}. The average length of the examined portions of the documents is 15 ± 11, 54 ± 4, and 73 ± 0.5 sentences for Penn Treebank, PhotoBook, and Spoken BNC, respectively.[7].

Control runs Deep learning models are known to exploit peculiarities of the data distribution that humans would not find relevant. In this case, we are concerned that our language model may be able to make use of irrelevant contextual features to produce more accurate $P(S|C)$ predictions. This would lead to an artificial decrease in $H(S|C)$. To control for this eventuality, we obtain $H(S|C)$ estimates for a given sentence using 3 control contexts, following the same procedure described previously for the true context.[8] We randomly sample one control context from the target corpus and two from a corpus with the same modality (i.e., never mixing monologue and dialogue). This ensures that the control contexts are truly independent with respect to the target sentence (e.g., with respect to topic, referents, and style).

	GPT-2 pre-trained	GPT-2 finetuned
Penn Treebank	28.03	21.89
PhotoBook	43.42	14.93
BNC Spoken	66.47	8.69

Table 2: Word-level perplexity of the GPT-2 models on 30% held-out portions of the corpora.

5 Analysis of Language Model Estimates

In this section, we report the estimates and patterns of sentence information content directly computed with the finetuned GPT-2 language models for our three corpora. Recall that we are directly estimating both $H(S)$ and $H(S|C)$ from data, in contrast to previous work, where $H(S|C)$ is never computed empirically. Before using the $H(S|C)$ estimates to test the ERC and UID hypotheses, we validate them by comparison with those obtained using random control contexts (see Section 4.3). If the language model relies on irrelevant contextual features, we would expect the estimates obtained with the true context to be virtually indistinguishable from those obtained with random contexts. This would mean that our $H(S|C)$ estimates are not reliable. In contrast, if the model does effectively exploit the actual context to estimate a sentence's entropy, we should see a clear difference between the true $H(S|C)$ estimate and the control runs.

As can be seen in Figure 1, true and control trends start diverging from sentence position 2. Control contexts produce a positive shift in the magnitude of $H(S|C)$ in all corpora: processing a sentence S in a random context is always harder than processing it in its true context. Moreover, because the control contexts are incoherent with respect to S, they cause $H(S|C)$ to be higher than $H(S)$: processing a sentence S in an incoherent context is harder than processing it with no context.[9] We also notice that while the magnitude of $H(S|C)$ depends on the veracity of the contexts, its fluctuations are largely determined by $H(S)$. This is particularly true for the control trends of $H(S|C)$, whose slope, too, is determined by $H(S)$.

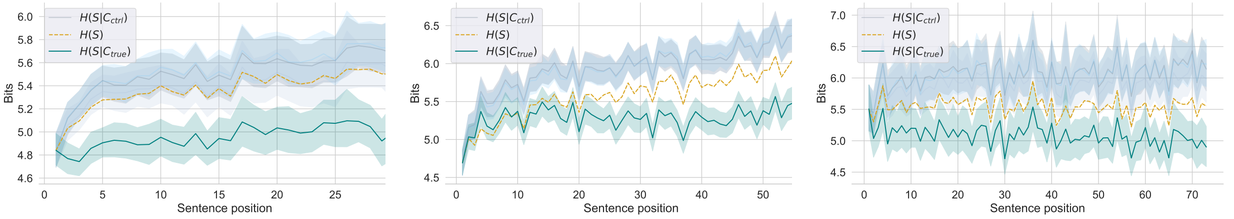
In sum, the $H(S|C)$ trends computed with the

[6] The held-out corpora, annotated with information content estimates, are provided in the supplementary material. Excerpts can be found in Appendix A.

[7] We have tried to substitute GPT-2 with the Transformer-XL language model (Dai et al., 2019) because of its unlimited context window size. In spite of its larger window, however, Transformer-XL yields higher perplexity than GPT-2 on all corpora, hence we sticked to GPT-2. Further reasons to discard Transformer-XL are discussed in Appendix B.1

[8] The length of the control contexts is always equal to the number of tokens in the true context.

[9] In Figure 1a and, partially, in 1c, we can see that one of the control runs is closer to $H(S)$; for this run the contexts are sampled from the target corpus (see Section 4.3) and appear to be less harmful for the language model estimates.



