CORTEX 181 (2024) 12–25

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

The unique contribution of uncertainty reduction during naturalistic language comprehension

Ming Song [a,b,c], *Jing Wang* [a,b,c,**] *and Qing Cai* [a,b,c,*]

[a] *Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), Affiliated Mental Health Center (ECNU), Institute of Brain and Education Innovation, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China*

[b] *Shanghai Changning Mental Health Center, Shanghai, China*

[c] *Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China*

ARTICLE INFO

Article history:
Received 22 March 2024
Reviewed: 21 June 2024
Revised 21 July 2024
Accepted 24 September 2024
Action Editor Valentina Borghesani
Published online 5 October 2024

Keywords:
Prediction
Surprisal
Entropy reduction
Language comprehension
Naturalistic stimuli
fMRI

ABSTRACT

Language comprehension is an incremental process with prediction. Delineating various mental states during such a process is critical to understanding the relationship between human cognition and the properties of language. Entropy reduction, which indicates the dynamic decrease of uncertainty as language input unfolds, has been recognized as effective in predicting neural responses during comprehension. According to the *entropy reduction hypothesis* (Hale, 2006), entropy reduction is related to the processing difficulty of a word, the effect of which may overlap with other well-documented information-theoretical metrics such as surprisal or next-word entropy. However, the processing difficulty was often confused with the information conveyed by a word, especially lacking neural differentiation. We propose that entropy reduction represents the cognitive neural process of information gain that can be dissociated from processing difficulty. This study characterized various information-theoretical metrics using GPT-2 and identified the unique effects of entropy reduction in predicting fMRI time series acquired during language comprehension. In addition to the effects of surprisal and entropy, entropy reduction was associated with activations in the left inferior frontal gyrus, bilateral ventromedial prefrontal cortex, insula, thalamus, basal ganglia, and middle cingulate cortex. The reduction of uncertainty, rather than its fluctuation, proved to be an effective factor in modeling neural responses. The neural substrates underlying the reduction in uncertainty might imply the brain's desire for information regardless of processing difficulty.

* *Corresponding author*. School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai, 200062, China.

** *Corresponding author*. School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai, 200062, China.

E-mail addresses: wangjing@psy.ecnu.edu.cn (J. Wang), qcai@psy.ecnu.edu.cn (Q. Cai).

https://doi.org/10.1016/j.cortex.2024.09.007

1. Introduction

When humans comprehend language with uncertainty about the upcoming utterance, they not only decode the semantic content but also constantly engage in numerous complex and automated cognitive processes, particularly real-time prediction. Such prediction happens at multiple levels of language, one of which is the level of word (Caucheteux et al., 2023; Heilbron et al., 2022; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018). To characterize the prediction-related mental processes and link the stimuli with neurocognitive activities, deep language models are used to infer the metrics based on the probability distribution of words in the vocabulary. These metrics have been found able to find shared computational principles for language processing in the human mind and large language models (Goldstein et al., 2022), including that they both are engaged in continuous next-word prediction and that they compare their pre-onset predictions to the incoming word to calculate the post-onset surprise. Information theory-derived metrics have been developed and linked to the mental and neural processes of language comprehension, but these metrics were mostly interpreted as indicators of complexity and processing difficulty in sentence comprehension (Hale, 2016; Wehbe et al., 2021). These metrics have also been linked to the 'information' gained from processing the sequence (Hale, 2006), but the relationship between various metrics and their specific implications has not been clarified.

Surprisal and entropy are the two most studied neural-linguistic metrics associated with prediction uncertainty. The surprisal of a word is determined by the deviation to its probability to occur in the context (Hale, 2001; Levy, 2008). Entropy (Shannon, 1948, 1951) of an upcoming word indicates the uncertainty, or the confidence of predicting the next word before it occurs, regardless of what the actual word would be. Surprisal can index prediction errors which leads to increasing processing effort (Demberg & Keller, 2008; Smith & Levy, 2013). Given "The day was breezy so the boy went outside to fly …", the continuation "an airplane" would elicit higher surprisal than "a kite" (DeLong et al., 2005). Such high surprisal has been linked to some neurobiological signatures, including N400 (Fitz & Chang, 2019; Hodapp & Rabovsky, 2021). By modeling blood oxygen level-dependent (BOLD) signals during comprehension of spoken language stimuli, previous work has also identified brain regions sensitive to word-level surprisal, mainly located in bilateral temporal gyri (Russo et al., 2020; Shain et al., 2020; Willems et al., 2016). Different from surprisal that reflects a backward-looking cognitive process (that is, the deviation to existing prediction), entropy is a forward-looking metric that quantifies the confidence of newly generated prediction, providing evidence that the next-word entropy and surprisal have distinctive effects in modeling BOLD signals during comprehension of spoken language stimuli (Willems et al., 2016).

Both surprisal and entropy depict important aspects of the mental states during language processing, but they characterize neither the state of comprehension *per se*, nor the outcome of successful comprehension. Correspondingly, entropy reduction (ER) which has been emphasized in recent years can quantify the levels of understanding by measuring the change of prediction uncertainty (Venhuizen et al., 2019) in continuous sentence processing. The unclosed sentence "In a breezy day, the boys went outside to …" becomes more predictable when the verb "fly" appears, resulting in a reduction of uncertainty about the rest of the sentence. The entropy reduction after seeing the word "fly" is higher than seeing the word "look", because "look" does not tend to increase one's confidence in predicting the following sequence since the word "to" is presented, whereas "fly" leads to more certain prediction of the following sequence. With the intuition that entropy reduction is a kind of "information gain", the entropy reduction hypothesis (ERH) suggests that entropy reduction is positively related to human sentence processing difficulty, as processing difficulty at a word is proportional to the information gain brought by it (Hale, 2006). However, the relatedness between processing difficulty and ER does not imply that ER merely reflects processing difficulty, nor does it imply that ER represents the same mental process as surprisal or entropy does. ER has been found to affect reading time independently from that of surprisal (Frank, 2013; Linzen & Jaeger, 2016; Lowder et al., 2018). Neural correlates of entropy reduction have been found in the temporal lobe via intracranial electrodes (Nelson, Dehaene, et al., 2017). An ERP study found that ER appeared to quantify different neural processes with compared to surprisal (Frank et al., 2015), but it failed to identify a corresponding ERP component. Hale et al. (2018) emphasized the ER calculated with a syntactic, rather than non-syntactic, neural language model yielded electrophysiological responses on anterior frontal electrodes. Although these electrode-based studies rarely involve areas beyond the cortical surface, they suggested a unique role of ER.

However, the operational definitions of information-theoretical metrics, particularly in terms of entropy reduction, have varied across studies. Most studies agreed that surprisal can be quantified as a word-based measure, but researchers made efforts to expand the concept of entropy beyond vocabulary, specifically to include uncertainty about the rest of the sentence (Hale, 2006). These uncertainty values are usually defined using probabilistic grammars[1] (Hale, 2006; Yun et al., 2015). Entropy reduction calculated by grammar-based approaches was found to be correlated with neurophysiological activities (Hale, 2006; Nelson, Dehaene, et al., 2017), and other studies have confirmed the effect of entropy reduction in reading time (Wu et al., 2010). Although some studies posit that a word's syntactic category might be undetermined on the first encounter (Frank, 2010, 2013), the issue could be solved by multipath parsing (Brennan et al., 2020; Franzluebbers et al., 2024; Hale et al., 2018). Word-based approaches proposed in (Frank, 2010, 2013) remain noteworthy for their emphasis on semantic content rather than sentence structure, and they avoid the difficulties associated with writing grammars, allowing more researchers to understand

[1] Probabilistic grammars are introduced pedagogically in Appendix C of Jurafsky and Martin's Speech and Language Processing 3rd Edition (Jurafsky & Martin, 2019) and Chapter 10 of Eisenstein's Introduction to Natural Language Processing (Eisenstein, 2019).

and expand its implications. Frank and colleagues measured simplified entropy reduction based on several future actual words and found effects on reading time (Frank, 2013), but they failed to identify a corresponding neural component (Frank et al., 2015). Lowder et al. (2018) used a cumulative cloze task of the next word to measure both surprisal and entropy reduction, suggesting that they capture different aspects of lexical predictability. A computational linguistic model defined in a limited semantic space also suggested that surprisal and ER reflect different aspects of incremental comprehension (Venhuizen et al., 2019), but it is difficult to expand the semantic space to cover a broader language. Overall, these methods had varying degrees of compromise in quantifying the uncertainty about the sentence.

