• EEG → important tool in neuro-linguistics because of very high time resolution and sensitivity to expectation violations.
↳ records the voltage fluctuations on the scalp of the brain.

* Relate time series data from EEG (from each of these electrodes) to intermediate states of language processing models.

Finding Syntax in Human Encephalography with Beam Search

(EEG)

John Hale[♠,△] **Chris Dyer**[♠] **Adhiguna Kuncoro**[♠,♣] **Jonathan R. Brennan**[♢]

[♠]DeepMind, London, UK
[♣]Department of Computer Science, University of Oxford
[♢]Department of Linguistics, University of Michigan
[△]Department of Linguistics, Cornell University

{jthale,cdyer,akuncoro}@google.com jobrenn@umich.edu

* Examines what sort of information needs to be represented inside the intermediate states.

* If the structure is explicitly represented inside the model then reliably can predict patterns in EEG data.

* Some sort of hierarchical sentence structure is needed inside the states.

Abstract

Recurrent neural network grammars (RNNGs) are generative models of (tree, string) pairs that rely on neural networks to evaluate derivational choices. Parsing with them using beam search yields a variety of incremental complexity metrics such as word surprisal and parser action count. When used as regressors against human electrophysiological responses to naturalistic text, they derive two amplitude effects: an early peak and a P600-like later peak. By contrast, a non-syntactic neural language model yields no reliable effects. Model comparisons attribute the early peak to syntactic composition within the RNNG. This pattern of results recommends the RNNG+beam search combination as a mechanistic model of the syntactic processing that occurs during normal human language comprehension.

• P600 → the signature human signal of syntactical processing.

* If the sentence structure is removed from the model, no such predictive structure is obtained.

* RNNGs to score elements in the beam.

RNNGs

• use a stack LSTM to summarize the syntactic context.
⇓
serves as the basis to decide what actions to take, generate word or open a phrase or close a phrase.

• Also used to decide which phrase to open?

• What word to generate.

1 Introduction

Computational psycholinguistics has "always been...the thing that computational linguistics stood the greatest chance of providing to humanity" (Kay, 2005). Within this broad area, cognitively-plausible parsing models are of particular interest. They are mechanistic computational models that, at some level, do the same task people do in the course of ordinary language comprehension. As such, they offer a way to gain insight into the operation of the human sentence processing mechanism (for a review see Hale, 2017).

As Keller (2010) suggests, a promising place to look for such insights is at the intersection of (a) incremental processing, (b) broad coverage, and (c) neural signals from the human brain.

The contribution of the present paper is situated precisely at this intersection. It combines a probabilistic generative grammar (RNNG; Dyer et al., 2016) with a parsing procedure that uses this grammar to manage a collection of syntactic derivations as it advances from one word to the next (Stern et al., 2017, cf. Roark, 2004). Via well-known complexity metrics, the intermediate states of this procedure yield quantitative predictions about language comprehension difficulty. Juxtaposing these predictions against data from human encephalography (EEG), we find that they reliably derive several amplitude effects including the P600, which is known to be associated with syntactic processing (e.g. Osterhout and Holcomb, 1992).

Comparison with language models based on long short term memory networks (LSTM, e.g. Hochreiter and Schmidhuber, 1997; Mikolov, 2012; Graves, 2012) shows that these effects are specific to the RNNG. A further analysis pinpoints one of these effects to RNNGs' syntactic composition mechanism. These positive findings reframe earlier null results regarding the syntax-sensitivity of human processing (Frank et al., 2015). They extend work with eyetracking (e.g. Roark et al., 2009; Demberg et al., 2013) and neuroimaging (Brennan et al., 2016; Bachrach, 2008) to higher temporal resolution.[1] Perhaps most significantly, they establish a general correspondence between a computational model and electrophysiological responses to naturalistic language.

Following this Introduction, section 2 presents recurrent neural network grammars, emphasizing their suitability for incremental parsing. Sections 3 then reviews a previously-proposed

[1]Magnetoencephalography also offers high temporal resolution and as such this work fits into a tradition that includes Wehbe et al. (2014), van Schijndel et al. (2015), Wingfield et al. (2017) and Brennan and Pylkkänen (2017).

2 components

• Incremental Parsing → hear the words one by one and infer the syntactical tree bit by bit in the order that the words appear and the order person heard while EEG was being recorded.

• Beam search → Since there are many possible trees that go with any sentence. use beam search and explore only some number k of top scoring trees.
→ computes neg. log. likelihood.
→ uses RNNGs to assign the neg. log. likelihood and decide what to inc. in beam.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2727–2736
Melbourne, Australia, July 15 - 20, 2018.