(a) Penn Treebank (b) PhotoBook (c) Spoken BNC

Figure 1: Contextualised information estimates with true and random control contexts. Bootstrapped 95% confidence bands. We also show the mean $H(S)$ values for reference; confidence bands are visible in Figure 2.

true data differ from the trends obtained with control runs according to reasonable expectations: the true $H(S|C)$ estimates are lower, and the control estimates higher, than the $H(S)$ estimates. This provides evidence that our empirical estimates of sentence information content for the three corpora are solid. We can now use these validated estimates to test the ERC and UID hypotheses.

6 Experiments

Taking the values for $H(S)$ (Eq. 4) and $H(S|C)$ (Eq. 5) estimated with the language model, we use Eq. 6 to compute context informativeness $I(S;C)$ for all sentences in our datasets within a fixed initial context window, as explained in Section 4.

Recall that the ERC and UID principles hypothesise that both $H(S)$ and $I(S;C)$ will increase with the position of S within a discourse, and that as a result, $H(S|C)$ will remain stable. This is expressed in Eq. 3, repeated here for convenience:

$$H(S|C) \equiv H(S) - I(S;C) \qquad [7]$$

In the following experiments we investigate whether this is indeed the case.

6.1 Is information content constant?

In Experiment 1, we test whether the positive effect of sentence position on decontextualised information content observed in earlier work (e.g., Genzel and Charniak, 2002, 2003; Xu and Reitter, 2018) corresponds to a comparable increase in context informativeness. Following Qian and Jaeger (2011) and Xu and Reitter (2018), we fit linear mixed-effect models using the logarithm of the decontextualised information content $H(S)$ as our response variable and the logarithm of sentence position as predictor, with a random intercept grouped by distinct documents/dialogues. Because we do not use Xu and Reitter's length-normalised metric, and length is known to have an effect on information content estimates (Keller, 2004), we include the logarithm of sentence length as an additional predictor. Our models also have a document-specific random slope for sentence position and sentence length to capture cross-document variation (Barr et al., 2013). We repeat the same procedure to also fit models using the logarithm of the contextualised information content $H(S|C)$, and the mutual information $I(S;C)$ as response variables.

The results of the linear mixed-effect models are summarised in Table 3; a full report of the results is shown in Table 7 (Appendix C). We now discuss each of the measures in turn.

Decontextualised information content ($H(S)$) Decontextualised information content significantly increases with sentence position in Penn Treebank and in PhotoBook. Its rate of increase is relatively low, as indicated by the coefficients of our linear mixed-effect model. In Spoken BNC, there is no effect of sentence position on $H(S)$.

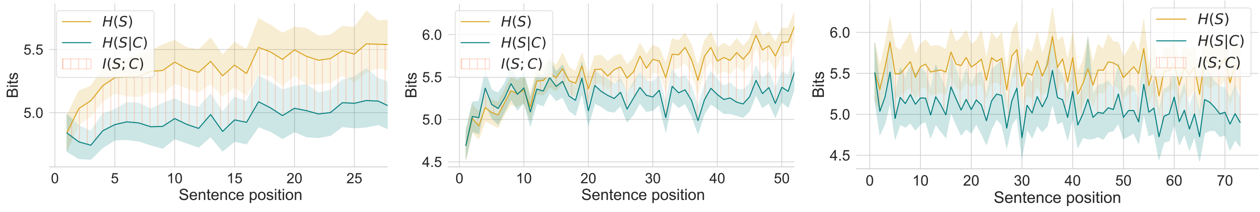
Context informativeness ($I(S;C)$) Context informativeness increases with sentence position in all corpora. Its rate of increase is higher than that of $H(S)$ (recall that these two quantities must increase at a similar rate for $H(S|C)$ to remain constant). In Penn Treebank and Spoken BNC, $I(S;C)$ increases very rapidly in the initial positions; in PhotoBook, the rate of increase is more regular and yields the strongest effect in our statistical models.

Contextualised information content ($H(S|C)$) We find no significant effect of sentence position on contextualised information content in Penn Treebank: $H(S|C)$ remains constant as predicted by the ERC principle. However, we observe a significant negative effect in both dialogue corpora.