We propose that the entropy reduction in the sequential language input results in a change in cognitive state that is beyond processing difficulty. In this study, we quantify simplified ER using actual word probabilities, inspired by the work of Frank and colleagues (Frank, 2013; Frank et al., 2015). There are two reasons for doing this: First, uncertainty about the whole sentence is reflected to some extent in entropy regarding future words, i.e., reflected by the confidence in what the actual following words are. Second, as surprisal was calculated based on the next-word prediction, it would benefit the comparison of metrics to quantify the measures in a unified form, such as using generative GPT models. This would help to clarify the relationship among different information measures at the same level. In this way, ER targets the fluctuation of uncertainty about the following actual words rather than the grammar roles, similar to how surprisal and entropy were used in previous studies (Russo et al., 2020; Willems et al., 2016). The fluctuation of uncertainty is a commonly studied cognitive process in various non-language tasks. It had independent incentive value in children learning (Feldstein, 1973; Nicki & Shea, 1971; Wentworth & Witryol, 1984) and played a role in promoting domain-general reward pursuit (Asutay et al., 2020; Daikoku, 2019; Gold et al., 2019; Kringelbach, 2005; Shen et al., 2015). Therefore, we expect that the effects of ER can be distinguished from that of surprisal, especially in brain areas that are not language-specific.

One brain area of particular interest is the ventromedial prefrontal cortex (vmPFC), which cannot be investigated by electrophysiological studies. vmPFC was involved in processing various forms of uncertainty (Bechara et al., 2000; Critchley et al., 2001; Hsu et al., 2005; Huettel et al., 2006; I. Levy et al., 2010), semantic composition (Bemis & Pylkkänen, 2011; Pylkkänen, 2019a; Pylkkänen et al., 2009) and reward processing (Ciaramelli et al., 2021; Grabenhorst & Rolls, 2011; Pujara et al., 2016; Rolls, 2022; Strait et al., 2014). Even when the task did not explicitly require predictions, vmPFC spontaneously tracked the prediction about the choice of the subject's partner in a referential communication game (Mi et al., 2021). vmPFC has also been found to reflect the stable trait of uncertainty aversion by how it expands the representational space of different concepts (Vives et al., 2023). We expect vmPFC to also be responsible for tracking the change of uncertainty, or in other words, the information gain, in naturalist speech processing, by differentiating the effect of entropy reduction from other difficulty-related measures.

The goal of this study is to identify the uniqueness of gaining information in continuous language processing. We hypothesize that the amount of information gained from the incoming word (characterized by entropy reduction) during naturalistic language comprehension predicts a unique proportion of neural responses that cannot be accounted for by processing difficulty (indexed by entropy and surprisal). We derived these metrics using the pre-trained GPT-2 model (Radford et al., 2019) and subsequently employed them to predict fMRI signals acquired during the listening of corresponding narrative stories (Bhattasali et al., 2020). We performed a region-of-interest analysis to investigate the roles of these metrics in accounting for vmPFC activations. Additionally, we identified other brain areas that were sensitive to the prediction difficulty during language processing based on the effects of entropy and surprisal. Crucially, we differentiated the effects of entropy reduction in predicting brain activities from that of the other two metrics to unbind information gain from processing difficulty in language comprehension. We expect that ER accounts for neural responses during continuous speech comprehension that are not characterized by the metrics of processing difficulty, particularly in vmPFC.

We used the Alice Dataset (Bhattasali et al., 2020), an open neural dataset collected when participants listened to the first chapter of *Alice in Wonderland*, to address our questions. This dataset has been widely used to investigate correlations between linguistic structure and brain activities. The electrophysiological data has been used to compare different computational models by investigating amplitude effects (Brennan, 2016; Brennan & Hale, 2019; Hale et al., 2018). The fMRI data has been applied to investigate the correlations between processing-complexity predictors and neural activities in temporal and frontal regions (Brennan et al., 2016, 2020; Hale et al., 2015; Li et al., 2016), but has discussed little about entropy reduction. There is difficulty in integrating all these results while focusing on different information measures rather than different computational models. Considering that language models have been proven to provide efficient modeling for next-word predictions (Goldstein et al., 2022), we calculated all the information measures based on the publicly available language model and re-analyzed the fMRI data in this study, which facilitates replication and comparison.

2. Materials and methods

2.1. *The Alice Dataset*

This study used the text stimuli and reanalyzed the fMRI data from the Alice Dataset (Bhattasali et al., 2020). Human subjects listened to the first chapter of *Alice in Wonderland* while being scanned. The audio stimulus presented to each participant lasted 12.4 min. The corresponding text transcription was segmented into 2132 words. Linguistic annotations of the words were provided, including the timestamps (onsets and offsets) on the auditory stimulus, the log-transformed lexical frequency (LgWF), and the prosodic break strength (PBS). The prosodic break strength is a perceptual judgment of break index strength (Beckman et al., 2010). The magnetic resonance

imaging as well as the electro-physiology data were provided in the dataset. We only use the fMRI data that were collected using a 3 T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil. We adhered to all the data exclusion criteria and followed the preprocessing procedures implemented in the original work (Bhattasali et al., 2020), and put the processed data into MNI space with a resolution of 2 mm. After excluding the data of subjects with excessive head movement and low behavioral performance in post-scan comprehension tests, data from 23 native English speakers were included in this study.

2.2. GPT-2-derived information-theoretical measures

Information-theoretical metrics of surprisal, entropy, and entropy reduction were estimated using a pre-trained transformer-based language model, the X-large version of GPT-2 (Radford et al., 2019), which was designed for generative language tasks. It can provide the probability distribution of the next word given a word sequence with no more than 1024 words. Empirical evidence showed that the correlation between human and GPT-2's word predictions improved as the contextual window increased, but a 100-word context performed well enough (Goldstein et al., 2022). Therefore, we used the 100-word context to predict the upcoming words. Given the preceding context, every word in the lexicon W can be assigned a probability P at word point t. The calculations of specific information-theoretical measures are described below.

2.2.1. Surprisal

Surprisal was estimated using the negative logarithm of the probability of presented word at time t as defined in Eq. (1):

$$S(t) = -log_2(p_w) \tag{1}$$

where p_w is the probability of the word w presented at word point t, calculated based on the preceding context words. A higher surprisal value indicates an unexpected word, making it challenging to integrate it into the existing internal representation of the known utterance. Since (Hale, 2001) and (Levy, 2008), the concept of surprisal has been associated with the cognitive resources required to process a new word.

2.2.2. Entropy

Entropy was defined as the uncertainty of prediction, the value of which would be maximized when all words were equally likely to appear given the preceding context. Following previous studies, we used the next word entropy to quantify the uncertainty at time t about the next word (the word at time $t+1$) (van Schijndel & Linzen, 2018; Willems et al., 2016). It was calculated as the weighted sum of the surprisal of all candidates, as shown in Eq. (2).

$$H(t+1;t) = -\sum_{v \in W} (p_v | w_1^t) * log_2(p_v | w_1^t) \tag{2}$$

where W is the candidate vocabulary for the next word in GPT-2 model, and p_v is the probability of any word v at time t+1. In the following, we use *entropy* to refer to next word entropy for simplicity.

In addition, to examine the impact of the convolution of metric series with hemodynamic response function, we also calculated the **current-word entropy**. The entropy values (next-word entropy) were shifted forward by one time stamp, that is, the next-word entropy at time t-1 was also the current-word entropy of the word at time t.