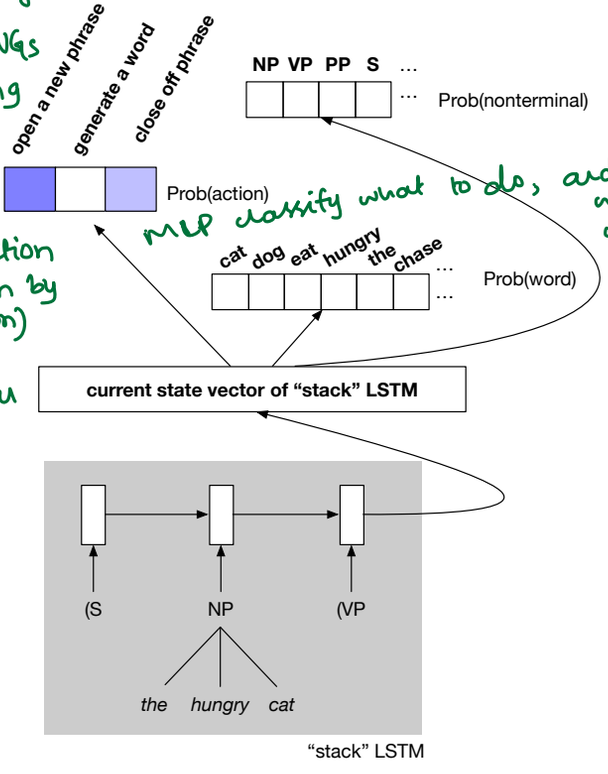


Figure 1: Recurrent neural network grammar configuration used in this paper. The absence of a lookahead buffer is significant, because it forces parsing to be incremental. Completed constituents such as [NP the hungry cat] are represented on the stack by numerical vectors that are the output of the syntactic composition function depicted in Figure 2.

beam search procedure for them. Section 4 goes on to introduce the novel application of this procedure to human data via incremental complexity metrics. Section 5 explains how these theoretical predictions are specifically brought to bear on EEG data using regression. Sections 6 and 7 elaborate on the model comparison mentioned above and report the results in a way that isolates the operative element. Section 8 discusses these results in relation to established computational models. The conclusion, to anticipate section 9, is that syntactic processing can be found in naturalistic speech stimuli if ambiguity resolution is modeled as beam search.

2 Recurrent neural network grammars for incremental processing

Recurrent neural network grammars (henceforth: RNNGs Kuncoro et al., 2017; Dyer et al., 2016) are probabilistic models that generate trees. The probability of a tree is decomposed via the chain rule in terms of derivational action-probabilities that are conditioned upon previous actions i.e. they are history-based grammars (Black et al., 1993). In the vanilla version of RNNG, these steps follow a depth-first traversal of the developing phrase structure tree. This entails that daughters are announced bottom-up one by one as they are completed, rather than being predicted at the same time as the mother.

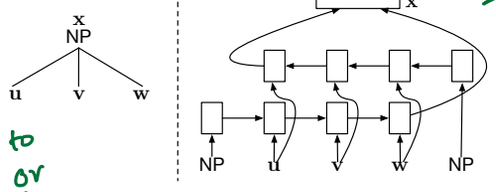


Figure 2: RNNG composition function traverses daughter embeddings $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$, representing the entire tree with a single vector $\mathbf{x}$. This Figure is reproduced from (Dyer et al., 2016).

Each step of this generative story depends on the state of a stack, depicted inside the gray box in Figure 1. This stack is "neuralized" such that each stack entry corresponds to a numerical vector. At each stage of derivation, a single vector summarizing the entire stack is available in the form of the final state of a neural sequence model. This is implemented using the stack LSTMs of Dyer et al. (2015). These stack-summary vectors (central rectangle in Figure 1) allow RNNGs to be sensitive to aspects of the left context that would be masked by independence assumptions in a probabilistic context-free grammar. In the present paper, these stack-summaries serve as input to a multi-layer perceptron whose output is converted via softmax into a categorical distribution over three possible parser actions: open a new constituent, close off the latest constituent, or generate a word. A hard decision is made, and if the first or last option is selected, then the same vector-valued stack–summary is again used, via multilayer perceptrons, to decide which specific nonterminal to open, or which specific word to generate.

Phrase-closing actions trigger a syntactic composition function (depicted in Figure 2) which

- After listening, participants took a comprehension quiz to make sure that they heard comprehension of stories above chance.
- EEG recorded at 500Hz with 61 active electrodes, M10 montage.
- epochs → content words from onset −0.3s to 1s.

