Multipath parsing in the brain

Berta Franzluebbers
University of Georgia
berta@uga.edu

Donald Dunagan
University of Georgia
dgd45125@uga.edu

Miloš Stanojević
Google DeepMind
stanojevic@google.com

Jan Buys
University of Cape Town
jbuys@cs.uct.ac.za

John T. Hale
University of Georgia
jthale@uga.edu

Abstract

Humans understand sentences word-by-word, in the order that they hear them. This incrementality entails resolving temporary ambiguities about syntactic relationships. We investigate how humans process these syntactic ambiguities by correlating predictions from incremental generative dependency parsers with timecourse data from people undergoing functional neuroimaging while listening to an audiobook. In particular, we compare competing hypotheses regarding the number of developing syntactic analyses in play during word-by-word comprehension: one vs more than one. This comparison involves evaluating syntactic surprisal from a state-of-the-art dependency parser with LLM-adapted encodings against an existing fMRI dataset. In both English and Chinese data, we find evidence for multipath parsing. Brain regions associated with this multipath effect include bilateral superior temporal gyrus.

1 Introduction

A major unsolved problem in computational psycholinguistics is determining whether human sentence comprehension considers a *single* analysis path[1] at a time, or whether it sometimes entertains *multiple* lines of reasoning about the structure of a sentence. These lines of reasoning typically correspond to syntactic ambiguities (e.g. between Subject and Modifier as shown in Figure 1). Multipath accounts of ambiguity resolution (e.g. Gibson, 1991; Jurafsky, 1996) are quite different from single-path accounts (e.g. Frazier and Fodor, 1978; Marcus, 1980). The debate within cognitive science has focused on the special case of garden path sentences leaving open the broader question: Is human comprehension ever multipath in the sense of Ranked Parallel parsing? If it is, where in the brain does such multipath parsing happen?

[1]The "paths" terminology (vs "serial" or "parallel") serves to categorize the cognitive issue as one of process rather than architecture (see Lewis, 2000b, §3.3.3).

Figure 1: A sentence fragment from our stimulus text showing temporary syntactic ambiguity about the correct relationship between *The* and *desert*, which is resolved by hearing the next word (blue).

We take up the call to answer this question (e.g. Lewis, 2000a; Gibson and Pearlmutter, 2000) by combining broad coverage incremental parsing (e.g. Nivre, 2004) with information-theoretical complexity metrics to derive predictions about neuroimaging. Analyzing an entire book (Li et al., 2022) makes it possible to examine – for the first time – all of the ambiguities that are implicated by a given conception of syntax and a given parsing strategy – not just those attested in classic garden path materials (Mason et al., 2003; Hopf et al., 2003; Rodd et al., 2010).

We extend the state of the art in incremental generative dependency parsing by using a large language model (BLOOM; Le Scao et al., 2022b) together with parameter efficient fine-tuning using adapters (Pfeiffer et al., 2020). We fix the parsing strategy while varying the number of allowable analysis paths and connect neural data with parser actions via surprisal (for a review of this information-theoretical metric see Hale, 2016 or neuroimaging studies such as Brennan et al., 2016; Henderson et al., 2016; Shain et al., 2020; Brennan et al., 2020). This sets up a contrast between a multipath model that considers (at most) five paths at a time, versus a single-path model that can only consider one at a time. The results, reported in section 5, ultimately support the multipath view, as the surprisals from the five-way parser are better-correlated with the neuroimaging data. This obtains for both English and Chinese.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12215–12229

Action	Before	After	Arc	Probability
Shift	$(\sigma\|i, j)$	$(\sigma\|i\|j, j+1)$	-	$p_{tr}(\text{sh}\|h_i, h_j)p_{gen}(w_{j+1}\|h_i, h_j)$
Left-arc	$(\sigma\|i, j)$	(σ, j)	$j \to i$	$p_{tr}(\text{re}\|h_i, h_j)p_{dir}(\text{la}\|h_i, h_j)$
Right-arc	$(\sigma\|l\|i, j)$	$(\sigma\|l, j)$	$l \to i$	$p_{tr}(\text{re}\|h_i, h_j)p_{dir}(\text{ra}\|h_i, h_j)$

Table 1: The arc-hybrid transition system of Kuhlmann et al. (2011) defines the possible parser actions (shift, left-arc, and right-arc) as transitions from previous to current parser states. States are indicated by (stack, current index) tuples, e.g. w_i is the token on top of the stack (σ), and w_j is the current index token at time step j. The probabilities associated with these actions are decomposed into complementary shift (sh) and reduce (re) transitions, and complementary left-arc (la) and right-arc (ra) arc directions. The shift action also includes the probability of generating the next token w_{j+1}.