However, the entropy about the next word is not the same as the entropy about the rest of the sentence according to Hale (2006), and there was difficulty in defining the uncertainty about a sentence. We followed Frank (2013) in calculating a simplified entropy of sentences using actual words, but we deviated from their approach by summing the entropy of upcoming words instead of applying the chain rule. We use H_s defined by Eq. (3) to refer to the entropy of a sentence.

$$H_s(t) = \sum_{i=1}^{n} H(t+i;t) \tag{3}$$

We adopted n = 4 following the work of Frank and colleagues (Frank, 2013; Frank et al., 2015) so that no more than four upcoming words were taken into account when computing the entropy of the sentence. Note that we padded a placeholder at the end of a sequence to get one step further predictions of the GPT-2 model, which was different from Frank's calculation.

2.2.3. Entropy reduction

The change in prediction between point t and point $t+1$ can be quantified as the decrease of uncertainty, which measures information gained from word t. ER brought by word t can be defined with Eq. (4). Because we were interested in the unidirectional effect of uncertainty decrease, the negative values were set to zeros (Lowder et al., 2018).

$$ER(t) = \max\{H_s(t-1) - H_s(t), 0\} \tag{4}$$

When calculating the measures for the 2132-word stimuli provided in the Alice dataset, we kept the nearest 100 words in the preceding context to make sure that all predictions were based on a fixed length of context. As a result, the first 30 data points in the neural signals that corresponded to the first 100 words were not considered in the prediction model. To preserve information about paragraph segmentations in the model, we retained the period symbols at the end of the paragraphs in the sequences as paragraph delimiters. However, measures of these symbols did not enter subsequent analyses. We note that the language model does not simply segment words by explicit spaces; long, compound words might be split into word pieces or commonly repeated tokens (e.g., *waistcoat* to *waist* and *coat*). Such segmentation of compound words occurred in 1.9% (41/2132) of the words. In these cases, the measures of the first token were used to represent the word.

2.3. Analysis of the information-theoretical metrics

2.3.1. Comparisons between metrics derived from the real and shuffled sequences

We calculated the information-theoretical metrics, including surprisal, entropy, and entropy reduction for the real sequence of the transcription text of the story. To illustrate that the metrics might capture certain properties of the real

sequences in language and differentiate ER from the other metrics, the information-theoretical metrics were computed for sequences of randomly shuffled words from the original text. The two sets of metrics were expected to differ because the shuffled sequence should induce high uncertainty, little compositionality, and little information gain. We utilized kernel density estimation to visualize the probability distribution of different metrics. The shuffling and computation processes were repeated 500 times to provide a comprehensive description of the distinctions between the real and shuffled sequences. For each metric, we computed the mean value for each shuffled sequence, established a normal distribution of mean values, and tested the mean value of the actual sequence against this distribution.

2.3.2. *Bold signal simulations with the metrics*

Following previous work (Brennan et al., 2016; Russo et al., 2020; Willems et al., 2016), we established point events for each metric using the onset of each word provided in the dataset. The intensity of these events is proportional to the metric at the corresponding points. The values were Z-scored among each kind of metric. The events were convolved with a canonical hemodynamic response function (HRF) using the Python package "nipy" (nipy.org) and were summed to yield estimated BOLD responses for each indicator separately. The estimated responses were then down-sampled to .5 Hz to align with the observed responses which were collected every 2 s. These estimated BOLD series were used as predictors of corresponding neural signals in the subsequent process. As a result, we obtained 342 data points after the convolution and down-sampling, corresponding to the presentation time of 2032 words for further analysis.

2.3.3. *Correlations between the metrics and between the simulated neural response*

In addition to prosodic break strength and word frequency, which were provided in the Alice dataset, we computed two other word-level metrics of no interest that are important to modeling neural signals during narrative comprehension: word duration (WD) and root mean squared (RMS) amplitude of the word sound among the duration of word presentation. The WD was calculated as the time difference between the offset and onset of a word; the RMS was calculated as the root mean square of sound power during word presentation, using the audio stimuli provided in the dataset. We calculated the correlations of the metrics derived from the real sequence, including WD, RMS, PBS, LgWF, surprisal, entropy, current-word entropy, and ER. The correlations were calculated with the metric sequences of 2002 words, and correlations among the simulated data series (with a length of 342) corresponding to different metrics were also calculated to show the change brought by the convolution procedure.

2.4. *ROI-based analysis on the VMPFC*

Because of the established role of vmPFC in uncertainty processing, semantic composition, and reward processing, we particularly investigated the vmPFC to disentangle the potential effects of different information-theoretical metrics in this region. The location of vmPFC was determined using neurosynth meta-analysis (neurosynth.org) by selecting the intersection of voxels revealed by the terms *vmpfc* and *cortex vmpfc*.

We performed likelihood ratio tests (LRT) on linear mixed-effects models with and without each predictor of interest. BOLD time courses extracted from each subject's data were averaged across all the voxels in vmPFC and were concatenated across subjects. Predictor series corresponding to the BOLD time courses were treated as the fixed effects and the subject was treated as random effect. The model containing WD, RMS, PBS, LgWF, and subject identity was created as the base model. Each of the target predictors (surprisal, entropy, or ER) was respectively entered to create a nested model. Likelihood ratio test was performed to assess the effect of including the target predictor. Next, to test the unique contribution of each target predictor in addition to the other two, some other base models were built by including two of the predictors, whereas the alternative model included all three predictors. Thus, for each predictor of interest, two LRTs were performed to examine its sensitivity and the unique contribution as compared to the other metrics in explaining the vmPFC responses.

2.5. *Whole-brain analysis*

We then examined the effects of the information-theoretical metrics in accounting for neural responses in any gray matter area. As was done for the nested model comparison, two contrasts were performed in the whole brain for each predictor of interest. That is, in one analysis, the sensitivity of the target predictor was tested. In the other, the unique contribution of each target metric was tested by controlling for the effect of the other two measures. The first-level general linear model estimated the voxel-wise coefficients for and the variance explained by the set of information-theoretical metrics within each subject, with WD, RMS, PBS, and LgWF included as control variables. To disentangle the effects of correlated variables, we first used all the control variables to predict the BOLD signals and recorded the residuals. Then, the predictor of interest was directly used to predict these residuals. This procedure differs from residualizing the target predictor by control predictors used by Brennan et al. (2016) as some analyses with residualized variables cannot be meaningfully interpreted (Wurm & Fisicaro, 2014). In both methods, the effects of control variables were regressed out, but by predicting the residuals of the BOLD signals, the target predictor remained unchanged, ensuring the reliability and interpretation of its coefficients. The values of the convolved target predictor were shuffled, and the shuffled values were then used to predict the residual signals for comparison.

At the group level, beta values from the first-level GLMs from each participant were evaluated with paired-sample t-tests between the real models and the shuffled models. To further ensure the reliability of the reported results, paired-sample t-tests were also applied to the variance explained by the real and shuffled models. We report voxels with a p-value of less than .005 in beta tests and a p-value of less than .05 in variance tests as statistically significant. A cluster-wise correction was applied to the efficiency map to correct for multiple comparisons. We ran 3dFWHMx in AFNI (https://afni.

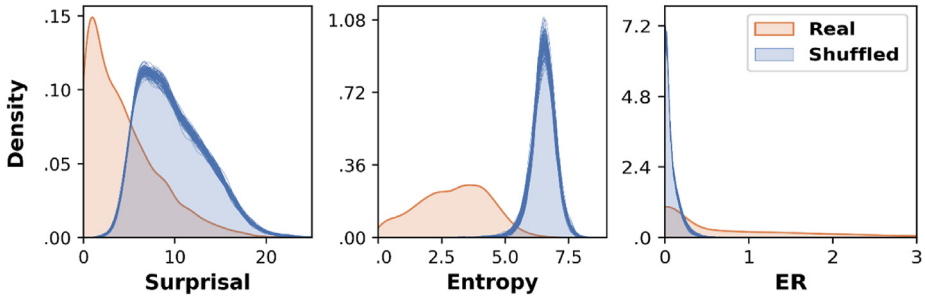

Fig. 1 – Distributions of surprisal, entropy, and entropy reduction of words in the real text and the randomly shuffled sequences. Each line in blue was a kernel density estimation for a shuffled sequence.

nimh.nih.gov/) to estimate the smoothness of the data with a file of a subject that contains the residuals of everything that was not modeled. The output numbers were then used with 3dClustSim to get the number of contiguous voxels that were needed for a cluster at the significant level of $p < .05$ with 10,000 iterations of a Monte Carlo simulation.