Data Preprocessing

- ICA to remove eye-movement artifacts and visual inspection to reject trials with excessive noise (M=13.5%)

squeezes a sequence of subtree vectors into one single vector. This happens by applying a bidirectional LSTM to the list of daughter vectors, prepended with the vector for the mother category following §4.1 of Dyer et al. (2016).

The parameters of all these components are adaptively adjusted using backpropagation at training time, minimizing the cross entropy relative to a corpus of trees. At testing time, we parse incrementally using beam search as described below in section 3.

3 Word-synchronous beam search

Method to relate parser states to EEG data

- Linear Regression → forward modelling → Predicting the neural signal from the parser states.
- → Take the word-by-word processing complexity metrics and form a matrix X (predictors) in LR against Y (voltage from EEG cap at an electrode at a time).

Beam search is one way of addressing the search problem that arises with generative grammars — constructive accounts of language that are sometimes said to "strongly generate" sentences. Strong generation in this sense simply means that they derive both an observable word-string as well as a hidden tree structure. Probabilistic grammars are joint models of these two aspects. By contrast, parsers are programs intended to infer a good tree from a given word-string. In incremental parsing with history-based models this inference task is particularly challenging, because a decision that looks wise at one point may end up looking foolish in light of future words. Beam search addresses this challenge by retaining a collection called the "beam" of parser states at each word. These states are rated by a score that is related to the probability of a partial derivation, allowing an incremental parser to hedge its bets against temporary ambiguity. If the score of one analysis suddenly plummets after seeing some word, there may still be others within the beam that are not so drastically affected. This idea of ranked parallelism has become central in psycholinguistic modeling (see e.g. Gibson, 1991; Narayanan and Jurafsky, 1998; Boston et al., 2011).

As Stern et al. (2017) observe, the most straightforward application of beam search to generative models like RNNG does not perform well. This is because lexical actions, which advance the analysis onwards to successive words, are assigned such low probabilities compared to structural actions which do not advance to the next word. This imbalance is inevitable in a probability model that strongly generates sentences, and it causes naive beam-searchers to get bogged down, proposing more and more phrase structure rather than moving on through the sentence. To address it, Stern et al. (2017) propose a word-synchronous variant of beam search. This variant keeps searching through structural actions until "enough" high-scoring parser states finally take a lexical action, arriving in synchrony at the next word of the sentence. Their procedure is written out as Algorithm 1.

Algorithm 1 Word-synchronous beam search with fast-tracking. After Stern et al. (2017)

1: *thisword* ← input beam
2: *nextword* ← ∅
3: **while** $|nextword| < k$ **do**
4: *fringe* ← successors of all states $s \in$ *thisword* via any parsing action
5: prune *fringe* to top k
6: *thisword* ← ∅
7: **for** each parser state $s \in$ *fringe* **do**
8: **if** s came via a lexical action **then**
9: add s to *nextword*
10: **else** ▷ must have been structural
11: add s to *thisword*
12: **end if**
13: **end for**
14: **end while**
15: **return** *nextword* pruned to top $k_{\text{word}} \ll k$

cardinality

distance # of times around this

In Algorithm 1 the beam is held in a set-valued variable called *nextword*. Beam search continues until this set's cardinality exceeds the designated action beam size, k. If the beam still isn't large enough (line 3) then the search process explores one more action by going around the while-loop again. Each time through the loop, lexical actions compete against structural actions for a place among the top k (line 5). The imbalance mentioned above makes this competition fierce, and on many loop iterations *nextword* may not grow by much. Once there are enough parser states, another threshold called the word beam k_{word} kicks in (line 15). This other threshold sets the number of analyses that are handed off to the next invocation of the algorithm. In the study reported here the word beam remains at the default setting suggested by Stern and colleagues, $k/10$.

Stern et al. (2017) go on to offer a modification of the basic procedure called "fast tracking" which improves performance, particularly when the action beam k is small. Under fast tracking, an additional step is added between lines 4 and 5 of

- $EEG_i^{e,s,t} \sim$ position index$_i$ + acoustic (speaker sound power) + word freq$_{i-1}$ + word freq$_i$ + word freq$_{i+1}$ } nuisance co-regressor don't want to model these
  + SURPRISAL$_{LSTM}$ } regressor of surprisal from LSTM
  + {SURPRISAL$_{RNNG}$ + DISTANCE$_{RNNG}$}

	k=100	k=200	k=400	k=600	k=800	k=1000	k=2000
Fried et al. (2017) RNNG ppl unknown, −fast track	74.1	80.1	85.3	87.5	88.7	89.6	not reported
this paper ppl=141, −fast track	71.5	78.81	84.15	86.42	87.34	88.16	89.81
this paper ppl=141, $k_{ft} = k/100$	87.1	88.96	90.48	90.64	90.84	90.96	91.25

Table 1: Penn Treebank development section bracketing accuracies (F1) under Word-Synchronous beam search. These figures show that an incremental parser for RNNG can perform well on a standard benchmark. "ppl" indicates the perplexity of over both trees and strings for the trained model on the development set, averaged over words In all cases the word beam is set to a tenth of the action beam, i.e. $k_{\text{word}} = k/10$.