Summary The results of Experiment 1 empirically confirm Genzel and Charniak's assump-

	$H(S)$	$H(S\|C)$	$I(S;C)$
Penn Treebank	$\beta = 2.94\text{e}{-2}, p < 0.001$	$\beta = 0.23\text{e}{-2}, p > 0.05$	$\beta = 12.08\text{e}{-2}, p < 0.001$
PhotoBook	$\beta = 4.07\text{e}{-2}, p < 0.001$	$\beta = -1.63\text{e}{-2}, p < 0.001$	$\beta = 27.94\text{e}{-2}, p < 0.001$
Spoken BNC	$\beta = -0.05\text{e}{-2}, p > 0.05$	$\beta = -2.89\text{e}{-2}, p < 0.001$	$\beta = 6.31\text{e}{-2}, p < 0.001$

Table 3: Coefficients of linear mixed-effect models using 1) the logarithm of $H(S)$, 2) the logarithm of $H(S|C)$, and 3) $I(S;C)$ as response variables. The logarithms of sentence position and sentence length are the predictors and they are both assigned a per-document random slope; the models also include a per-document random intercept.



(a) Penn Treebank

(b) PhotoBook

(c) Spoken BNC

Figure 2: Decontextualised information content $H(S)$, contextualised information content $H(S|C)$, and context informativeness $I(S;C)$ against sentence position. Bootstrapped 95% confidence intervals.

tion (2002) that context informativeness increases throughout discourse. $H(S)$ and $I(S;C)$, however, do not always increase together, and when they do, they grow at a different rate. In Penn Treebank, the difference in rate is sufficiently low to keep $H(S|C)$ constant but this is not the case in the dialogue corpora: in PhotoBook $I(S;C)$ increases much faster than $H(S)$, and in Spoken BNC, $H(S)$ does not increase at all. The regression coefficients are rather small but comparable to those found in prior work (Qian and Jaeger, 2011; Xu and Reitter, 2018; Giulianelli et al., 2021). In sum, we find that the ERC principle holds in our corpus of written monologue, but it incorrectly predicts the rate of information in our two dialogue corpora.

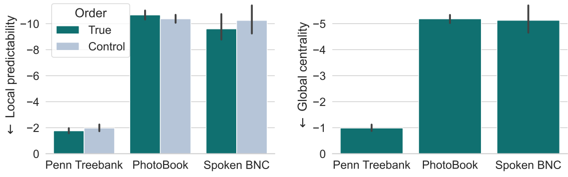
6.2 Is information content uniform?

Experiment 1 suggests that constancy may not be the best descriptor for patterns of contextualised information content, particularly in dialogue. In Experiment 2, we test whether these patterns can be described as uniform. Collins (2014) proposes two criteria to assess uniformity: local predictability and global centrality. *Local predictability* measures whether information content changes in a slow and predictable way from one linguistic unit to the next, as this is expected to reduced the addressee's processing effort and the chances of miscommunication. *Global centrality* measures to what extent the information estimates cluster around a fixed value; this criterion is directly derived from the noisy channel model, predicting that language is transmitted at a stable rate, close to the channel capacity (Shannon, 1948). These measures were originally defined by Collins (2014) to test for uniformity of word-level information content within a sentence; here, we apply them at the sentence level within a discourse. Since they assess uniformity according to different criteria, it is sufficient for one of them to hold to consider information profiles uniform.

We measure global centrality and local predictability of $H(S|C)$ within each document of a corpus. In particular, we calculate *local predictability* as the (negative) mean squared difference in $H(S|C)$ between two consecutive sentences:

$$LP = -\frac{1}{N}\sum_{i=2}^{N}\left(H\left(S_i|C_i\right) - H\left(S_{i-1}|C_{i-1}\right)\right)^2 \quad [8]$$

where N is the number of sentences in a document. We also compute local predictability on 100 randomly shuffled versions of a document, and compare the true and control scores. If local uniformity of information has an influence on speakers' choices, we should find a significant difference between true and control local predictability scores.



(a) Local predictability

(b) Global centrality

Figure 3: Per-document uniformity of contextualised information content $H(S|C)$. Bootstrapped 95% confidence intervals.

Global centrality is the (negative) variance of contextualised information content of all sentences in a document:

$$GC = -\frac{1}{N}\sum_{i=1}^{N}\left(H\left(S_i|C_i\right) - \mu\right)^2 \qquad [9]$$

where μ is the mean information content over the sentences in a document.