3. Results

3.1. *The information-theoretical metrics*

We utilized kernel density estimation to get the distribution of each metric for both the real sequence and 500 shuffled sequences. The distributions were plotted in Fig. 1, showing that ER had a different pattern from the other two metrics. The real sequence had lower surprisal values ($M = 4.266$, $SD = 3.833$, $p < .0001$ against the shuffled means) and lower entropy values ($M = 2.953$, $SD = 1.377$, $p < .0001$ against the shuffled means) than that in shuffled sequences, suggesting higher predictability of the story text than the shuffled sequence. Values of entropy reduction were higher for the real text ($M = .683$, $SD = .995$, $p < .0001$ against the shuffled means), indicating the information gain as the text unfolds. These results suggested that ER could differ significantly from surprisal, despite both being positively linked to processing difficulty to some extent. Moreover, all three metrics in the actual sequence demonstrated significantly higher standard deviations compared to their distributions in shuffled sequences ($p < .0001$), indicating a more dynamic state associated with language processing. These information-theoretical metrics were used below as inferences of different mental states to account for the neural responses to the actual story.

Several text metrics, such as word frequency and surprisal, were naturally correlated over the text (Table 1; Fig. 2). Moreover, the correlations between some of the metrics changed significantly before and after the convolution procedure, as the convolution helped to mitigate sensitivity to temporal variations. For illustration, we presented the correlations between entropy and current-word entropy before and after convolution (Fig. 2). The entropy (of the next word) and current-word entropy vectors are mismatched by *one* time point, where the entropy at time t-1 was the current-word entropy of the word at time t. The correlation between entropy and current-word entropy was only .225 in their raw measures, but became .995 after convolution (Fig. 2d). Thus, some of the seemingly high correlation was a natural consequence of the convolution.

3.2. *Effects of information-theoretical metrics in vmPFC*

The *a priori* defined vmPFC area included 2248 voxels with a resolution of 2 mm. Likelihood ratio tests revealed that all three information-theoretical metrics contributed significantly to fitting the mixed linear models, but only surprisal and ER had significant unique contributions in addition to the other metrics (Table 2). Importantly to the interest of this study, the predictive effect of ER was maintained when entropy and surprisal were pre-included, indicating its independence from the other two metrics.

3.3. *Brain regions sensitive to each of the metrics in the whole brain*

We investigated brain areas that were sensitive to the information-theoretical metrics, including surprisal, entropy, and ER respectively, by comparing the coefficients of the real-effect model with the shuffled-effect model. Consistent with previous studies (Russo et al., 2020; Willems et al., 2016), responses of voxels along the bilateral lateral superior and

Table 1 – Pearson correlations between all pairs of text metrics over time before (outside brackets) and after (inside brackets) convolution with a hemodynamic response function.

	WD	RMS	PBS	LgWF	Surprisal	ER	Entropy
RMS	-.096 (-0.238)	_	_	_	_	_	_
PBS	.516 (.459)	-.228 (-.145)	_	_	_	_	_
LgWF	-.715 (-.676)	.025 (.201)	-.339 (-.334)	—	_	_	_
Surprisal	.240 (.383)	.148 (-.076)	-.040 (.412)	-.318 (-.281)	_	_	_
ER	.067 (-.006)	.080 (-.126)	-.180 (.068)	-.177 (-.031)	.292 (.076)	—	—
Entropy	.044 (.136)	-.015 (-.046)	.203 (.235)	-.042 (-.080)	.077 (.711)	-.568 (.015)	—
Current-word Entropy	.141 (.127)	.128 (-.042)	-.074 (.226)	-.198 (-.078)	.528 (.704)	.489 (.034)	.225 (.995)

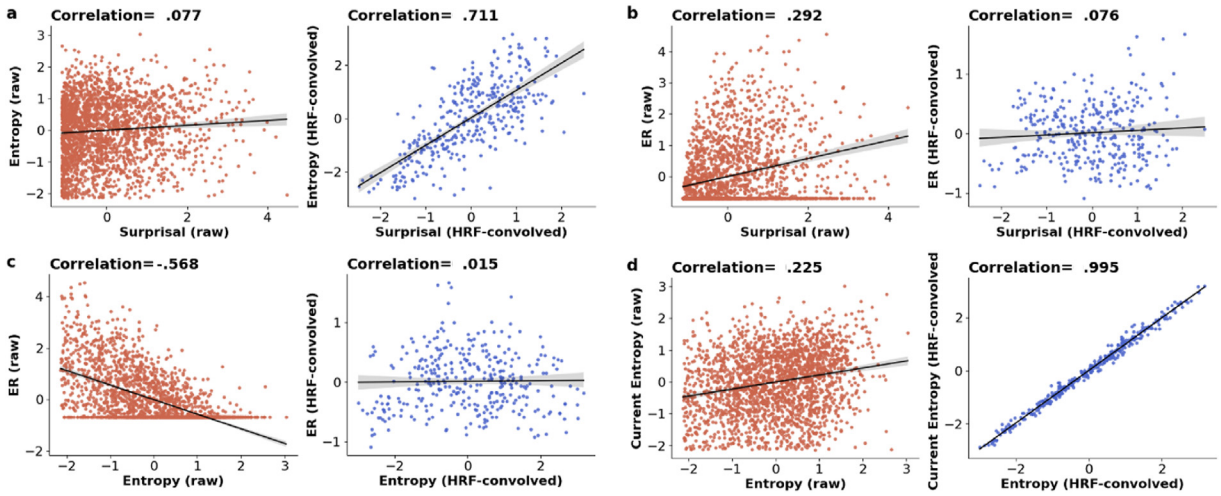

Fig. 2 – Scatter plots of pairs of information-theoretical metrics (surprisal, entropy, entropy reduction, and current-word entropy) over time before and after convolution with a hemodynamic response function. The correlation between current-word entropy and entropy was displayed to illustrate the effect of HRF convolution on overriding the actual relations between the two variables. The black line is the line of best fit. The shaded areas indicate the 95% confidence intervals.

middle temporal gyri (STG and MTG), including the anterior temporal lobe, were significantly predicted by surprisal (Fig. 3a). Moreover, we found that the bilateral middle frontal gyri (MFG) and right middle occipital gyrus were negatively associated with surprisal, which might reflect processing smoothness. The majority (259 out of 283) of the voxels that were positively sensitive to entropy were also sensitive to surprisal, located in the left anterior temporal lobe. The right MFG and the right supramarginal gyrus were negatively predicted by entropy (Fig. 3b). The results of entropy were partially consistent with (Willems et al., 2016), in which only the negatively associated areas were reported.

Importantly, we found areas that were sensitive to ER comprised the positively associated voxels in bilateral vmPFC (overlapped with the vmPFC ROI by 45%, i.e. 135 out of 297 voxels) and the left inferior frontal gyrus (LIFG), and the negatively associated voxels in bilateral insula, thalamus, basal ganglia (BG), and middle cingulate cortex (MCC; Fig. 3c).

3.4. *The unique contribution of each metric in accounting for the whole-brain responses*

The findings above showed that surprisal, entropy, and ER were all able to account for neural responses in different brain regions. We further investigated the unique contribution of each metric beyond the effect of the other two measures.

Unique contributions of ER in addition to surprisal and entropy were found in all the clusters that had been identified as sensitive to ER (Fig. 3f as compared to Fig. 3c). The effects of ER were positive on the vmPFC clusters (overlapped with the vmPFC ROI by 47%, i.e. 119 out of 254 voxels) and the LIFG, and negative on the bilateral insula, thalamus, BG, and MCC. Surprisal accounted for a unique proportion of variance in the left anterior lobe beyond the other two metrics (Fig. 3d). This area was also found in the surprisal-sensitivity analysis (Fig. 3a), but the effect size was reduced when the model included the other two metrics. No region was found to be uniquely explained by entropy. Overall, both entropy reduction and surprisal had unique contributions in explaining the neural responses during language comprehension, but entropy reduction independently explained the variances in the vmPFC and LIFG.