Algorithm 1 such that some small number k_{ft} of parser states are promoted directly into *nextword*. These states are required to come via a lexical action, but in the absence of fast tracking they quite possibly would have failed the thresholding step in line 5.

Table 1 shows Penn Treebank accuracies for this word-synchronous beam search procedure, as applied to RNNG. As expected, accuracy goes up as the parser considers more and more analyses. Above $k = 200$, the RNNG+beam search combination outperforms a conditional model based on greedy decoding (88.9). This demonstration reemphasizes the point, made by Brants and Crocker (2000) among others, that cognitively-plausible incremental processing can be achieved without loss of parsing performance.

4 Complexity metrics

In order to relate computational models to measured human responses, some sort of auxiliary hypothesis or linking rule is required. In the domain of language, these are traditionally referred to as complexity metrics because of the way they quantify the "processing complexity" of particular sentences. When a metric offers a prediction on each successive word, it is an *incremental* complexity metric.

Table 2 characterizes four incremental complexity metrics that are all obtained from intermediate states of Algorithm 1. The metric denoted DISTANCE is the most classic; it is inspired by the count of "transitions made or attempted" in Kaplan (1972). It quantifies syntactic work by counting the number of parser actions explored by Algorithm 1 between each word i.e. the number of times around the while-loop on line 3. The information theoretical quantities SURPRISAL and ENTROPY came into more widespread use later. They quantify unexpectedness and uncertainty, respectively, about alternative syntactic analyses at a given point within a sentence. Hale (2016) reviews their applicability across many different languages, psycholinguistic measurement techniques and grammatical models. Recent work proposes possible relationships between these two metrics, at the empirical as well as theoretical level (van Schijndel and Schuler, 2017; Cho et al., 2018).

metric	characterization
DISTANCE	count of actions required to synchronize k analyses at the next word
SURPRISAL	log-ratio of summed forward probabilities for analyses in the beam
ENTROPY	average uncertainty of analyses in the beam
ENTROPY Δ	difference between previous and current entropy value

Table 2: Complexity Metrics

The SURPRISAL metric was computed over the word beam i.e. the k_{word} highest-scoring partial syntactic analyses at each successive word. In an attempt to obtain a more faithful estimate, ENTROPY and its first-difference are computed over *nextword* itself, whose size varies but is typically much larger than k_{word}.

5 Regression models of naturalistic EEG

Electroencephalography (EEG) is an experimental technique that measures very small voltage fluctuations on the scalp. For a review emphasizing its implications vis-á-vis computational models, see Murphy et al. (2018).

We analyzed EEG recordings from 33 participants as they passively listened to a

* 33 participants. First chapter of Alice Adventures in Wonderland
* 8-question comprehension quiz later on → all part. scored > chance.

spoken recitation of the first chapter of Alice's Adventures in Wonderland.[2] This auditory stimulus was delivered via earphones in an isolated booth. All participants scored significantly better than chance on a post-session 8-question comprehension quiz. An additional ten datasets were excluded for not meeting this behavioral criterion, six due to excessive noise, and three due to experimenter error. All participants provided written informed consent under the oversight of the University of Michigan HSBS Institutional Review Board (#HUM00081060) and were compensated $15/h.[3]

* 61 electrodes
* data filtered 0.5–40Hz and baseline corrected against 100ms pre-word interval
* Separated into epochs for content words & function words

Data were recorded at 500 Hz from 61 active electrodes (impedences < 25 kΩ) and divided into 2129 epochs, spanning -0.3–1 s around the onset of each word in the story. Ocular artifacts were removed using ICA, and remaining epochs with excessive noise were excluded. The data were filtered from 0.5–40 Hz, baseline corrected against a 100 ms pre-word interval, and separated into epochs for content words and epochs for function words because of interactions between parsing variables of interest and word-class (Roark et al., 2009).

