Entropy Reduction correlates with temporal lobe activity

Matthew J. Nelson,
Stanislas Dehaene[*] and **Christophe Pallier**[†]
Neurospin research center
CEA Saclay
Gif-sur-Yvette 91191 France
{matthew.nelson, stanislas.dehaene, christophe.pallier}@cea.fr

John T. Hale
Linguistics Department
Cornell University
Ithaca, NY 14853 USA
jthale@cornell.edu

Abstract

Using the Entropy Reduction incremental complexity metric, we relate high gamma power signals from the brains of epileptic patients to incremental stages of syntactic analysis in English and French. We find that signals recorded intracranially from the anterior Inferior Temporal Sulcus (aITS) and the posterior Inferior Temporal Gyrus (pITG) correlate with word-by-word Entropy Reduction values derived from phrase structure grammars for those languages. In the anterior region, this correlation persists even in combination with surprisal co-predictors from PCFG and ngram models. The result confirms the idea that the brain's temporal lobe houses a parsing function, one whose incremental processing difficulty profile reflects changes in grammatical uncertainty.

1 Introduction

Incremental complexity metrics connect word-by-word processing data to computational proposals about how parsing might work in the minds of real people. Entropy Reduction is such a metric. It relates the comprehension difficulty that people experience at a word to decreases in uncertainty regarding the grammatical alternatives that are in play at any given point in a sentence (for a review, see Hale 2016). Entropy Reduction plays a key role in accounts of many classic psycholinguistic phenomena (Hale 2003; 2004; 2006) including the difficulty profile of prenominal relative clauses (Yun et al., 2015). It has connected a wide range of behavioral measures to many different theoretical ideas about incremental processing, both with controlled stimuli (Linzen and Jaeger, 2016; Wu et al., 2010) and in naturalistic texts (Frank, 2013). Entropy Reduction and related metrics of grammatical uncertainty have also proved useful in the analysis of EEG data by helping theorists to interpret well-known event-related potentials (beim Graben et al., 2008; beim Graben and Drenhaus, 2012).

This paper applies Entropy Reduction (henceforth: ER) to another type of tightly time-locked brain data: high gamma power electrical signals recorded from the brains of patients awaiting resective surgery for intractable epilepsy. While experimental participants are reading sentences, entropy reductions from phrase structure grammars predict changes in this measured neural signal. This occurred at sites within the temporal lobe that have been implicated, in various ways, in language processing (Fedorenko and Thompson-Schill, 2014; Pallier et al., 2011; Dronkers et al., 2004). The result generalizes across both French and English speakers. The absence of similar correlations in a control condition with word lists suggests that the effect is indeed due to sentence-structural processing. A companion paper explores algorithmic models of this processing (Nelson et al., Under review).

The remainder of this paper is organized into five sections. Section 2 first introduces intracranial recording techniques, as they were applied in our study. Section 3 details the language models that we used, including both hierarchical phrase structure grammars and word-level Markov models. Section 4 goes on to explain the statistical methods, including a complementary "sham" analysis of the word-list control condition where no sentence structure exists. Section 5 reports the results of these analyses (e.g. Table 2 on page 8). Section 6 concludes.

[*,†] additional affiliation: Université Paris 11
[*] additional affiliation: Collège de France

Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017), pages 1–10,
Valencia, Spain, April 7, ©2017 Association for Computational Linguistics

Site	Number of patients	Language	Recording type
Stanford Medical Center	3	English	ECoG
Massachusetts General Hospital	1	English	Depth
Pitié-Salpêtrière Hospital	8	French	Depth

Table 1: Recording site information.

2 Methods: Intracranial recording

2.1 Overview

In intracranial recording, neurological patients volunteer to perform a task while electrodes, implanted in their brains for clinical reasons, continuously monitor neural activity. It offers the most direct measure possible of neural activity in humans, and as such is attractive to researchers from across many disciplines (Fedorenko et al., 2016; Martin et al., 2016; Rutishauser et al., 2006). Recordings can be made either from the cortical surface (referred to here as ECoG, short for electrocorticogram) or from beneath the cortical surface (referred to here as depth recordings). For both types, what is recorded is a spatial average of extracellular potentials generated by neurons in the vicinity of the recording site. This is the same signal as the EEG signal, which has a millisecond temporal resolution, but with a spatial resolution far improved beyond that of EEG. Despite these benefits, there are also limitations to the technique. The recordings are only made in certain hospitals under quite specialized conditions. The number of subjects recorded from are therefore typically smaller than in studies using non-invasive brain-imaging methods. Also, the signals are obtained from patients with brain pathologies, primarily epilepsy. Nevertheless, the latter concern can be mitigated by screening out participants who perform poorly on clinical tests of language function, by discarding data from regions that are later determined to be pathological, or from trials with epileptic activity (see § 2.3.1).