3.5. *The reduction rather than the fluctuation as a contributing factor*

To clarify whether the effect of entropy reduction was specific to the *decrease* of entropy, or whether the entropy fluctuation in any direction was able to account for the neural responses, we further examined the effects of entropy fluctuation (EF; calculated by Hs (t-1) - Hs (t), and the negative values were kept; Z-scored). The same procedure of the whole-brain voxel-wise analysis was implemented.

No clusters survived the cluster-wise correction in the analysis for EF. We counted the number of significant voxels in voxel-wised tests (that is, the cluster-wise correction was not applied) and examined the increase in explained variance across subjects in the predicted voxels. The results showed that ER not only predicted a large number of voxels but also explained significantly greater variance in predicted

Table 2 – Overall and unique effects of entropy, surprisal, and entropy reduction in explaining the mean responses in vmPFC during speech comprehension.

	Entropy	Surprisal	ER	Entropy (unique)	Surprisal (unique)	ER (unique)
LRT	.019*	.007**	.028*	.841	.288	.037*

*: $p < .05$. **: $p < .01$.

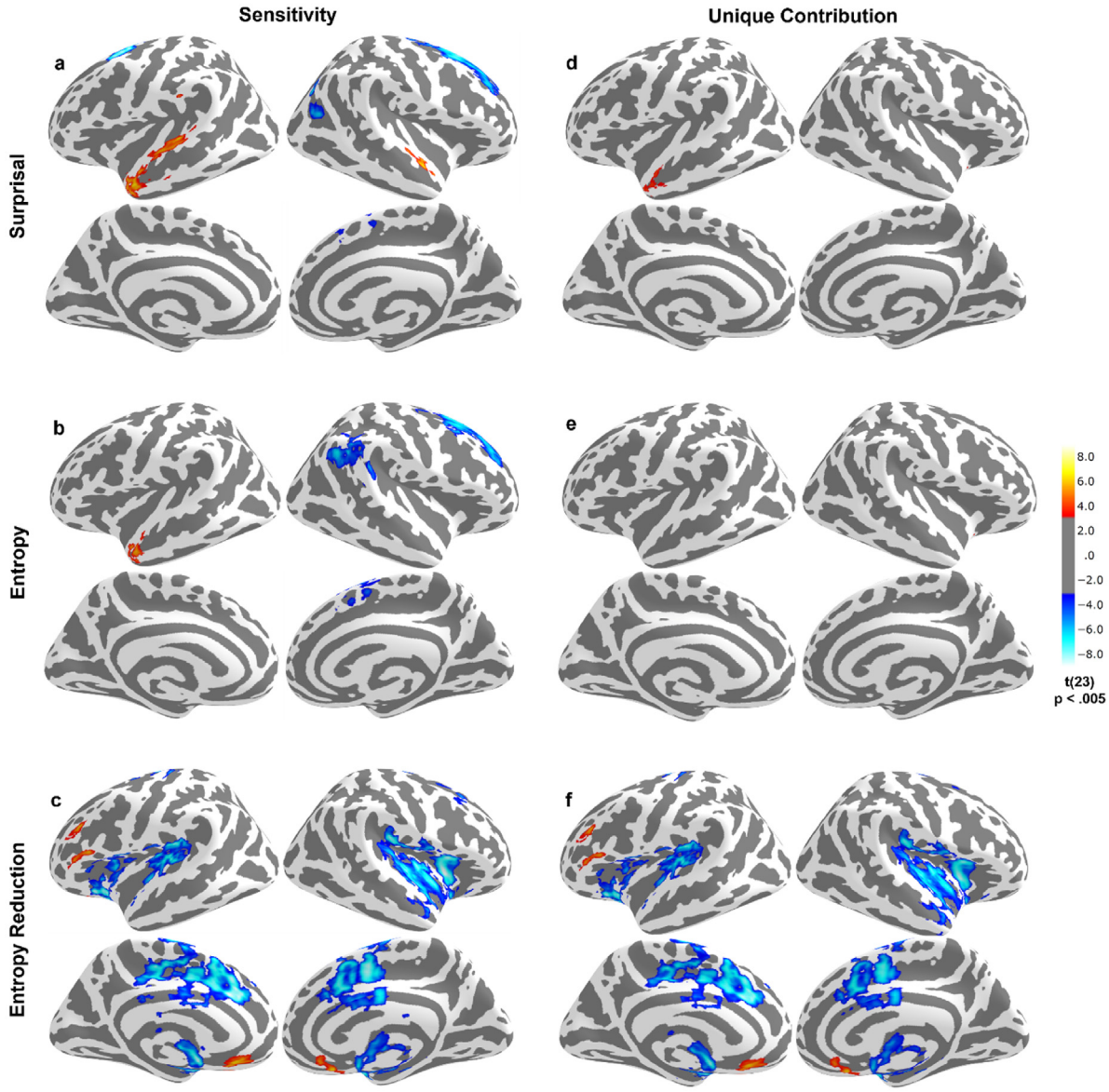

Fig. 3 – The sensitivity (left column) and unique contribution (right column) to the whole brain. Voxel-wise $p < .005$, cluster-wise FWE corrected $p < .05$. The voxels in blue were of negative coefficients and the voxels in orange were of positive coefficients.

voxels, even in the voxels predicted by EF ($t = 49.375$, $p < .0001$; Fig. 4). These results indicated that the reduction, rather than the fluctuation, of uncertainty is an effective factor in modeling neural responses during language comprehension.

3.6. *The relations between the surprisal and the other two metrics*

As most brain regions that were sensitive to entropy were also those sensitive to surprisal, we further analyzed the effect of surprisal to clarify its relationship in competition with entropy and ER respectively. Similar to the unique contribution analysis, we looked for voxels predicted by one metric while controlling for the effect of the other. When entropy was pre-included, the clusters that could be additionally explained by surprisal reduced both in number and size as compared to the surprisal-sensitive map (Fig. 5a, as compared to Fig. 3a). Entropy showed no effects in predicting neural responses in addition to surprisal. Some areas sensitive to surprisal turned out to be not significant after the inclusion of entropy, suggesting that entropy served as a predictor insufficiently alternative to surprisal. The effects of surprisal were not deficient by the pre-inclusion of ER, confirming that they are independent contributing factors.

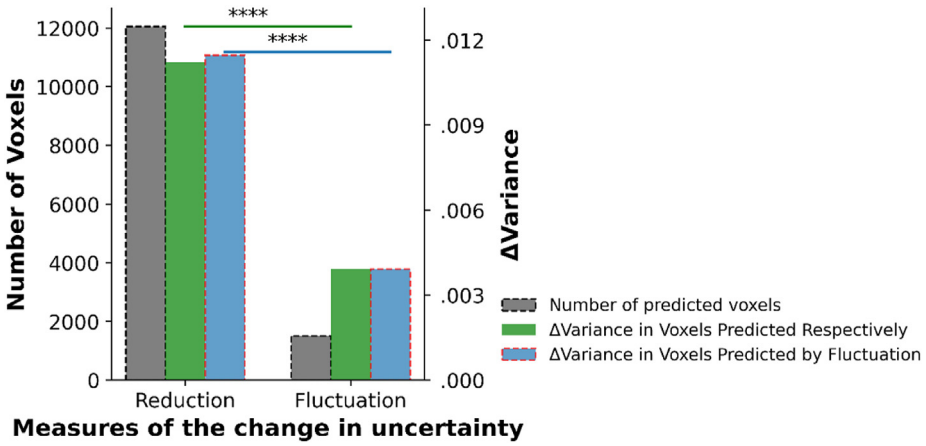

Fig. 4 – Effects of two measures characterizing the change in uncertainty. The grey bars represent the number of predicted voxels by individual measures (the cluster-wise correction was not applied); the green bars indicate the mean variance explained by the measures (in corresponding predicted voxels); the blue bars indicate the mean variance explained by the measures (in voxels predicted by EF).