Our key results are visualised in Figure 3. We now discuss the two measures of uniformity in turn.

Local predictability We find the highest degree of local predictability in the Penn Treebank articles; $H(S|C)$ estimates for PhotoBook and Spoken BNC show much lower levels of uniformity according to this criterion (see Fig. 3a). Surprisingly, for all three corpora, the local predictability of the true documents is not significantly different from that of shuffled documents: this suggests that, within discourse, the pressure for maintaining the levels of information content locally similar is not as pronounced as it is within a sentence (e.g., Jaeger and Levy, 2007; Collins, 2014).

Global centrality The written texts of the Penn Treebank exhibit a higher degree of global centrality than both written and spoken dialogues (see Fig. 3b). This is in line with our findings for Experiment 1. As reported in Section 6.1, we find no effect of sentence position on $H(S|C)$ in the Penn Treebank, and indeed now we observe that all information estimates in the Penn Treebank documents tend to cluster around a fixed value: in this corpus, information is transmitted at a constant and uniform rate. In the dialogue corpora, where the rate of increase of $H(S)$ and $I(S;C)$ is significantly different, $H(S|C)$ values are less uniformly distributed according to the global centrality criterion.

Summary Experiment 2 shows that sentence information content is significantly more uniform in written monologue than in written and spoken dialogue, both at a local and at a global level. A possible explanation for this may be the fact that, whereas in newspaper articles uniformity depends on the linguistic choices of a single speaker, dialogue utterances are produced online by two speakers, which makes it harder to keep levels of information content locally and globally predictable. Furthermore, comparing the local predictability scores of original and shuffled documents, we find that local predictability is not a good predictor of information transmission patterns in discourse.

7 Discussion and Conclusions

In this study, we have examined some central tenets of the classic information-theoretic model of communication. In contrast to previous work, we have used language models to obtain information content estimates ($H(S|C)$) for sentences *within their discourse context*, and we have measured context informativeness ($I(S;C)$) as the reduction in sentence surprisal contributed by discourse with respect to out-of-context estimates ($H(S)$). This has allowed us to directly model the information transmission profiles of written texts and written and spoken dialogues and, thereby, to test whether they follow the rational communicative strategies predicted by the theory.

We have found that in American English newspaper articles, $H(S|C)$ remains stable as predicted by the theory. This is not the case, however, for spoken British English open domain dialogues, nor for written English task-oriented dialogues: here, $H(S|C)$ decreases, albeit moderately, as sentence position grows. We suggest that this is the result of the uneven rates of increase measured for $H(S)$ and $I(S;C)$—the latter increases faster than the former in all corpora under examination. We find the strongest $I(S;C)$ increase in the PhotoBook dialogues, where topic is determined by a game's image domain (see Section 3) and, by task design, participants produce multiple subsequent sentences to describe the same images over game rounds. Correct interpretation of subsequent references (McDonald, 1978) requires indeed access to the shared knowledge accumulated by speakers during dialogue. We observe the second strongest $I(S;C)$ increase in the Penn Treebank articles, where topic is consistent throughout the text but new information keeps being conveyed from the beginning to the end of the discourse. The weakest increase takes place in the Spoken BNC: topic is more likely to change during the course of an open domain dialogue and, with topic shifts, the previously established common ground becomes less relevant for the prediction of new linguistic material.

The lower rates of increase of $H(S)$, on the other hand, can be due to the limits imposed on lexical choice by grammar and style. In PhotoBook, where participants write freely in a chat interface, the increase is stronger than in the more formal newspaper articles of the Penn Treebank. However, the stable $H(S)$ trends in the Spoken BNC suggest that this is only one side of the coin. The theory predicts that when context is more informative, speakers will increase the density of their sentences to be more efficient, but speakers do not need to be always efficient in open domain conversations, where the pure information transmission goal is perhaps overweighted by more social goals that are not taken into account by the theory.