2 Related Work

This paper follows a line of research which aims to characterize human language processing by evaluating word-by-word difficulty predictions against neuroimaging data from the brain. Hale et al. (2022) reviews many studies that fit into this tradition. Most of them consider just a single gold standard or single system-assigned analysis. Hale et al. (2018) and Crabbé et al. (2019) are notable exceptions to this general trend because they derive predictions from multiple analyses that would be considered as part of a beam. Hale and Crabbé work with phrase structure. By contrast, the present study uses dependency parsing. This choice is motivated by prior work in neurolinguistics (e.g. Bornkessel-Schlesewsky and Schlesewsky, 2015). For instance, Li and Hale (2019) relate a dependency-based structural distance metric to hemodynamic activity in left posterior temporal lobe, among other brain areas. Their study used spoken English materials. Lopopolo et al. (2021) apply a related metric to Dutch. Oota et al. (2023) use graph neural networks to embed dependency analyses of sentence-initial substrings. Analyzing brain responses to written, rather than spoken stimuli, via the encoding approach of Reddy and Wehbe (2021), Oota et al. identify many of the same brain areas as the studies mentioned above.

The single-path vs multipath question is itself motivated by prior work with eyetracking data. By varying the number of paths available to a parser, Boston et al. (2011) find support for Ranked Parallel parsing (in the sense of Gibson 1991 and Jurafsky 1996, further characterized below in §3.4).

Apart from these cognitive considerations, a complementary motivation for this research is to investigate the parsing improvements that come by leveraging large language models (LLMs; see accuracy scores in Table 2). Such models are trained on datasets that far exceed classic treebanks in size. This allows LLMs to capture distributional regularities at a vast scale. However, it is challenging to explain at an algorithmic level what these models are doing.

Instead of deriving a processing complexity metric directly from the output of a large language model, or correlating LLM internal representations with fMRI data, as in Schrimpf et al. (2021) and Caucheteux and King (2022), this work uses initial substring encodings from an LLM to inform an incremental parser. Our project also differs from Eisape et al. (2022), who probe LLM representations with a view towards inferring a single, unlabelled dependency analysis. In contrast, we use LLM representations to score the set of all possible alternative analyses. Our scientific goal is to relate such parser states to human brain states, as observed via neuroimaging.

3 Dependency Parsing

3.1 Parser Architecture

The construction of our dependency parser relies on the work of Buys and Blunsom (2018), which employs the transition-based parsing system detailed in Nivre (2008) and exactly enumerates all parser paths. Following Kuhlmann et al. (2011), dynamic programming is used to sum over all paths. We use the arc-hybrid transition system shown in Table 1 because it showed good parsing performance in Buys and Blunsom (2018). The generative parser assigns a probability to each of the words in the sentence incrementally, like any other language model, in contrast to discriminative parsers which only predict transition actions. The *shift* action predicts the next word on the buffer, which corresponds to the *buffer-next* model in Buys and Blunsom (2018).

We update Buys and Blunsom by encoding sequences not with an LSTM but with representations from the BigScience Large Open-science Open-access Multilingual Language Model (BLOOM; Le Scao et al., 2022b), chosen based on its open-access status and training on both English and Chinese data. These BLOOM representations encode an entire initial substring up to and including a particular word (or subword token). A set of classifiers over these substring representations estimates the probability of each transition, word and arc label. The classifiers correspond to the subscripted probabilities in the rightmost column of Table 1.

We use the pre-trained BLOOM-560 model. Larger BLOOM models showed no significant increase in accuracy. This corroborates Oh and Schuler (2023) and Pasquiou et al. (2023), who find that smaller models provide an equal (or better) fit to human data.