Linear regression was used per-participant, at each time-point and electrode, to identify content-word EEG amplitudes that correlate with complexity metrics derived from the RNNG+beam search combination via the complexity metrics in Table 2. We refer to these time series as Target predictors.

Each Target predictor was included in its own model, along with several Control predictors that are known to influence sentence processing: sentence order, word-order in sentence, log word frequency (Lund and Burgess, 1996), frequency of the previous and subsequent word, and acoustic sound power averaged over the first 50 ms of the epoch.

All predictors were mean-centered. We also constructed null regression models in which the rows of the design matrix were randomly permuted.[4] β coefficients for each effect were tested against these null models at the group level across all electrodes from 0–1 seconds post-onset, using a non-parametric cluster-based permutation test to correct for multiple comparisons across electrodes and time-points (Maris and Oostenveld, 2007).

6 Language models for literary stimuli

We compare the fit against EEG data for models that are trained on the same amount of textual data but differ in the explicitness of their syntactic representations.

At the low end of this scale is the LSTM language model. Models of this type treat sentences as a sequence of words, leaving it up to backpropagation to decide whether or not to encode syntactic properties in a learned history vector (Linzen et al., 2016). We use SURPRISAL from the LSTM as a baseline.

RNNGs are higher on this scale because they explicitly build a phrase structure tree using a symbolic stack. We consider as well a degraded version, RNNG$_{-\text{comp}}$ which lacks the composition mechanism shown in Figure 2. This degraded version replaces the stack with initial substrings of bracket expressions, following Choe and Charniak (2016); Vinyals et al. (2015). An example would be the length 7 string shown below

(S | (NP | the | hungry | cat |)$_{NP}$ | (VP

Here, vertical lines separate symbols whose vector encoding would be considered separately by RNNG$_{-\text{comp}}$. In this degraded representation, the noun phrase is not composed explicitly. It takes up five symbols rather than one. The balanced parentheses (NP and)$_{\text{NP}}$ are rather like instructions for some subsequent agent who might later perform the kind of syntactic composition that occurs online in RNNGs, albeit in an implicit manner.

In all cases, these language models were trained on chapters 2–12 of Alice's Adventures in Wonderland. This comprises 24941 words. The stimulus that participants saw during EEG data collection, for which the metrics in Table 2 are calculated, was chapter 1 of the same book, comprising 2169 words.

RNNGs were trained to match the output trees provided by the Stanford parser (Klein and Manning, 2003). These trees conform to the Penn Treebank annotation standard but do not explicitly mark long-distance dependency or include any empty categories. They seem to adequately represent basic syntactic properties such

[2] https://tinyurl.com/alicedata

[3] A separate analysis of these data appears in Brennan and Hale (2018); datasets are available from JRB.

[4] Temporal auto-correlation across epochs could impact model fits. Content-words are spaced 1 s apart on average and a spot-check of the residuals from these linear models indicates that they do not show temporal auto-correlation: AR(1) < 0.1 across subjects, time-points, and electrodes.

as clausal embedding and direct objecthood; nevertheless we did not undertake any manual correction.

During RNNG training, the first chapter was used as a development set, proceeding until the per-word perplexity over all parser actions on this set reached a minimum, 180. This performance was obtained with a RNNG whose state vector was 170 units wide. The corresponding LSTM language model state vector had 256 units; it reached a per-word perplexity of 90.2. Of course the RNNG estimates the joint probability of both trees *and* words, so these two perplexity levels are not directly comparable. Hyperparameter settings were determined by grid search in a region near the one which yielded good performance on the Penn Treebank benchmark reported on Table 1.

7 Results

To explore the suitability of the RNNG + beam search combination as a cognitive model of language processing difficulty, we fitted regression models as described above in section 5 for each of the metrics in Table 2. We considered six beam sizes $k = \{100, 200, 400, 600, 800, 1000\}$. Table 3 summarizes statistical significance levels reached by these Target predictors; no other combinations reached statistical significance.

LSTM	not significant	
SURPRISAL	$k = 100$	$p_{cluster} = 0.027$
DISTANCE	$k = 200$	$p_{cluster} = 0.012$
SURPRISAL	$k = 200$	$p_{cluster} = 0.003$
DISTANCE	$k = 400$	$p_{cluster} = 0.002$
SURPRISAL	$k = 400$	$p_{cluster} = 0.049$
ENTROPY Δ	$k = 400$	$p_{cluster} = 0.026$
DISTANCE	$k = 600$	$p_{cluster} = 0.012$
ENTROPY	$k = 600$	$p_{cluster} = 0.014$

Table 3: Statistical significance of fitted Target predictors in Whole-Head analysis. $p_{cluster}$ values are minima for each Target with respect to a Monte Carlo cluster-based permutation test (Maris and Oostenveld, 2007).