2.2 Patients

Patients from three different hospitals (Table 1) were recorded while awaiting resective surgery as part of their clinical treatment for intractable epilepsy. Written informed consent was obtained from all participants. Experiments were approved by the corresponding review boards at each institution.

2.3 Recordings

Intracranial voltages were low-pass filtered with a 200 Hz to 300 Hz cutoff and sampled at either 1525.88 Hz (SMC) or 512 Hz (MGH and PS). Electrode positions were localized using the method described in Dykstra et al (2011) and Hermes et al (2010) and converted to standard MNI coordinates. Only left hemisphere electrodes were analyzed.

2.3.1 Channel and artifact removal

In intracranial experiments, a portion of channels often show either flat or extremely noisy recorded signals. In both cases this suggests problems with the recording contact and the channel should in general not be analyzed. As mentioned above, channels recording from tissue that was determined to be pathological should also not be analyzed. Here, raw data for each channel were visually inspected for artifacts, such as large jumps in the data, and for channels with little to no signal variation apparent above the noise levels. 7.9% of channels were removed from further analysis in this manner. 10.5% of channels were clinically determined as showing epileptic activity and were also removed from further analysis.

2.3.2 Referencing

To eliminate potential effects from common-mode noise in the online reference and volume conduction duplicating effects in nearby electrodes, recordings were re-referenced offline using a bipolar montage in which the difference in voltage signals between neighboring electrodes was calculated and used for further analysis. For ECoG grids, such differences were computed for all nearest neighbor pairs along both dimensions of the grid. Electrodes that were identified as noisy or pathological were systematically excluded before pairwise differencing. This procedure resulted in 288 bipolar pairs of ECoG electrodes, and 433 bipolar pairs of depth electrodes available for analysis in this dataset. We took the location of each bipolar pair to be the midpoint between the two individual electrode locations. We hence-

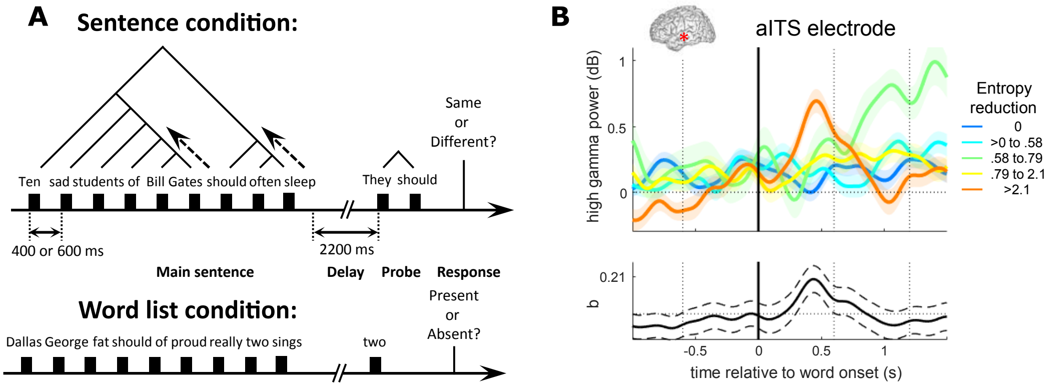

Figure 1: (A) Experimental setup: patients saw sentences or word lists, of variable length, and judged whether a subsequent probe matches or does not match. (B) High-gamma power profile at various levels of Entropy Reduction for the word at time 0. The lower part of panel (B) shows the fitted regression coefficient for Entropy Reduction and its 95% confidence interval across time.

forth refer to these bipolar pairs as electrodes for simplicity. All results presented in this study were essentially unchanged when using an average reference montage.

2.4 Tasks

There were two tasks presented in separate blocks: one in which the stimuli were simple sentences in the participant's native language, and another where the stimuli were randomly-ordered word lists. Figure 1A schematically depicts this arrangement.

In the main sentence task blocks, patients were presented with a sentence of variable length (up to 10 words), followed after a delay of 2.2 seconds by a shorter probe sentence (2-5 words). On 75% of trials, this probe was related to the previous sentence by processes of substitution and ellipsis. For example, a stimulus sentence like "Bill Gates slept in Paris" was followed by probes such as "he did" or "he slept there." On the remaining 25% of trials the probe shared this form, but was unrelated in meaning to the stimulus e.g. "they should." The participants were instructed to press one key if the probe had the SAME meaning and another key if the meaning of the probe was DIFFERENT. This matching task is meant to engage participants' memory for the entire sentence, rather than just one part.