Instead of fine-tuning the entire pre-trained BLOOM model on dependency parsing, we apply a Pfeiffer adapter after each layer (Pfeiffer et al., 2020). This bottleneck adapter introduces a new linear layer which reduces the dimension from 1024 down to 64 and back up to 1024, for input into the next pre-trained BLOOM layer. The adapter approach enables parameter-efficient fine-tuning while preserving the original language modelling representations as much as possible in the generative parser.

The goal of these design choices is to deliver linguistically plausible dependency analyses of sentence-initial substrings. Some aspects of the overall architecture play the role of auxiliary hypotheses that do not map on to the brain. As subsection 3.5 lays out in further detail, the cognitive claim is limited to proposals that (a) human parsing involves recognizing dependency relations via a schema like Table 1 and (b) surprising parser actions, within this schema, call for greater hemodynamic resources than do unsurprising ones. Proposal (a) is situated at Marr's middle level of analysis (Marr, 1982).

3.2 Parser Training

The parser is trained on English and Chinese treebanks annotated with Stanford Dependencies (De Marneffe and Manning, 2008). The English model is trained on the Penn Treebank (PTB) version 3, with the standard split of training on sections 02-21, development on section 22, and testing on section 23. The Chinese model is trained on Chinese Treebank version 7, with the standard split

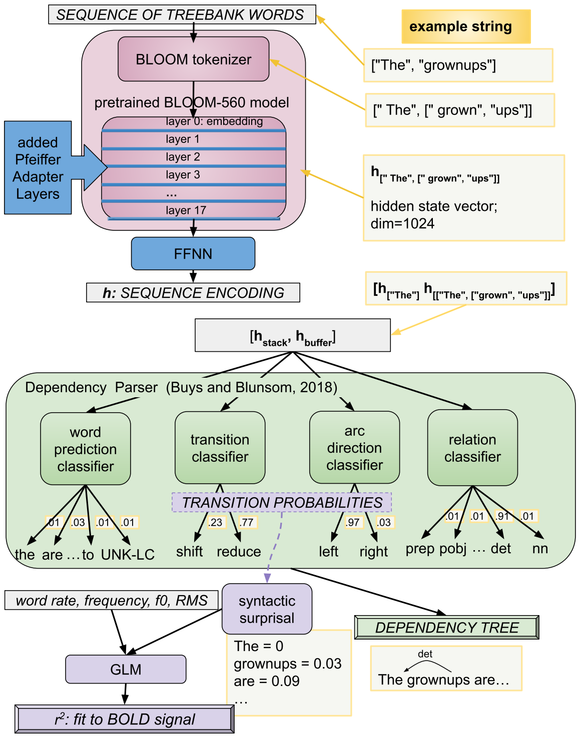

Figure 2: Pipeline Overview: Pretokenized text from the treebank corpora is pre-processed by adding a leading space to each token. The pre-trained BLOOM model is fine-tuned using the Pfeiffer Adapter, and an additional single-layer feedforward neural network completes the encoding step. The dependency parser takes tuples of sequence encodings (stack (σ), buffer (β)) as input, and its system of classifiers is trained from random weights. Note that the word prediction classifier predicts over the size of the training data vocabulary + Berkeley unknown tokens instead of the BLOOM vocabulary. The final dependency tree output is evaluated in Table 2. After training, we use the parser's transition probabilities (see Table 1) to calculate syntactic surprisal (Section 3.5), which is added to other regressors (see Section 4.4), in a General Linear Model (GLM) which predicts BOLD signal (Section 5).

of files 0-2082 for training, development on files 2083-2242, and testing on files 2243-2447. This results in 39,832 training sentences for English and 19,457 training sentences for Chinese.

The parser's next word prediction classifier predicts over a limited vocabulary, where words seen only once in the training data are replaced with unknown word tokens according to the rules in the Berkeley Parser. This results in 23,830 word types for English and 19,671 for Chinese. The LLM encoder, however, uses the BLOOM sub-word tokenizer and the BLOOM vocabulary, which is of size 250,680.[2]

[2]This vocabulary size mismatch is one reason we did not directly assign next-word probabilities using the LLM.