7.1 Whole-Head analysis

Surprisal from the LSTM sequence model did not reliably predict EEG amplitude at any timepoint or electrode. The DISTANCE predictor did derive a central positivity around 600 ms post-word onset as shown in Figure 3a. SURPRISAL predicted an early frontal positivity around 250 ms, shown in Figure 3b. ENTROPY and ENTROPY Δ seemed to drive effects that were similarly early and frontal, although negative-going (not depicted); the effect for ENTROPY Δ localized to just the left side.

7.2 Region of Interest analysis

We compared RNNG to its degraded cousin, RNNG$_{-\text{comp}}$, in three regions of interest shown in Figure 4. These regions are defined by a selection of electrodes and a time window whose zero-point corresponds to the onset of the spoken word in the naturalistic speech stimulus. Regions "N400" and "P600" are well-known in EEG research, while "ANT" is motivated by findings with a PCFG baseline reported by Brennan and Hale (2018).

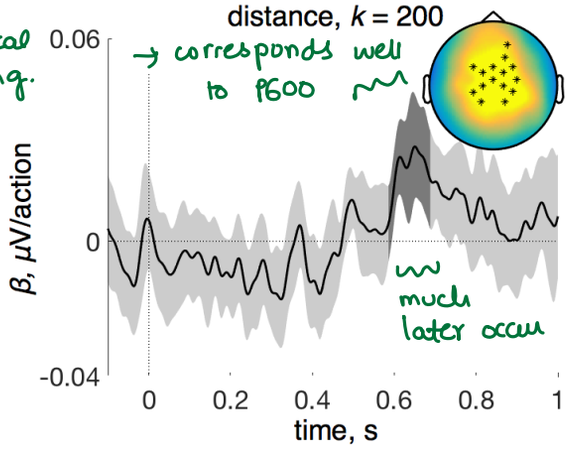
Single-trial data were averaged across electrodes and time-points within each region and fit with a linear mixed-effects model with fixed effects as described below and random intercepts by-subjects (Alday et al., 2017). We used a stepwise likelihood-ratio test to evaluate whether individual Target predictors from the RNNG significantly improved over RNNG$_{-\text{comp}}$, and whether a RNNG$_{-\text{comp}}$ model significantly improve a baseline regression model. The baseline regression model, denoted $\emptyset$, contains the Control predictors described in section 5 and SURPRISAL from the LSTM sequence model. Targets represent each of the eight reliable whole-head effects detailed in Table 3. These 24 tests (eight effects by three regions) motivate a Bonferroni correction of $\alpha = 0.002 = 0.05/24$.

Statistically significant results obtained for DISTANCE from RNNG$_{-\text{comp}}$ in the P600 region and for SURPRISAL for RNNG in the ANT region. No significant results were observed in the N400 region. These results are detailed in Table 4.

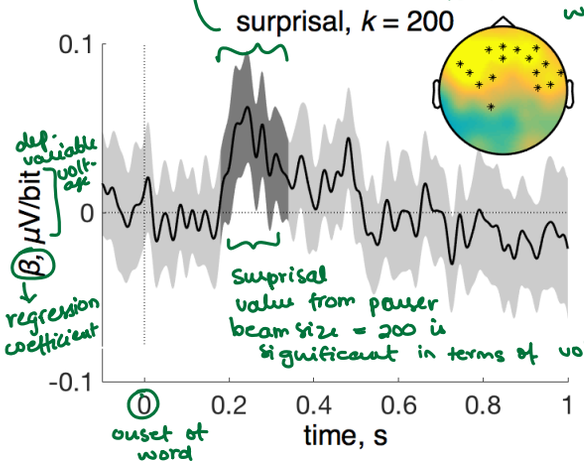
8 Discussion

Since beam search explores analyses in descending order of probability, DISTANCE and SURPRISAL ought to be yoked, and indeed they are correlated at $r = 0.33$ or greater across all of the beam sizes k that we considered in this study. However they are reliably associated with different EEG effects. SURPRISAL manifests at anterior electrodes relatively early. This seems to be a different effect from that observed by Frank et al. (2015). Frank and colleagues relate N400 ampli-

* LSTM surprisal → not significant
* Surprisal → early frontal positivity around 250ms
* Distance → central positivity around 600 ms post-word onset
* Entropy & Entropy Δ → derive similar early & frontal effects but -ve going.



(a) DISTANCE derives a P600 at $k = 200$.



(b) SURPRISAL derives an early response at $k = 200$.