In the word-list task block, patients were presented with the same words used in the preceding sentence task block, but in random order. To avoid any attempt at sentence reconstruction, words were shuffled both within and across sentences. Then following the same delay as in the sentence task, the patients were presented with a one word probe, and asked to identify whether or not that word was in the preceding list. This control task has the same perceptual and motor demands as the main task but with no sequential expectations or sentence-structural interpretation of the stimuli. Sentence and word list tasks were presented in alternating blocks of 80 trials, with 2 to 3 sentence-task blocks and 1 word list block recorded for each patient.

In both sentence and word list conditions, words were presented one at a time at a fixed location on a screen to discourage eye movements. The temporal rate was adapted to individual patients' natural pace, either 400ms (4 patients) or 600ms (8 patients) per word.

3 Materials: language models

We consider two types of language models. The first type comprises linguistically-motivated probabilistic context-free phrase structure grammars (PCFG) based on X-bar theory (Sportiche et al., 2013; Jackendoff, 1977). Figure 2 shows an English example. The hierarchical analyses assigned by this first type of model contrast with those of a second type: word bigram models fitted to Google Ngrams (Michel et al., 2011). Within each type, there are specific English and French versions.

The PCFGs are derived from a computer program that created the stimuli for the intracranial recordings. This program randomly generates well-formed X-bar structures using uniform dis-

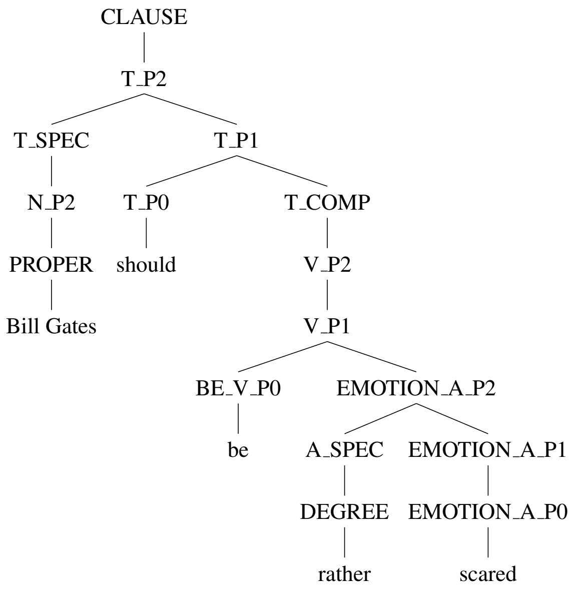

Figure 2: Example analysis from X-bar grammar. P$\{0, 1, 2\}$ annotations indicate bar level, i.e V_P2 means verb phrase, and *should* is analyzed as a projection of tense (T).

tributions. It decides, for instance, on the number of adjuncts present in a particular phrase, the status of each verb as infinitival, transitive or copular, and on nominal properties such as case, person and number. Applying relative frequency estimation to a sample of these trees, we inferred a PCFG that matches the distributions present in the experimental stimuli. For more on this estimation procedure, see Chi (1999).

These language models serve to predict comprehension difficulty via three different incremental complexity metrics, described below.

3.1 Entropy Reduction

Entropy Reduction (ER) is a complexity metric that tracks uncertainty regarding the proper analysis of sentences. If a word comes in that *decreases* grammatical uncertainty, then the metric predicts effort in proportion to the degree to which uncertainty was reduced. Hale (2016) reviews this metric, its motivation and broader implications. Here we characterize precisely the particular ERs that figure in our modeling by reference to a generic sentence w consisting of two concatenated substrings, u and v. Let $w = uv$ be generated by a PCFG G so that $w \in L(G)$ and denote by D_u the set of derivations that derive the k-word initial substring $u_{0\cdots k}$ as a prefix. This initial substring corresponds to the words that the comprehender has already heard, and may be of any length. The existence of at least one grammatical completion, v, restates the requirement that u be a viable prefix. Since G is a probabilistic grammar, each member $d \in D_u$ has a probability. If the Shannon entropy $H(D_u)$ of this set is reduced in the transition from one initial substring to the next, then information-processing work has been done and neural effort is predicted. We compute the predictions of this metric, in both languages, using the freely-available Cornell Conditional Probability Calculator, or CCPC for short (Chen et al., 2014). This program calculates a probability for each derivation $d \in D_u$, conditioned on the prefix string u. It uses exhaustive chart parsing to compute the total probability of D_u following Nederhof and Satta (2008). In order to focus on sentence-structural aspects of comprehension, we follow previous work such as Demberg and Keller (2008) and Yun et al. (2015) in computing this metric at the pre-terminal, rather than word, level.