Another empirical finding that is not in line with the analysed theoretical framework is that uniformity of information content across consecutive sentences (local predictability) is not a good predictor of the information transmission profiles of the texts and dialogues we analysed. Local uniformity may be more relevant for lower-level linguistic signals as they come in a much faster succession: speakers want to avoid sudden changes in information density to reduce comprehension effort; yet, at the discourse level, changes in surprisal are less abrupt as they are spread throughout an entire sentence, thus giving the addressee time to adapt gradually to the higher information content of the larger transmission unit. Global centrality seems to be a more faithful criterion of uniformity, in particular for the articles of the Penn Treebank. In other words, sentences are not so much produced to limit the difference in information content with respect to the previous sentence, but rather to maintain the overall transmission rate stable in the articles. Both dialogue corpora show a significantly lower degree of uniformity than the Penn Treebank, measured both as local predictability and global centrality: in dialogue, an efficient strategy of information exchange needs to be coordinated between two speakers, which can make it more difficult to obtain uniform information profiles.

In conclusion, our study suggests that the classical model of communication may be too simplistic for discourse, where the units of information are more complex. A first issue has to do with identifying the relevant contextual components, which are determined, at least, by the internal structure of the discourse (Genzel and Charniak, 2003) and by topic shifts (Qian and Jaeger, 2011; Xu and Reitter, 2018). We have indeed shown in related work (Giulianelli et al., 2021) that theoretically motivated contextual units exhibit clearer UID trends in task-oriented dialogue.

Second, the predictions made by this model rely on estimates of comprehension effort of a static addressee whereas true addressees adapt on-the-fly: e.g., van Schijndel and Linzen (2018) show that endowing a language model with a simple adaptation mechanism improves predictions of human reading times compared to a non-adaptive model. Moreover, the classic framework assumes a single addressee across documents while, especially for dialogue, communication is shaped by the identity and the characteristics of multiple addressees (Brennan and Clark, 1996; Brown-Schmidt et al., 2015).

Finally, the framework condenses production and comprehension effort in a single estimate. Future work should study strategies of information transmission in discourse using a model of communication, such as the Rational Speech Act model (Frank and Goodman, 2012), that includes production costs more explicitly and that allows accompanying cognitive costs with social costs—e.g., those related to the goal of the linguistic interaction. Zaslavsky et al. (2021) recently showed that the RSA model optimises the trade-off between expected utility and communicative effort, and that it is directly related to Rate-Distortion theory (Shannon, 1948)—the branch of information theory that formalises the effect of limited transmission resources on communicative success.

Acknowledgements

We would like to thank Arabella Sinclair for a valuable discussion of the experimental results, Rochelle Choenni and Ece Takmaz for their useful feedback and suggestions, as well as the anonymous CoNLL-2021 reviewers for their helpful comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

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.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

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.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Sarah Brown-Schmidt, Si On Yoon, and Rachel Anna Ryskin. 2015. People as contexts in conversation. In *Psychology of Learning and Motivation*, volume 62, chapter 3, pages 59–99. Elsevier.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

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

Meghan Clayards, Michael K. Tanenhaus, Richard N. Aslin, and Robert A. Jacobs. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809.

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

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg, Asad Sayeed, Philip Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, Yanchao Yu, Verena Rieser, and Oliver Lemon. 2016. Information density and overlap in spoken dialogue. *Computer speech & language*, 37:82–97.

Gabriel Doyle and Michael Frank. 2015a. 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.

Gabriel Doyle and Michael C. Frank. 2015b. 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.

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*.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

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.

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 (EMNLP)*, pages 65–72.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

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 (EMNLP)*. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

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

T. Florian Jaeger and Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Frederick Jelinek, Lalit Bahl, and Robert Mercer. 1975. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256.

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 (EMNLP)*, pages 317–324.

Roger Levy. 2008. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 234–243.

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. 2009. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC 2014. *International Journal of Corpus Linguistics*, 22(3):319–344.

David D. McDonald. 1978. Subsequent Reference: Syntactic and Rhetorical Constraints. In *Theoretical Issues in Natural Language Processing-2*.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Marcus P. Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 LDC99T42 Web Download. Linguistic Data Consortium.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

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

Marten van Schijndel and Tal Linzen. 2018. A Neural Model of Adaptation in Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Alejandro Vega and Nigel Ward. 2009. Looking for entropy rate constancy in spoken dialog. Technical Report UTEP-CS-09-19, University of Texas El Paso.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A Cognitive Regularizer for Language Modeling. *CoRR*, abs/2105.07144.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Noga Zaslavsky, Jennifer Hu, and Roger P Levy. 2021. A Rate–Distortion view of human pragmatic reasoning. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 347–348.