Figure 3: Plotted values are fitted regression coefficients and 95% confidence intervals, statistically significant in the dark-shaded region with respect to a permutation test following Maris and Oostenveld (2007). The zero point represents the onset of a spoken word. Insets show electrodes with significant effects along with grand-averaged coefficient values across the significant time intervals. The diagram averages over all content words in the first chapter of Alice's Adventures in Wonderland.

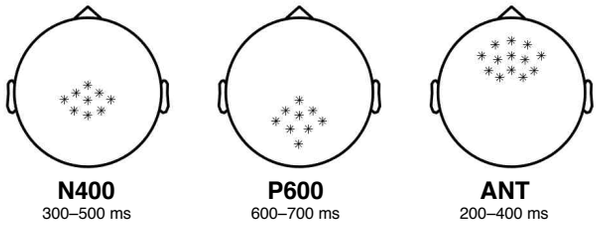


Figure 4: Regions of interest. The first region on the left, named "N400", comprises central-posterior electrodes during a time window 300–500 ms post-onset. The middle region, "P600" includes posterior electrodes 600–700 ms post-onset. The rightmost region "ANT" consists of just anterior electrodes 200-400 ms post-onset.

tude to word surprisals from an Elman-net, analogous to the LSTM sequence model evaluated in this work. Their study found no effects of syntax-based predictors over and above sequential ones. In particular, no effects emerged in the 500–700 ms window, where one might have expected a P600. The present results, by contrast, show that an explicitly syntactic model can derive the P600 quite generally via DISTANCE. The absence of an N400 effect in this analysis could be attributable to the choice of electrodes, or perhaps the modality of the stimulus narrative, i.e. spoken versus read.

The model comparisons in Table 4 indicate that the early peak, but not the later one, is attributable to the RNNG's composition function. Choe and Charniak's (2016) "parsing as language modeling" scheme potentially could explain the P600-like wave, but it would not account for the earlier peak. This earlier peak is the one derived by the RNNG under SURPRISAL, but only when the RNNG includes the composition mechanism depicted in Figure 2.

This pattern of results suggests an approach to the overall modeling task. In this approach, both grammar and processing strategy remain the same, and alternative complexity metrics, such as SURPRISAL and DISTANCE, serve to interpret the unified model at different times or places within the brain. This inverts the approach of Brouwer et al. (2017) and Wehbe et al. (2014) who interpret different *layers* of the same neural net using the same complexity metric.

9 Conclusion

Recurrent neural net grammars indeed learn something about natural language syntax, and what they learn corresponds to indices of human language processing difficulty that are manifested in electroencephalography. This correspondence, between computational model and human electrophysiological response, follows from a system that lacks an initial stage of purely string-based processing. Previous work was "two-stage" in the sense that the generative model served to

Conclusion
* Parsing work is detectable on the scalp in EEG. → from all the different sentence types across the first chapter of Alice in the wonderland.
* Adequate model needs to use hierarchical representations in some way.

Questions
* What mechanisms of sentence processing would account best for the human processing difficulty profile.
* Can we look more specifically at more specific computational alternatives that either do or do not fit.

	RNNG$_{-\text{comp}}$ > ∅ (→ LSTM baseline)			RNNG > RNNG$_{-\text{comp}}$ (↳ failure of LSTM to model the same signal.)		
	χ^2	df	p	χ^2	df	p
DISTANCE, "P600" region						
$k = 200$	13.409	1	**0.00025**	4.198	1	0.04047
$k = 400$	15.842	1	**<0.0001**	3.853	1	0.04966
$k = 600$	13.955	1	**0.00019**	3.371	1	0.06635
SURPRISAL, "ANT" region						
$k = 100$	3.671	1	0.05537	13.167	1	**0.00028**
$k = 200$	3.993	1	0.04570	10.860	1	**0.00098**
$k = 400$	3.902	1	0.04824	10.189	1	**0.00141**
ENTROPY Δ, "ANT" region						
$k = 400$	10.141	1	**0.00145**	5.273	1	0.02165

Table 4: Likelihood-ratio tests indicate that regression models with predictors derived from RNNGs with syntactic composition (see Figure 2) do a better job than their degraded counterparts in accounting for the early peak in region "ANT" (right-hand columns). Similar comparisons in the "P600" region show that the model improves, but the improvement does not reach the $\alpha = 0.002$ significance threshold imposed by our Bonferroni correction (bold-faced text). RNNGs lacking syntactic composition do improve over a baseline model (∅) containing lexical predictors and an LSTM baseline (left-hand columns).

rerank proposals from a conditional model (Dyer et al., 2016). If this one-stage model is cognitively plausible, then its simplicity undercuts arguments for string-based perceptual strategies such as the Noun-Verb-Noun heuristic (for a textbook presentation see Townsend and Bever, 2001). Perhaps, as Phillips (2013) suggests, these are unnecessary in an adequate cognitive model. Certainly, the road is now open for more fine-grained investigations of the order and timing of individual parsing operations within the human sentence processing mechanism.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. 1607441 and No. 1607251. We thank Max Cantor and Rachel Eby for helping with data collection.