3.2 Surprisal

The surprisal of a word, in the sense of Hale (2001), links measurable comprehension difficulty to the (negative log) total probability eliminated in the transition from $u_{0\cdots k}$ to $u_{0\cdots k+1}$. We used the CCPC to compute surprisals at the preterminal level from PCFG models. Surprisals from word-bigram models were obtained simply by negative log-transforming the conditional probability of a successor word given the previous word.

3.3 Bigram entropy

This metric is entropic like ER, but ignores structure and deals only with the conditional probability distribution of the next word. We determined this entropy using the counts of all of the bigrams in the Google N-grams database starting with one of the words in our stimuli. This amounted to over 9.2 million unique bigrams in English and 3.3 million in French. In the analysis to follow, these word-bigram models serve as a comparison to the grammatical predictors rather than any sort of positive proposal about human sentence comprehension.

4 Analysis

4.1 Broadband high gamma power

We analyzed the broadband high-gamma power (HGP), which is broadly accepted in the neurophysiological field as reflecting the average activation and firing rates of the local neuronal population around a recording site (Ray and Maunsell, 2011; Miller et al., 2009). We calculated the HGP using wavelet analyses implemented in the FieldTrip toolbox (Oostenveld et al., 2011). We used a wavelet width of 5 and calculated the spectral power over the frequency window spanning from 70 to 150 Hz sampled in the time domain at 1/4 of the raw sampling rate. The resulting power at each time point was then transformed to a decibel scale relative to the entire experiment mean power for each channel for subsequent analyses. The shading of traces in Figure1B reflects the standard errors of the mean across trials.

4.2 Regression analyses

At the single-electrode level, we performed linear regression analyses with each word as the basic unit of observation. The dependent variable was the HGP, averaged over a window from 200 to 500 ms following each word. It is in this time window, more or less, that linguistic effects have been found in behavioral, EEG and MEG data (Pylkkänen et al., 2014; Bemis and Pylkkänen, 2013; Sahin et al., 2009; Friederici, 2002).

We considered the word-by-word Entropy Reduction (ER), as the covariate of interest. To this, we added two other covariates of no interest. One differentiates closed class and open class words, while another summarizes baseline neural activity. We used for the baseline value the average HGP in a 1-second interval before the onset of the first word of a particular stimulus main sentence. This approach, in which the baseline is included as a covariate, improves over the classical subtraction approach because it only accounts for the variance in the dependent variable in common with the baseline term. However for display purposes, Figure 1B depicts the classical subtraction of signal-minus-baseline.

The models shown in Table 3 include an additional covariate of interest: bigram entropy, bigram surprisal and phrase structure surprisal, as introduced above in section 3. The four regression models were thus:

$$\text{HGP} \sim 1 + \text{ER} + \text{Word Class} + \text{Baseline} \tag{I}$$

$$\begin{aligned}\text{HGP} \sim 1 + \text{ER} &+ \text{bigram entropy} \\ &+ \text{Word Class} + \text{Baseline}\end{aligned} \tag{II}$$

$$\begin{aligned}\text{HGP} \sim 1 + \text{ER} &+ \text{bigram surprisal} \\ &+ \text{Word Class} + \text{Baseline}\end{aligned} \tag{III}$$

$$\begin{aligned}\text{HGP} \sim 1 + \text{ER} &+ \text{PCFG surprisal} \\ &+ \text{Word Class} + \text{Baseline}\end{aligned} \tag{IV}$$

We observed the same patterns of results described in this paper when including all of the parameters in one larger model.

4.3 Word list sham analyses

If uncertainty about grammatical structures is indeed driving ER effects when the stimulus is a sentence, then these effects should be stronger than corresponding effects for the same words presented in a random, non-sentential order. To test this, we assigned sham ER values to the word list condition that matched the value in the sentence condition in one of two ways. In Method 1 (word identity matching), each word in the word-list condition was matched to the same word when it occurred in the list condition (possibly at a different position). In Method 2 (word ordinal position matching), each trial in the word-list condition was matched to a trial of the same length in the sentence-task condition. The ER values of the sentence-task trial were then assigned to the word-list trial, matched by ordinal position. We then compared the effect of the real ER values in the sentence task versus the sham values assigned to the word-list task by computing the interaction of that variable across tasks for each sham assignment method. These analyses control for the possibility that either ordinal word position or individual word identity underlie the effects observed in the sentence-task condition.