Corpus	Model	Dev		Test	
		LAS	UAS	LAS	UAS
PTB-3	Buys and Blunsom (2018)	88.66	91.19	88.54	91.01
PTB-3	English BLOOM (brain analysis)	90.26	92.71	90.32	92.62
CTB-7	Chinese BLOOM (brain analysis)	77.07	83.65	74.71	81.66

Table 2: Labeled attachment score (LAS) and unlabeled attachment score (UAS) for the English PTB corpus and Chinese CTB corpus

Token sequences are encoded by the LLM, selecting the encoding of the right-most sub-word of each word for parser predictions. This allows sequence encodings to be as detailed as possible while still limiting the vocabulary during training. It should be noted, though, that using this encoding for words that are unknown to the parser's next word classifier prevents the parser from being strictly generative. More information on strictly generative models can be found in the appendix.

While training, sentences are shuffled at each epoch, and within batches (batch size = 16), which are created from sentences of the same length. The BLOOM-560 model has 24 hidden layers, after encoding.

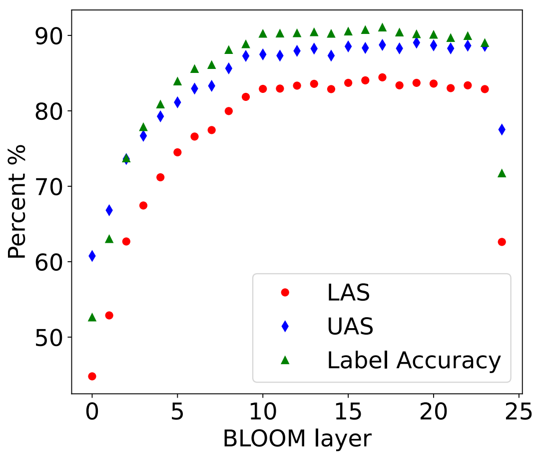

Figure 3: Labeled Accuracy Score (LAS), Unlabeled Accuracy Score (UAS), Label Accuracy for our parser trained and evaluated (dev set) on the Universal Dependencies ParTUT corpus. Each model uses a different BLOOM layer as a sequence encoder.

We select the 17th layer as the representation to input into the model, which was decided based on optimization of development set accuracy on the ParTUT Universal Dependencies corpus (De Marneffe et al., 2021). We use this smaller corpus (2,090 sentences) in order to conserve resources when investigating the utility of BLOOM representations for dependency parsing. As seen in Figure 3, BLOOM layers 10 to 17 have similar accuracy levels. It is possible that more extensive optimization of this layer choice might provide small improvements in accuracy beyond that reported in Table 2. Further training hyperparameters can be found in the Appendix.

3.3 Parser Accuracy

Cognitively plausible parsers must be incremental, in the sense that they must not use information about as-yet-unprocessed words occurring later in the sentence. We compare the accuracy of the incremental generative parser proposed here to that of Buys and Blunsom (2018), which employs the same transition system with encodings based on an LSTM with random initial weights.

Parser accuracy is evaluated on the development set after each epoch, and the highest scoring epoch is reported in Table 2. However, the cognitive modeling results reported in Section 5 rely on earlier training checkpoints. The epoch is chosen by maximizing r^2 correlation with the fMRI data. The labeled attachment score and unlabeled attachment score (LAS/UAS) for the epochs chosen are 89.75/92.37 for English, and 77.06/83.59 for Chinese. We find that the objective of obtaining minimal loss does not correspond to optimal correlation with fMRI activity.

3.4 Multiple paths

Multipath parsing, as a psycholinguistic claim about human sentence comprehension, is simply the idea that ambiguity-resolution pursues more than one alternative at the same time. This contrasts with Frazier and Fodor's (1978) conception, in which a single parse is developed along one path. We consider here a version of the multipath idea that Kurtzman (1984) calls "strong" parallelism, in which paths can persist from word to word. Gorrell (1987) emphasizes that multiple alternatives are not equally available (page 84). This is naturally formalized by ranking the paths. Such ranking can be implemented with beam search, at the heart of parsing systems like Roark (2001), which is widely applied in computational psycholinguistics and neurolinguistics.