Appendix

A Corpus Excerpts

Tables 4, 5, and 6 show excerpts of a Penn Treebank article, a PhotoBook dialogue, and a Spoken BNC dialogue. The article (Table 4) is annotated with sentence positions and information content estimates. The dialogues (Tables 5 and 6) are annotated with utterance positions, speaker identifiers, and information content estimates.

B Language Models

We experiment with GPT-2 (Radford et al., 2019), an autoregressive Transformer-based (Vaswani et al., 2017) language model, and we rely on HuggingFace's implementation with default tokenizers and default parameters (Wolf et al., 2020).[10] We

[10] The pre-trained model is named `gpt2` in HuggingFace.

Position	Sentence	$H(S)$	$H(S\|C)$
1	Storage Technology Corp. said it signed a letter of intent to acquire M4 Data Inc. of Britain.	3.89	3.89
2	Terms weren't disclosed.	2.26	2.11
3	Storage Technology said M4's magnetic tape storage equipment will complement its tape cartridge products.	7.64	6.55
4	M4 sells to the original equipment manufacturer market world-wide and has about $20 million in annual sales.	5.75	5.50
5	A Storage Technology spokesman said the transaction should be completed in one to two months.	4.45	3.81

Table 4: An annotated Penn Treebank article (document id: 15).

Position	Speaker	Utterance	$H(S)$	$H(S\|C)$
1	A	Do you have a boy in an orange shirt jumping near a boat ?	3.64	3.64
2	B	Yes.	4.86	5.12
3	A	do you have a miltary boat that shows a man climbing a ladder?	4.25	4.03
4	B	I don't have that one.	1.28	1.47
5	B	I have a woman in a white hat, red boat and blue life vest.	3.62	3.29
6	A	I dont have that	2.69	2.87
7	A	do you have a man in a vest and tie at night against the railing	4.64	4.30
8	B	Yes.	4.86	5.20
9	A	any other questions?	4.05	3.84
10	A	do you see two ladies with a panda bear doll on a boat ?	4.87	4.82
11	B	Yes.	4.86	3.85
12	A	do you see the military man climbing the ladder from the raft in a helmet	4.85	4.42
13	B	Yep. I have that one, too.	2.77	2.32
14	A	do you see a lady in kayak and whit hat red kayak?	4.31	3.97
15	B	I don't have that one this time.	1.51	1.33
16	A	do you have questions?	4.14	4.52
17	B	I have an Asian sitting near several stacks of wood.	6.08	5.63
18	A	no i dont have that	2.78	2.70

Table 5: The first two rounds of a PhotoBook dialogue (dialogue id: 1861).

Position	Speaker	Utterance	$H(S)$	$H(S\|C)$
1	S0018	so how come you're back so early? I thought you had a tennis lesson	3.50	3.50
2	S0019	oh well so did I	5.75	5.74
3	S0019	and having made the arrangement with last Tuesday carefully explaining to him that I couldn't do tomorrow because of the funeral he said well okay I can do twelve o'clock on Monday fine so I toddles along at twelve o'clock today to be told that 's on a course at	3.64	3.79
4	S0018	oh no	5.28	5.22
5	S0019	but had obviously not bothered to write it down	6.08	5.83
6	S0018	so he'd just completely forgotten you?	5.72	5.13
7	S0019	yes in a word	7.36	6.48
8	S0018	Did you phone him?	6.05	6.36
9	S0019	no I didn't I allowed myself a little bit of time to not be quite so cross and I had er half an hour with well more than half an hour three-quarters of an hour with one of the other coaches there	3.95	3.71
10	S0018	what he just happened to be free?	5.71	6.06

Table 6: The first ten turns of a Spoken BNC dialogue (dialogue id: SVNL).

use the model's maximum sequence length, 1024. As the pre-trained model yields relatively high perplexity on the target corpora, we finetune[11] it on 70% of each target corpus and leave out 30% of the dataset to compute the model's evaluation perplexity and to conduct our statistical analysis. The training and held-out portions of the corpora are specified in the main paper. GPT-2 is finetuned for 20 epochs with a learning rate of $1e-04$ and batches of size 8. Because 20 epochs do not yield a substantial perplexity reduction for the Spoken BNC dialogues, we finetuned the model for 20 additional epochs. The perplexity of the pre-trained and finetuned models on the target corpora is reported in the main paper.