4.4 Statistical tests

Mixed-effects regression models have become standard in computational psycholinguistics. However sample sizes in intracranial studies are not usually as large, for the reasons mentioned above in subsection 2.1. In such a scenario multilevel models typically gain little beyond classical varying-coefficient models (Gelman and Hill, 2007). We therefore pursued a statistical approach that independently assesses statistical significance across electrodes and participants using two different testing procedures. To make inferences about particular brain areas rather than analyzing the entire heterogeneous sample at once, we pursued this approach in a regions of interest (ROIs) based analysis. Both procedures use as inputs the z-scores of coefficients from the above multiple regression analysis for each electrode located within a given ROI. We derived the z-scores from the p-values of the t-statistics of the coefficients, which account for the degrees of freedom of each test.

The first test tests for significance across electrodes ignoring participant identification using Stouffers z-score method (Zaykin, 2011). This method tests the significance of the z-score sum across electrodes with an assumption of independence between electrodes. Though its independence assumption is likely violated in these data, the test provides a useful benchmark. This test is complemented by the second test that does not make this assumption.

The second test tests for significance across participants (i.e. treating participants as a random factor) using a randomization/simulation procedure that proceeded as follows. For each participant, we observed the highest (and lowest) z-score for all electrodes in the ROI, and calculated the average of these scores across participants that had any electrodes in the ROI. We then simulated independent random z-scores sampled from a standard normal distribution for every electrode in the ROI, with each simulated electrode assigned to a subject to give the same distribution of the number of electrodes per each subject in the ROI found in the real data. With each iteration of the simulation we calculated the mean of the highest simulated z-scores across subjects in the same manner as with the real data, repeating this 100,000 times to obtain a simulated null-distribution of the across-subject mean best z-score expected by chance. The mean highest (and absolute value of the lowest) z-scores across subjects in the actual data were then compared to this null distribution to ascertain the probability of recording such a value of equal or greater extremity in the sample by chance.

By testing whether the effect is consistently observed across multiple participants, this second test avoids concerns about dependence between electrodes. This test benefits from the sensitivity afforded by testing for the best electrode in each subject, especially appropriate in an intracranial recording scenario with a relatively small number of electrodes that are not necessarily positioned at the ideal location for a given effect in each subject. The first test complements this by showing significance over the entire pool of electrodes, not relying on subjects' best electrodes.

Note that an alternate approach for the first test would be to count the number of electrodes in each region with a positive effect significant at the 0.05 level, and use a binomial test to assess the probability of observing at least that many significant electrodes by chance given the total number of electrodes in that region. We prefer Stouffer's z-score method because it does not rely on an arbitrary 0.05 threshold to determine the overall p-value, and because it takes into account the total contribution of every z-score in the sample. We thus chose to report the Stouffers z-score test results only, though we note that the proportions of significant electrodes here support the same patterns of significance.

4.5 Regions of interest (ROI) definition

We defined ROIs independently of our theoretical predictors by finding local maxima of the difference between sentences and baseline activity. The procedure to find these locations proceeded as follows: For each electrode a z-score of the contrast between activation during the sentence and during the baseline period was calculated. A potential ROI center was systematically placed at all possible locations in a 3D grid in the cortex, with 1 mm between possible locations in all directions. The ROI radius was fixed to 22 mm. At each position, the electrodes within the fixed distance from the ROI center were grouped to calculate two independent z-scores. These z-scores were calculated using much the same procedure as above, except that a t-test across the means of subjects was used to assess the across-subjects' z-score, rather than simulations. This avoidance of numerical simulation saved computing time. The two z-scores were combined via a weighted average to determine a composite z-score for each ROI, with a weight of 0.25 and 0.75 assigned to the across-electrodes and across-subjects z-scores respectively. Local maxima of the composite z-score were then detected and ordered according to the composite z-score, discarding local maxima within 22 mm of an-

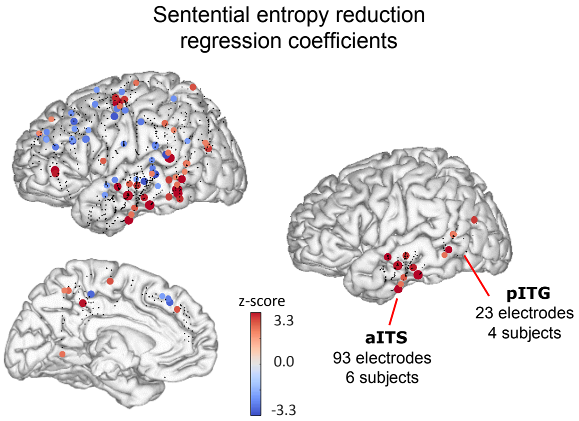

Figure 3: Entropy reduction regression coefficients. Each dot represents an electrode (aggregated across subjects), with size and color saturation reflecting the corresponding z-score for the coefficient corresponding to ER. Z-scores below an absolute value threshold of 1.96 are plotted with a black dot. Electrodes located >10mm from midline are projected onto the lateral surface plot (top), others are projected onto the midline plot (bottom). (Right) The electrodes included in the aITS and pITG ROIs are shown.