3.5 Complexity Metric

In order to correlate parser predictions with brain activity, and in particular to take multiple parser paths into consideration, those predictions must somehow be summarized and quantified. One way to do that is via the surprisal complexity metric, revived by Hale (2001) as a method of predicting sentence-processing difficulty word-by-word. The basic formula is the logarithm of the reciprocal of a probability.

$$\log_2\left(\frac{1}{p(y)}\right) = -\log_2(p(y)) \tag{1}$$

In Equation 1, $p(y)$ denotes the probability of an arbitrary outcome y. In incremental sentence comprehension the relevant outcomes are words. The surprisal of a word at position w_i which follows a string ending in w_{i-1}, is a ratio of the probability at the current word with the previous probability:

$$-\log_2\left(\frac{p(w_i, w_{i-1}, ..., w_1))}{p(w_{i-1}, ..., w_1))}\right) \tag{2}$$

3.5.1 Syntactic Surprisal

Word probabilities, denoted p_{gen} in Table 1, are affected by nonsyntactic factors. As Figure 4 shows, these nonsyntactic factors can overshadow the difference between syntactic alternatives. For this reason, we follow Roark et al. (2009) in decomposing surprisal into lexical and syntactic expectations. Discarding lexical expectations, we focus exclusively on differences among competing syntactic analyses by adopting their *syntactic surprisal* measure, given below as Equation 3. Syntactic surprisal does not consider word probability at the current time step (i), but the complete path probability is used in the denominator ($i - 1$) to ensure that the correct path is selected.

In addition, we limit the sum to the top k transition sequences. Here, j indexes discrete ranks of a ranked parallel parser, and $t_{a(i)}$ refers to all the parser transitions in the path that ends at generation of token w_i. At each time step i, $t_{a(i)}$ includes 1 shift transition, and zero to many possible reduce transitions.

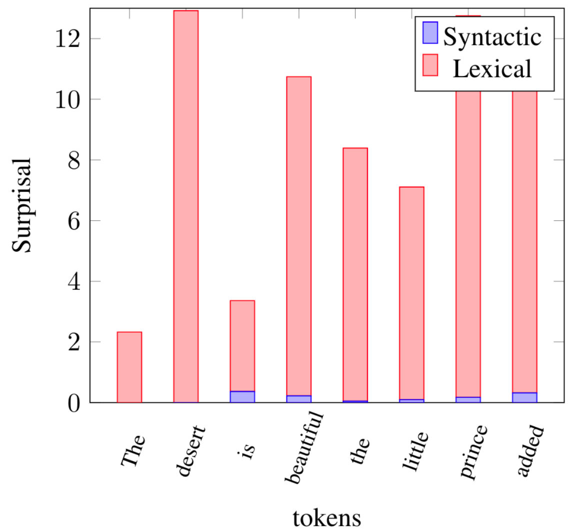

Figure 4: A sentence from *The Little Prince*, showing incremental surprisal from joint probability decomposed into lexical (red), and syntactic surprisal (blue)

$$\mathrm{SynS}_k(w_i) = -\log_2 \frac{\sum_{j=1}^{k} p\left(t_{a(i)} \ldots t_{a(1)} \mid w_{i-1} \ldots w_1\right)_j}{\sum_{j=1}^{k} p\left(t_{a(i-1)} \ldots t_{a(1)} \mid w_{i-1} \ldots w_1\right)_j} \tag{3}$$

In Equation 3, syntactic surprisal is parameterized by a maximum number of allowable analysis paths, k. Figure 5 elaborates an example calculation using this definition.

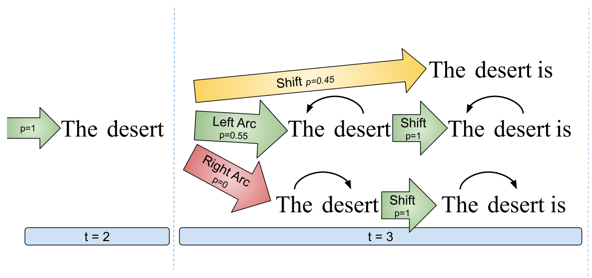

Figure 5: A sentence fragment from *The Little Prince* showing shift-reduce parser actions with associated probabilities. Word generation probabilities are not shown here, but would be included in the full path calculation. In this example, the syntactic surprisal at time t = 3 is -$\log_2$(0.55/1) = 0.86 for k=1, and for k=2 the top two paths are added: -log((0.55+0.45)/1) = 0.