4. Discussion

In this study, we distinct the effect of information gain of a word (measured by entropy reduction) from processing difficulty (measured by surprisal and entropy) from neural responses during naturalistic narrative processing. Information-theoretical metrics of real and random sequences were derived with a pre-trained GPT-2 model and were used to model neural responses in the brain. In the ROI analysis, we found that entropy, surprisal, and entropy reduction all contributed to predicting the mean neural activities of the ventromedial prefrontal cortex. Then we conducted whole brain analysis to differentiate the effects of entropy reduction in predicting brain activities from that of the other two metrics. We found that ER contributed independently to the other two metrics, indicating a differentiation between information gain and processing difficulty brought by the uncertainty in language. Besides, the analysis of the relations between surprisal and the other two metrics revealed that the effects of entropy in accounting for neural signals during natural comprehension were homologous with that of surprisal.

There are two essential properties of natural language as compared to a random sequence of words: high predictability and informativeness. High predictability is demonstrated by the lower surprisal and entropy in real text compared to a random sequence (Fig. 1), indicating that in the case of a random sequence, surprisal does not necessarily correspond to increased information. The predictability of natural language, which might be the result of recognizing the familiar and generalizing to the similar, makes predicting in language an efficient solution for fast and accurate comprehension despite the noisiness, ambiguity, and speed of our linguistic input (Kleinschmidt & Jaeger, 2015). The high informativeness means that successful comprehension leads to adaption to the novel input and reduction in uncertainty about the future input. Therefore, the degree of entropy reduction during prediction is a sign of how successful the comprehension is, and

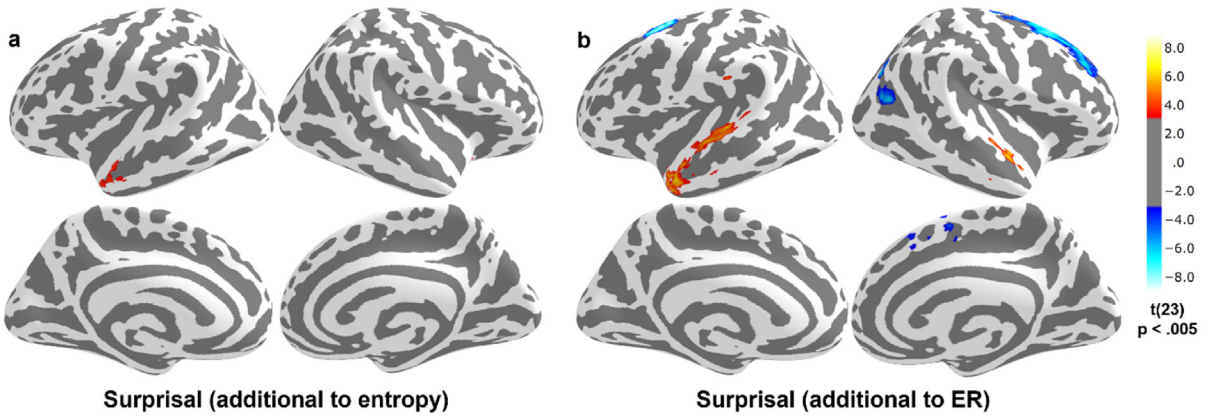

Fig. 5 – Effects of surprisal with one of the other metrics pre-included in the GLMs. (a) the additive effects of surprisal with entropy pre-included. (b) He additive effects of surprisal with ER pre-included. Voxel-wise $p < .005$, cluster-wise FWE corrected $p < .05$. The voxels in blue were of negative coefficients and the voxels in orange were of positive coefficients.

the reduction of entropy has been suggested to reflect end-state confirmation that happens with successful comprehension (Venhuizen et al., 2019).

Entropy reduction positively affected the activations in LIFG and vmPFC beyond the effects of entropy and surprisal (Fig. 3f), which might reflect the pursuit of information regardless of processing difficulty. The activation in LIFG was related to increasing demand for semantic integration (Zhu et al., 2013) and selection among competing semantic alternatives (Moss et al., 2005; Thompson-Schill et al., 1997). It seems to support the ERH that linked ER to processing difficulty, but the sensitivity to long-distance dependencies in LIFG differs from the conceptual combinatory effects reflected by LATL (Pylkkänen, 2019b), thus there's a difference between the processes reflected by ER and that of surprisal. The key is that the semantic alternatives are selected only when the informative comprehension happens, corresponding to the fact that ER appears at only some of the time points and high-gamma power in IFG increases when words can be merged at both the middle and the end of sentences (Nelson, El Karoui, et al., 2017). It is not necessary to limit the activation in LIFG to language-specific comprehension, as it contains a language-selective region and a domain-general MD region (Fedorenko & Blank, 2020). The activation in vmPFC, a domain-general brain area, further suggested that ER can be seen as a certain reward in communication that changed the comprehender's belief about the future. This area was found to be sensitive to uncertainty fluctuation and belief updating previously (Majumdar et al., 2023; Mi et al., 2021), and involved in reward processing (Ciaramelli et al., 2021; Grabenhorst & Rolls, 2011; Rolls, 2022; Strait et al., 2014). This kind of reward, or ongoing uncertainty estimations about future outcomes, might be related to affective experience (Asutay et al., 2020; Majumdar et al., 2023; Stefanova et al., 2020), enhancing the role of vmPFC in processing uncertainty as it is a hub to the generation of affective meaning (Roy et al., 2012).

Another piece of evidence of the reward-related consequence of entropy reduction was the brain areas negatively activated by it, including the bilateral insula, thalamus, basal ganglia (BG), and middle cingulate cortex (MCC). The insula is involved in maintaining and manipulating information, which is crucial for language comprehension (Menon & Uddin, 2010). The thalamus and MCC regulate both cognitive and emotional processing (Fouragnan et al., 2018), while the BG monitors the reward prediction error (O'Doherty et al., 2004). Decreased activations in these areas indicate a lower demand for cognitive resources when long-distance dependencies are solved and uncertainty about the sentence is reduced. The critical point is that uncertainty is aversive, correlating with negative affect, such as vigilance and anxiety (Hirsh et al., 2012; Jackson et al., 2015; Whalen, 2007). Motivation to avoid high uncertainty increases resource allocation in the process of reward pursuit to facilitate uncertainty reduction (Gold et al., 2019; Shen et al., 2015).

Our results on the effects of surprisal and entropy showed important differences from those of previous work (Russo et al., 2020; Willems et al., 2016). Besides bilateral temporal lope, we found broad areas in the frontal lobe negatively activated by surprisal, which was ignored by previous research with the hypothesis that only more positive coeffects were related to surprisal. Regions where the signals were negatively associated with surprisal largely overlapped with the multiple demand network (Duncan, 2010), suggesting that the integration of high-surprisal words may be competitive to other cognitive processes for resources. Willems et al. (2016) reported that signals in several frontal lobe areas were negatively associated with the next-word entropy, but we speculate that it was part of the effects of surprisal. The high correlation between entropy and surprisal after convolution was intermediated by the current entropy at time t (Fig. 2d), as high uncertainty in prediction naturally leads to high prediction error (surprisal). Further evidence is that the activation of the next word entropy in this study was nearly a subset of areas activated by surprisal and the effects of entropy disappeared after the pre-inclusion of surprisal (Fig. 3b–e). However, we do not claim that the next word entropy has no effects beyond surprisal in other studies, especially those that are sensitive to subtle temporal differences.