other local maxima with a higher z-score. The two highest-scoring local maxima with an anterior/posterior MNI coordinate more posterior than -8 were selected as the ROI centers. These had MNI coordinates: -37,-16,-27 (aITS) and -47,-66,1 (pITG). Figure 3 shows the locations of electrodes within each of these regions.

5 Results & Discussion

ER was observed to correlate with an increase in activity as suggested in Figure 1B on page 3, where data from just one electrode are plotted. Figure 3, above, shows the distribution of the effect across the entire sample. Groups of positive coefficients were observed in the aITS and pITG ROIs, which, as Table 2 shows, were significant across subjects and electrodes. A comparison with sham ER values assigned to the word list task showed that the effect in both areas was significantly higher than word identity matched sham values in the word list task (Table 2, middle). The coefficients in aITS but not pITG were significantly higher than ordinal position matched sham values in the word list task (Table 2, bottom).

In additional multiple regression models, we included other entropy- and surprisal- based predictors alongside ER in two-parameter models. Table 3 shows that there was a significant negative effect of bigram entropy and a positive effect of bigram surprisal in both aITS and pITG, with no effect of PCFG surprisal in either region. ER is still significant in combination with each of these covariates, except with bigram surprisal in pITG, which was significant across subjects but not across electrodes. Overall, we find that ER is positively correlated with temporal activity after accounting for lexical effects and surprisal in its conventional version.

6 Conclusion

Intracranial recordings from patients reading sentences show a correlation with ER in anterior Inferior Temporal Sulcus (aITS) and posterior Inferior Temporal Gyrus (pITG). This occurred even when potential contributions to neural activity from word identity or ordinal position in sentences were accounted for in a control task where there was no syntactic structure. Additionally, aITS and pITG showed a negative response to bigram entropy and a positive response to bigram surprisal. However, the ER effect persisted in aITS when combined with these and other potentially competing effects. These results converge with other findings based on reading time (Wu et al., 2010; Linzen and Jaeger, 2016) that suggest that downward changes in grammatical uncertainty can serve as an approximate quantitative index of human processing effort. We did also observe a positive effect of lexical bigram surprisal, especially in pITG, which has been observed in other work (Nelson et al., Under review), though we focus here on ER.

These results add a precise anatomical localization to this earlier body of work, converging well with findings from MEG (Brennan and Pylkkänen, 2016; van Schijndel et al., 2015), PET (Mazoyer et al., 1993; Stowe et al., 1998) and fMRI (Brennan et al., 2012). As with any correlational modeling result, there is no suggestion of exhaustivity. We do not claim that X-bar grammars are the only ones whose Entropy Reductions would model HGP. But they suffice to do so. This lends credence to the idea, recently underlined by van Schijndel and Schuler (2015), that phrase structure of some sort must figure in realistic models of word-by-word human sentence comprehension.

	Sentences					
	Stouffer's Z		Mean highest Z-score		Mean lowest Z-score	
region	value	p	value	p	value	p
ant. Inferior Temporal Sulcus	**6.01**	**< 0.001**	**3.53**	**< 0.001**	-0.99	0.997
post. Inferior Temporal Gyrus	**4.69**	**< 0.001**	**2.42**	**< 0.001**	-0.87	0.862
	Sentences vs. Word lists- word identity match					
ant. Inferior Temporal Sulcus	**4.23**	**< 0.001**	**2.18**	**0.013**	-1.53	0.642
post. Inferior Temporal Gyrus	**3.42**	**< 0.001**	1.77	0.050	-0.87	0.862
	Sentences vs. Word lists- word ordinal position match					
ant. Inferior Temporal Sulcus	**2.80**	**0.005**	**2.21**	**0.009**	-1.61	0.497
post. Inferior Temporal Gyrus	1.95	0.051	1.49	0.204	-1.22	0.491