For our estimates of information content, we include sentence beginning symbols as contextual cues but their information content is not computed.

B.1 Transformer-XL

Although excluding high sentence positions is in line with prior work measuring decontextualised information content (e.g., Genzel and Charniak, 2002, 2003; Xu and Reitter, 2018), we have tried to substitute GPT-2 with the Transformer-XL language model (Dai et al., 2019) because of its unlimited context window size. In spite of its larger window, however, Transformer-XL yields higher perplexity than GPT-2 on all corpora. Moreover, to make finetuning computationally feasible, we had to limit the context window size to values close to 1024; this is likely to make the model unable to use very long-distance dependencies at inference time, making it more similar but less performant than GPT-2. Indeed, Transformer-XL models finetuned with a fixed context size of 1024 yield higher perplexity than the corresponding finetuned GPT-2 models.

C Experimental Results

Table 7 summarises the results of our statistical analysis, as introduced in Section 6.1. Our linear mixed-effect models include the logarithm of the information theoretic estimate of interest (con-

[11] We use HuggingFace's finetuning script https://github.com/huggingface/transformers/blob/master/examples/pytorch/language-modeling/run_clm.py.

		Fixed effects			Random effects (Std. Dev.)	
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
PTB: $H(S)$	Intercept	1.966	0.025	<0.001	0.332	0.186
	Position	0.029	0.004	<0.001	0.043	
	Length	-0.125	0.006	<0.001	0.076	
PTB: $H(S\|C)$	Intercept	1.878	0.026	<0.001	0.320	0.204
	Position	0.002	0.004	0.545	0.037	
	Length	-0.107	0.007	<0.001	0.076	
PTB: $I(S;C)$	Intercept	0.711	0.048	<0.001	0.587	0.397
	Position	0.121	0.007	<0.001	0.058	
	Length	-0.173	0.013	<0.001	0.164	
PB: $H(S)$	Intercept	1.786	0.010	<0.001	0.183	0.337
	Position	0.041	0.002	<0.001	0.042	
	Length	-0.181	0.003	<0.001	0.056	
PB: $H(S\|C)$	Intercept	1.986	0.010	<0.001	0.190	0.397
	Position	-0.016	0.003	<0.001	0.039	
	Length	-0.250	0.003	<0.001	0.065	
PB: $I(S;C)$	Intercept	-1.089	0.027	<0.001	0.559	0.846
	Position	0.279	0.007	<0.001	0.134	
	Length	0.355	0.009	<0.001	0.199	
BNC: $H(S)$	Intercept	1.813	0.015	<0.001	0.144	0.287
	Position	-0.001	0.003	0.875	0.027	
	Length	-0.080	0.004	<0.001	0.038	
BNC: $H(S\|C)$	Intercept	1.729	0.025	<0.001	0.241	0.492
	Position	-0.029	0.006	<0.001	0.060	
	Length	-0.051	0.006	<0.001	0.065	
BNC: $I(S;C)$	Intercept	0.446	0.049	<0.001	0.351	1.154
	Position	0.063	0.012	<0.001	0.075	
	Length	-0.104	0.011	<0.001	0.087	

Table 7: Results of linear mixed-effect models on the Penn Treebank articles (PTB), the PhotoBook written dialogues (PB), and the Spoken BNC dialogues (BNC).

textualised information content $H(S)$, decontextualised information content $H(S|C)$, or context informativeness $I(S;C)$) as the response variable; the logarithm of sentence position and the logarithm of sentence length as predictors; a random intercept grouped by distinct documents/dialogues; and a document-specific random slope for sentence position and sentence length. The Random effects columns show the standard deviation of the random effects (Coeff.) and the residual standard deviation.

D Computing Infrastructure

The models were trained and evaluated on a computer cluster with Debian Linux OS. Parallelization over four GPUs was implemented for the finetuning of GPT-2. All information content computations were executed using used a single GPU. The GPU nodes are GPU GeForce 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1.