Further research is needed to investigate the relationship between entropy reductions calculated by different methods. A decrease in grammatical uncertainty was found to be positively correlated with temporal lobe activity (Nelson, Dehaene, et al., 2017) and had a significant effect on anterior frontal electrodes (Hale et al., 2018). Frank and colleagues (Frank, 2013; Frank et al., 2015) confirmed that ER computed by looking several tokens into the future was independent of surprisal, but they failed to identify a corresponding neural component. This may be due to the difficulty in obtaining word probabilities of a much larger number of word candidates compared to grammar candidates. We overcame this quantification obstacle with the GPT-2 model, which is not grammar-based, and found the simplified ER showed effects not only in the IFG but also in the vmPFC and other brain areas. Our results seemed to be different from those in electrode-based studies, but it is not conclusive whether these two kinds of measure refer to the same phenomenon. The positive activation in IFG might be related to the effect on anterior frontal electrodes found by Hale, et al. (2018). The node-opening and node-closing effects, where high-gamma power increased with each successive word in a sentence but decreased suddenly whenever words could be merged into a phrase (Nelson, El Karoui, et al., 2017), provided a syntactical perspective that could be linked to our word-based methods. The closing of nodes leads to a reduction of uncertainty, thus the high-gamma power related to the increasing number of nodes closing agreed with our results to some degree. However, the opening and closing of nodes did not strictly correspond to the increasing and reducing of entropy. More studies with both natural and carefully designed stimuli are needed to clarify this relationship.

Besides, it is crucial to examine how individual differences in uncertainty tolerance affect neural responses during language processing. For instance, curiosity, which is based on temporal uncertainty about the future and the pursuit of a reduction of uncertainty in the future, can be experienced as both negative and positive, and higher curiosity reliably promotes the patience to let information unfold over time (Hsiung et al., 2023). Humans devote a substantial part of their time to seeking and consuming information, to progressively reduce uncertainty about the world around us, and to accrue

information that makes us feel good (van Lieshout et al., 2020). This implies that a variety of emotions, from the basic ones to the sophisticated feelings such as humor and aesthetic affects in language comprehension, could be associated with expectations formed about these dynamics.

In summary, this study reveals that entropy reduction represents the cognitive neural process of information gain that can be dissociated from processing difficulty. The identification of neural substrates that mark the decrease in uncertainty might imply the brain's desire for information regardless of processing difficulty. The processing difficulty brought by the prediction error and the information gain tracked by the entropy reduction are both important aspects of language processing. They differ human language from the random sequence and might be effective measures for the evaluation of the smoothness and informativity of language.

Data/code availability

The primary data used in the analysis, including the text of the story, the audio stimulus, and the fMRI data, are publicly available in the Alice dataset (https://openneuro.org/datasets/ds002322/versions/1.0.4). Additional files and codes for the analysis are available at https://osf.io/2zp7y/.

CRediT authorship contribution statement

Ming Song: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jing Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Qing Cai:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding information

This work was funded by the National Natural Science Foundation of China (31970987 to QC; 32100857 to JW); the Fundamental Research Funds for the Central Universities (YBNLTS2024-043).

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

The authors declare no conflict of interest. This work was funded by the National Natural Science Foundation of China (31970987 to QC; 32100857 to JW); the Fundamental Research Funds for the Central Universities, and the ECNU Academic Innovation Promotion Program for Excellent Doctoral Students (YBNLTS2024-043). We would like to thank the anonymous reviewers for their valuable comments and suggestions, which greatly improved the quality of this paper. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses was pre-registered before the research was conducted.

REFERENCES

Asutay, E., Genevsky, A., Hamilton, J. P., & Västfjäll, D. (2020). Affective context and its uncertainty drive momentary affective experience. *Emotion*. https://doi.org/10.1037/emo0000912

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295–307. https://doi.org/10.1093/cercor/10.3.295

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2010). The original ToBi system and the evolution of the ToBi framework. In *Prosodic typology: The phonology of intonation and phrasing, 1–37*. https://doi.org/10.1093/acprof:oso/9780199249633.003.0002

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience, 31*(8), 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011

Bhattasali, S., Brennan, J. R., Luh, W. M., Franzluebbers, B., & Hale, J. (2020). The Alice datasets: FMRI & EEG observations of natural language comprehension. In *Lrec 2020 - 12th international conference on language resources and evaluation, conference proceedings* (pp. 120–125). https://aclanthology.org/2020.lrec-1.15.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass, 10*(7), 299–313. https://doi.org/10.1111/lnc3.12198

Brennan, J. R., Dyer, C., Kuncoro, A., & Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia, 146*, Article 107479. https://doi.org/10.1016/j.neuropsychologia.2020.107479

Brennan, J., & Hale, J. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *Plos One, 14*(1). https://doi.org/10.1371/journal.pone.0207741

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W. M., & Hale, J. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language, 157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Caucheteux, C., Gramfort, A., & King, J. R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour, 7*(3), 430–441. https://doi.org/10.1038/s41562-022-01516-2

Ciaramelli, E., De Luca, F., Kwan, D., Mok, J., Bianconi, F., Knyagnytska, V., Craver, C., Green, L., Myerson, J., & Rosenbaum, R. S. (2021). The role of ventromedial prefrontal cortex in reward valuation and future thinking during intertemporal choice. *eLife, 10*, 1–17. https://doi.org/10.7554/ELIFE.67387

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron, 29*(2), 537–545. https://doi.org/10.1016/S0896-6273(01)00225-2

Daikoku, T. (2019). Depth and the uncertainty of statistical knowledge on musical creativity fluctuate over a composer's lifetime. *Frontiers in Computational Neuroscience, 13*, Article 441385. https://doi.org/10.3389/fncom.2019.00027

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences, 14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Eisenstein, J. (2019). *Introduction to natural language processing*. MIT press.

Fedorenko, E., & Blank, I. A. (2020). Broca's area is not a natural kind. *Trends in Cognitive Sciences, 24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001

Feldstein, J. H. (1973). Effects of uncertainty reduction, material rewards, and variety on children's choice behavior. *Journal of Experimental Child Psychology, 15*(1), 125–136. https://doi.org/10.1016/0022-0965(73)90136-7

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology, 111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping, 39*(7), 2887–2906. https://doi.org/10.1002/hbm.24047

Frank, S. (2010). Uncertainty reduction as a measure of cognitive processing effort. In J. T. Hale (Ed.), *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics* (pp. 81–89). Association for Computational Linguistics. https://aclanthology.org/W10-2010.

Frank, S. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science, 5*(3), 475–494. https://doi.org/10.1111/tops.12025

Frank, S., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language, 140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Franzluebbers, B., Dunagan, D., Stanojević, M., Buys, J., & Hale, J. T. (2024). *Multipath parsing in the brain* (arXiv:2401.18046). *arXiv*. http://arxiv.org/abs/2401.18046.

Gold, B. P., Gold, B. P., Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., & Zatorre, R. J. (2019). Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience, 39*(47), 9397–9409. https://doi.org/10.1523/JNEUROSCI.0428-19.2019

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience, 25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences, 15*(2), 56–67. https://doi.org/10.1016/j.tics.2010.12.004

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *2nd Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL 2001*. https://doi.org/10.3115/1073336.1073357

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science, 30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass, 10*(9), 397–412. https://doi.org/10.1111/lnc3.12196

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. R. (2018). Finding syntax in human encephalography with beam search. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1*, 2727–2736. https://doi.org/10.18653/v1/p18-1254

Hale, J., Lutz, D. E., Luh, W. M., & Brennan, J. R. (2015). Modeling fMRI time courses with linguistic structure at various grain sizes. In *6th workshop on cognitive modeling and computational linguistics, CMCL 2015 at the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2015 - proceedings* (pp. 89–97). https://doi.org/10.3115/v1/w15-1110

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32), Article 2020.12.03.410399. https://doi.org/10.1073/pnas.2201968119

Hirsh, J. B., Mar, R. A., & Peterson, J. B. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review, 119*(2), 304–320. https://doi.org/10.1037/a0026767

Hodapp, A., & Rabovsky, M. (2021). The N400 ERP component reflects an error-based implicit learning signal during language comprehension. *European Journal of Neuroscience, 54*(9), 7125–7140. https://doi.org/10.1111/ejn.15462

Hsiung, A., Poh, J. H., Huettel, S. A., & Adcock, R. A. (2023). Curiosity evolves as information unfolds. *Proceedings of the National Academy of Sciences of the United States of America, 120*(43), Article e2301974120. https://doi.org/10.1073/pnas.2301974120