Table 2: Hypothesis tests for fitted regression coefficients on model I for Entropy Reduction predictor by region of interest (ROI). The first two columns report statistics obtained using Stouffer's Z-score method, pooling electrodes across human participants. Subsequent columns report the highest and lowest z-score values on a per-participant basis, averaged across participants. The p-values for the mean highest and lowest z-scores were determined using simulations, see § 4.4. The middle and lower tables show the interaction of the regressor across the sentence and word list tasks after assigning sham values to the word list that were matched with the sentence condition values of the same word identity (Middle) and word ordinal position (Bottom). Positive values in these cases indicates a more positive coefficient in the sentence task.

	aITS					
	Stouffer's Z		Mean highest Z-score		Mean lowest Z-score	
predictor	value	p	value	p	value	p
(II) Entropy reduction	**5.43**	**< 0.001**	**3.22**	**< 0.001**	-1.05	0.994
Bigram entropy	**-2.53**	**0.011**	1.32	0.881	**-2.34**	**0.002**
(III) Entropy reduction	**4.70**	**< 0.001**	**2.58**	**< 0.001**	-1.08	0.991
Bigram surprisal	**2.31**	**0.021**	**2.53**	**< 0.001**	-1.27	0.932
(IV) Entropy reduction	**6.05**	**< 0.001**	**3.38**	**< 0.001**	-0.91	0.999
Phrase structure surprisal	-1.16	0.246	1.24	0.951	-1.46	0.741

	pITG					
	Stouffer's Z		Mean highest Z-score		Mean lowest Z-score	
predictor	value	p	value	p	value	p
(II) Entropy reduction	**4.11**	**< 0.001**	**2.38**	**< 0.001**	-0.97	0.779
Bigram entropy	**-2.33**	**0.02**	1.54	0.166	**-2.26**	**0.002**
(III) Entropy reduction	1.63	0.104	**1.98**	**0.013**	-1.61	0.119
Bigram surprisal	**7.56**	**< 0.001**	**5.61**	**< 0.001**	-0.29	0.999
(IV) Entropy reduction	**4.87**	**< 0.001**	**2.22**	**0.002**	-0.5	0.990
Phrase structure surprisal	-1.29	0.196	0.97	0.777	-1.7	0.078

Table 3: Hypothesis tests for fitted regression coefficients for two parameter models including a different co-factor of interest with the Entropy Reduction regressor. Each pair of rows corresponds to the two coefficients of a different two parameter model. Results are shown in the same format as Table 2.

Acknowledgments

This research was supported by an ERC grant "NeuroSyntax" to S.D., the Agence Nationale de la Recherche (ANR, France, Contract # 140301), the Fondation Roger de Spoelberch, and the Fondation Bettencourt Schueller. We thank Peter beim Graben for helpful comments on the manuscript. We would also like to thank Josef Parvizi, Lionel Naccache and Sydney Cash for their collaboration. We thank Profs. Michel Baulac and Vincent Navarro, and Drs. Claude Adam, Katia Lehongre and Stephane Clémenceau for their help with SEEG recordings in Paris. For assistance in data collection we acknowledge Imen el Karoui, Kristof Giber, Vinitha Rangarajan, Sandra Gattas, Xiaofang Yang, the medical staff at Stanford Medical Center, Massachusetts General Hospital and the Pitié-Salpêtrière Hospital, as well as our 12 patients for their participation. We thank Laurent Cohen for input during the project's conception, and Hilda Koopman for assistance in designing the stimuli.

References

Douglas K. Bemis and Liina Pylkkänen. 2013. Flexible Composition: MEG Evidence for the Deployment of Basic Combinatorial Linguistic Mechanisms in Response to Task Demands. *PLoS ONE*, 8(9), September.

Jonathan R. Brennan and Liina Pylkkänen. 2016. Meg evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science*, page Online Version of Record published before inclusion in an issue.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J. Heeger, and Liina Pylkkänen. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163 – 173.

Zhong Chen, Jiwon Yun, Tim Hunter, and John Hale. 2014. Modeling sentence processing difficulty with a conditional probability calculator. In *Proceedings of the* 36^{th} *Annual Cognitive Science Conference*, pages 1856–1857.

Zhiyi Chi. 1999. Statistical properties of probabilistic context-free grammars. *Computational Linguistics, Volume25, number l, March 1999*, 25(1):131–160.

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

Nina F. Dronkers, David P. Wilkins, Robert D. Van Valin, Brenda B. Redfern, and Jeri J. Jaeger. 2004. Lesion analysis of the brain areas involved in language comprehension: Towards a new functional anatomy of language. *Cognition*, 92(1-2):145–177.