References

Phillip M. Alday, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. 2017. Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: event-related potentials reflect continuous model updates. *eNeuro*, 4(6).

Asaf Bachrach. 2008. *Imaging neural correlates of syntactic complexity in a naturalistic context*. Ph.D. thesis, MIT.

Ezra Black, Fred Jelinek, John Lafrerty, David M. Magerman, Robert Mercer, and Salim Roukos. 1993. Towards history-based grammars: Using richer models for probabilistic parsing. In *31st Annual Meeting of the Association for Computational Linguistics*.

Marisa Ferrara Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

Thorsten Brants and Matthew Crocker. 2000. Probabilistic parsing and psychological plausibility. In *Proceedings of 18th International Conference on Computational Linguistics COLING-2000*, Saarbrücken/Luxembourg/Nancy.

Jonathan R. Brennan and John T. Hale. 2018. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. Forthcoming.

Jonathan R. Brennan and Liina Pylkkänen. 2017. MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science*, 41(S6):1515–1531.

Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157-158:81–94.

Harm Brouwer, Matthew W. Crocker, Noortje J. Venhuizen, and John C. J. Hoeks. 2017. A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, 41:1318–1352.

Pyeong Whan Cho, Matthew Goldrick, Richard L. Lewis, and Paul Smolensky. 2018. Dynamic encoding of structural uncertainty in gradient symbols. In

Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 19–28.

Do Kook Choe and Eugene Charniak. 2016. Parsing as language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2331–2336, Austin, Texas.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Daniel Fried, Mitchell Stern, and Dan Klein. 2017. Improving neural parsing by disentangling model combination and reranking effects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–166, Vancouver, Canada.

Edward Gibson. 1991. *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. Ph.D. thesis, Carnegie Mellon University.

Alex Graves. 2012. *Supervised sequence labelling with recurrent neural networks*. Springer.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

John Hale. 2017. Models of human sentence comprehension in computational psycholinguistics. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ronald M. Kaplan. 1972. Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence*, 3:77–100.

Martin Kay. 2005. ACL lifetime achievement award: A life of language. *Computational Linguistics*, 31(4).

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. 2017. What do recurrent neural network grammars learn about syntax? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1249–1258. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190.

Tomáš Mikolov. 2012. *Statistical Language Models Based on Neural Networks*. Ph.D. thesis, Brno University of Technology.

Brian Murphy, Leila Wehbe, and Alona Fyshe. 2018. Decoding language from the brain. In Thierry Poibeau and Aline Editors Villavicencio, editors, *Language, Cognition, and Computational Models*, pages 53–80. Cambridge University Press.

Srini Narayanan and Daniel Jurafsky. 1998. Bayesian models of human sentence processing. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, University of Wisconsin-Madson.

Lee Osterhout and Phillip J. Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31:785–806.

Colin Phillips. 2013. Parser & grammar relations: We don't understand everything twice. In Montserrat Sanz, Itziar Laka, and Michael K. Tanenhaus, editors, *Language Down the Garden Path: The Cognitive and Biological Basis of Linguistic Structures*, chapter 16, pages 294–315. Oxford University Press.

Brian Roark. 2004. Robust garden path parsing. *Natural Language Engineering*, 10(1):1–24.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore.

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.

Marten van Schijndel and William Schuler. 2017. Approximations of predictive entropy correlate with reading times. In *Proceedings of CogSci 2017*, London, UK. Cognitive Science Society.

Mitchell Stern, Daniel Fried, and Dan Klein. 2017. Effective inference for generative neural parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1695–1700, Copenhagen, Denmark.

David J. Townsend and Thomas G. Bever. 2001. *Sentence comprehension : the integration of habits and rules*. MIT Press.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar.

Cai Wingfield, Li Su, Xunying Liu, Chao Zhang, Phil Woodland, Andrew Thwaites, Elisabeth Fonteneau, and William D. Marslen-Wilson. 2017. Relating dynamic brain states to dynamic machine states: Human and machine solutions to the speech recognition problem. *PLOS Computational Biology*, 13(9):1–25.