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neuroscience: Neural systems responding to degrees of uncertainty in human decision-making. *Science, 310*(5754), 1680–1683. https://doi.org/10.1126/science.1115327

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron, 49*(5), 765–775. https://doi.org/10.1016/j.neuron.2006.01.024

Jackson, F., Nelson, B. D., & Proudfit, G. H. (2015). In an uncertain world, errors are more aversive: Evidence from the error-related negativity. *Emotion, 15*(1), 12–16. https://doi.org/10.1037/emo0000020

Jurafsky, D., & Martin, J. H. (2019). *Speech and language processing 3rd edition draft*. Prentice-Hall.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695

Kringelbach, M. L. (2005). The human orbitofrontal cortex: Linking reward to hedonic experience. *Nature Reviews Neuroscience, 6*(9), 691–702. https://doi.org/10.1038/nrn1747

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural representation of subjective value under risk

and ambiguity. *Journal of Neurophysiology, 103*(2), 1036–1047. https://doi.org/10.1152/jn.00853.2009
Li, J., Brennan, J., Mahar, A., & Hale, J. (2016). Temporal lobes as combinatory engines for both form and meaning. In *Proceedings of the workshop on computational linguistics for linguistic complexity ({CL}4{LC})* (pp. 186–191).
Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science, 40*(6), 1382–1411. https://doi.org/10.1111/cogs.12274
Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science, 42*, 1166–1183. https://doi.org/10.1111/cogs.12597
Majumdar, G., Yazin, F., Banerjee, A., & Roy, D. (2023). Emotion dynamics as hierarchical Bayesian inference in time. *Cerebral Cortex, 33*(7), 3750–3772. https://doi.org/10.1093/cercor/bhac305
Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure & Function, 214*(5), 655–667. https://doi.org/10.1007/s00429-010-0262-0
Mi, Q., Wang, C., Camerer, C. F., & Zhu, L. (2021). Reading between the lines: Listener's vmPFC simulates speaker cooperative choices in communication games. *Science Advances, 7*. https://doi.org/10.1126/sciadv.abe6276
Moss, H. E., Abdallah, S., Fletcher, P., Bright, P., Pilgrim, L., Acres, K., & Tyler, L. K. (2005). Selecting among competing alternatives: Selection and retrieval in the left inferior frontal gyrus. *Cerebral Cortex, 15*(11), 1723–1735.
Nelson, M. J., Dehaene, S., Pallier, C., & Hale, J. (2017). Entropy reduction correlates with temporal lobe activity. In *Cmcl 2017 - cognitive modeling and computational linguistics at the 15th conference of the European chapter of the association for computational linguistics, EACL 2017 - proceedings* (pp. 1–10). https://doi.org/10.18653/v1/w17-0701
Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences of the United States of America, 114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114
Nicki, R. M., & Shea, J. F. (1971). Learning, curiosity, and social group membership. *Journal of Experimental Child Psychology, 11*(1), 124–132. https://doi.org/10.1016/0022-0965(71)90068-3
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452–454. https://doi.org/10.1126/science.1094285
Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002–1044. https://doi.org/10.1037/bul0000158
Pujara, M. S., Philippi, C. L., Motzkin, J. C., Baskaya, M. K., & Koenigs, M. (2016). Ventromedial prefrontal cortex damage is associated with decreased ventral striatum volume and response to reward. *Journal of Neuroscience, 36*(18), 5047–5054. https://doi.org/10.1523/JNEUROSCI.4236-15.2016
Pylkkänen, L. (2019a). Opinion piece Neural basis of basic composition: What we have learned from the red-boat studies and their extensions. https://doi.org/10.1098/rstb.2019.0299.
Pylkkänen, L. (2019b). The neural basis of combinatory syntax and semantics. In *Science, 366* (pp. 62–66). https://doi.org/10.1126/science.aax0050
Pylkkänen, L., Oliveri, B., & Smart, A. J. (2009). Semantics vs. World knowledge in prefrontal cortex. *Language and Cognitive Processes, 24*(9), 1313–1334. https://doi.org/10.1080/01690960903120176
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.
Rolls, E. T. (2022). The hippocampus, ventromedial prefrontal cortex, and episodic and semantic memory. *Progress in Neurobiology, 217*, Article 102334. https://doi.org/10.1016/j.pneurobio.2022.102334
Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences, 16*(3), 147–156. https://doi.org/10.1016/j.tics.2012.01.005
Russo, A. G., De Martino, M., Mancuso, A., Iaconetta, G., Manara, R., Elia, A., Laudanna, A., Di Salle, F., & Esposito, F. (2020). Semantics-weighted lexical surprisal modeling of naturalistic functional MRI time-series during spoken narrative listening. *Neuroimage, 222*, Article 117281. https://doi.org/10.1016/j.neuroimage.2020.117281
Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia, 138*. https://doi.org/10.1016/j.neuropsychologia.2019.107307
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/J.1538-7305.1948.TB01338.X
Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal, 30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x
Shen, L., Fishbach, A., & Hsee, C. K. (2015). The motivating-uncertainty effect: Uncertainty increases resource investment in the process of reward pursuit. *Journal of Consumer Research, 41*(5), 1301–1315. https://doi.org/10.1086/679418
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013
Stefanova, E., Dubljević, O., Herbert, C., Fairfield, B., Schroeter, M. L., Stern, E. R., Urben, S., Derntl, B., Wiebking, C., Brown, C., Drach -Zahavy, A., Kathrin Loeffler, L. A., Albrecht, F., Palumbo, R., Boutros, S. W., Raber, J., & Lowe, L. (2020). Anticipatory feelings: Neural correlates and linguistic markers. *Neuroscience and Biobehavioral Reviews, 113*, 308–324. https://doi.org/10.1016/j.neubiorev.2020.02.015
Strait, C. E., Blanchard, T. C., & Hayden, B. Y. (2014). Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. *Neuron, 82*(6), 1357–1366. https://doi.org/10.1016/j.neuron.2014.04.032
Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences, 94*(26), 14792–14797. https://doi.org/10.1073/pnas.94.26.14792
van Lieshout, L. L., de Lange, F. P., & Cools, R. (2020). Why so curious? Quantifying mechanisms of information seeking. *Current Opinion in Behavioral Sciences, 35*, 112–117. https://doi.org/10.1016/j.cobeha.2020.08.005
van Schijndel, M., & Linzen, T. (2018). Can entropy explain successor surprisal effects in reading?. http://arxiv.org/abs/1810.11481.
Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Semantic entropy in language comprehension. *Entropy, 21*(12). https://doi.org/10.3390/e21121159
Vives, M. L., de Bruin, D., van Baar, J. M., FeldmanHall, O., & Bhandari, A. (2023). Uncertainty aversion predicts the neural expansion of semantic representations. *Nature Human Behaviour, 7*(5), 765–775. https://doi.org/10.1038/s41562-023-01561-5
Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., Von Der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021).

Incremental Language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex, 31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Wentworth, N., & Witryol, S. L. (1984). Uncertainty and novelty as collative motivation in children. *The Journal of Genetic Psychology, 144*(1), 3–17. https://doi.org/10.1080/00221325.1984.10532446

Whalen, P. J. (2007). The uncertainty of it all. *Trends in Cognitive Sciences, 11*(12), 499–500. https://doi.org/10.1016/j.tics.2007.08.016

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van Den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex, 26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1189–1198). Association for Computational Linguistics. https://aclanthology.org/P10-1121.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does *not* do). *Journal of Memory and Language, 72*, 37–48. https://doi.org/10.1016/j.jml.2013.12.003

Yun, J., Chen, Z., Hunter, T., Whitman, J., & Hale, J. (2015). Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics, 24*(2), 113–148. https://doi.org/10.1007/s10831-014-9126-6

Zhu, Z., Feng, G., Zhang, J. X., Li, G., Li, H., & Wang, S. (2013). The role of the left prefrontal cortex in sentence-level semantic integration. *Neuroimage, 76*, 325–331. https://doi.org/10.1016/j.neuroimage.2013.02.060