Andrew R. Dykstra, Alexander M. Chan, Brian T. Quinn, Rodrigo Zepeda, Corey J. Keller, Justine Cormier, Joseph R. Madsen, Emad N. Eskandar, and Sydney S. Cash. 2011. Individualized localization and cortical surface-based registration of intracranial electrodes. *NeuroImage*.

Evelina Fedorenko and Sharon L. Thompson-Schill. 2014. Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126.

Evelina Fedorenko, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262.

Stefan L. Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494.

Angela D. Friederici. 2002. Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2):78–84, February.

Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. Google-Books-ID: c9xLKzZWoZ4C.

Peter beim Graben and Heiner Drenhaus. 2012. Computationelle neurolinguistik. *Zeitschrift für germanistische Linguistik*, 40(1):97–125. In German.

Peter beim Graben, Sabrina Gerth, and Shravan Vasishth. 2008. Towards dynamical system models of language-related brain potentials. *Cognitive Neurodynamics*, 2(3):229 – 255.

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

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123, March.

John Hale. 2004. The information-processing difficulty of incremental parsing. In Frank Keller, Stephen Clark, Matthew Crocker, and Mark Steedman, editors, *Proceedings of the ACL Workshop on Incremental Parsing: bringing engineering and cognition together*, pages 58–65.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):609–642.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, pages 1–16.

Dora Hermes, Kai J. Miller, Herke Jan Noordmans, Mariska J. Vansteensel, and Nick F. Ramsey. 2010. Automated electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of neuroscience methods*, 185(2):293–298.

Ray Jackendoff. 1977. *$\bar{X}$ Syntax: A Study of Phrase Structure*. MIT Press, Cambridge, Mass.

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6):1382–1411.

Stephanie Martin, José del R. Millán, Robert T. Knight, and Brian N. Pasley. 2016. The use of intracranial recordings to decode human language: Challenges and opportunities. *Brain and Language*, page Corrected Proof Available online 1 July 2016.

Bernard M. Mazoyer, Nathalie Tzourio, Victor Frak, Andre Syrota, Nina Murayama, Olivier Levrier, Georges Salamon, Stanislas Dehaene, Laurent Cohen, and Jacques Mehler. 1993. The cortical representation of speech. *Journal of Cognitive Neuroscience*, 5(4):467–479.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182, January.

Kai J. Miller, Larry B. Sorensen, Jeffrey G. Ojemann, and Marcel Den Nijs. 2009. Power-law scaling in the brain surface electric potential. *PLoS computational biology*, 5(12).

Mark-Jan Nederhof and Giorgio Satta. 2008. Computing partition functions of PCFGs. *Research on Language and Computation*, 6(2):139–162.

Matthew J. Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Lionel Naccache, John T. Hale, Christophe Pallier, and Stanislas Dehaene. Under review. Neurophysiological dynamics of phrase structure building during sentence processing.

Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 02.

Liina Pylkkänen, Douglas K. Bemis, and Estibaliz Blanco Elorrieta. 2014. Building phrases in language production: An MEG study of simple composition. *Cognition*, 133(2):371–384, November.

Supratim Ray and John H. R. Maunsell. 2011. Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLOS Biol*, 9(4):e1000610.

Ueli Rutishauser, Adam N. Mamelak, and Erin M. Schuman. 2006. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. *Neuron*, 49(6):805–813.

Ned T. Sahin, Steven Pinker, Sydney S. Cash, Donald Schomer, and Eric Halgren. 2009. Sequential Processing of Lexical, Grammatical, and Phonological Information Within Broca's Area. *Science*, 326(5951):445–449, October.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*. Wiley-Blackwell.

Laurie A. Stowe, Cees A. J. Broere, Anne M. J. Paans, Albertus A. Wijers, Gijsbertus Mulder, Wim Vaalburg, and Frans Zwarts. 1998. Localizing components of a complex task: Sentence processing and working memory. *Neuroreport*, 9(13):2995–2999.

Marten van Schijndel and William Schuler. 2015. Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL 2015*, Denver, Colorado, USA, June. Association for Computational Linguistics.

Marten van Schijndel, Brian Murphy, and William Schuler. 2015. Evidence of syntactic working memory usage in MEG data. In *Proceedings of CMCL 2015*, Denver, Colorado, USA, June. Association for Computational Linguistics.

Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. 2010. Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1189–1198, Uppsala, Sweden, July. Association for Computational Linguistics.

Jiwon Yun, Zhong Chen, Tim Hunter, John Whitman, and John Hale. 2015. Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics*, 24(2):113–148.

Dmitri V. Zaykin. 2011. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24(8):1836–1841, August.