The limitation to k paths in Equation 3 follows Boston et al. (2011). Unlike Boston et al., we retain all analyses and simply chose the top k at each successive word. This allows paths that would have been lost forever in true beam search to later rejoin the beam. Such restoration has essentially the same effect as backtracking, in cases where it would lead to a higher-scoring analyses. "Reanalysis" of this sort could be helpful in modeling the comprehension of garden-path sentences (see e.g. Fodor and Ferreira, 1998).

3.5.2 Why Five Paths?

Addressing the single-path vs multipath question entails choosing some number of paths, k to represent multipath parsing in general. Empirical literature such as Gibson and Pearlmutter (2000) has been primarily concerned with the distinction between 1 and 2 paths. By contrast, accurate NLP systems sometimes consider tens or even hundreds of paths.[3] This mismatch calls for a setting of k that is small, but clearly different from 1. The syntactic surprisal metric itself also mitigates against settings of k that are too large. With this metric, more ranks quickly eat up the available probability mass, up to 1.0, when considering all possible paths. In this limiting case syntactic surprisal would trivially equal zero for all tokens (see Figure 6). For these reasons, we selected five paths to contrast with single-path parsing.[4]

4 fMRI Methods

4.1 Participants

As detailed in Li et al. (2022), the English dataset includes 49 participants (30 female, mean age = 21.3, range = 18-37), and the Chinese dataset includes 35 participants (15 female, mean age = 19.9, range = 18-24).

4.2 Data Acquisition

The English audio stimulus is an English translation of *The Little Prince*, read by Karen Savage. The Chinese audio stimulus is a Chinese translation of *The Little Prince*, read by a professional female Chinese broadcaster. The English and Chinese audiobooks are 94 and 99 minutes in length, respectively. The presentations were divided into nine sections, each lasting around ten minutes. Participants listened passively to the nine sections and completed four quiz questions after each section (36 questions in total). These questions were used to confirm participant comprehension of the story.

The English and Chinese brain imaging data were acquired with a 3T MRI GE Discovery MR750 scanner with a 32-channel head coil. Anatomical scans were acquired using a T1-weighted volumetric magnetization prepared rapid gradient-echo pulse sequence. Blood-oxygen-level-dependent (BOLD) functional scans were acquired using a multi-echo planar imaging sequence with online reconstruction (TR = 2000 ms; TE's = 12.8, 27.5, 43 ms; FA = 77°; matrix size = 72 x 72; FOV = 240.0 mm x 240.0 mm; 2x image acceleration; 33 axial slices, voxel size = 3.75 x 3.75 x 3.8 mm).

4.3 Data Preprocessing

The English and Chinese fMRI data were preprocessed using AFNI version 16 (Cox, 1996). The first 4 volumes in each run were excluded from analyses to allow for T1-equilibration effects. Multi-echo independent components analysis (ME-ICA), was used to denoise data for motion, physiology, and scanner artifacts (Kundu et al., 2012). Images were then spatially normalized to the standard space of the Montreal Neurological Institute (MNI) atlas, yielding a volumetric time series resampled at 2 mm cubic voxels.

4.4 Statistical Analysis

The goal of the analysis is to compare surprisal values from single-path and multipath parsers against the same observed fMRI timecourses. We follow Crabbé et al. (2019) in evaluating goodness of fit pairwise using cross-validated coefficient of determination (r^2) maps.

For each subject individually, the fMRI BOLD signal is modeled by a General Linear Model (GLM) at each voxel. The following five regressors are included in the GLM: word rate, fundamental frequency (f0), word frequency, root mean square intensity (RMS), and syntactic surprisal of the top k paths. Word rate is a timing function marking the offset of each spoken word; f0 is the fundamental frequency, or pitch of the audio; word frequency is obtained from words in a movie subtitles database (Brysbaert and New,

[3] Many of these paths lead to similar syntactic structures. Grouping them in a cognitively realistic way brings up foundational questions regarding the mental representation of grammatical relations (see Bresnan and Kaplan 1982, Sturt 1996, and §4.4 of Brasoveanu and Dotlačil 2020, among others). These questions, along with the proper formulation of reanalysis or repair operations (e.g. Lewis, 1992; Buch-Kromann, 2004) are important directions for future work.

[4] This goal of this study is to adjudicate between single-path and multipath parsing. This question has remained open within the literature on human sentence processing for quite some time. An alternative approach seeks to maximize correlations with fMRI data by e.g. searching for the optimal number of paths. This suggests a natural follow-up.

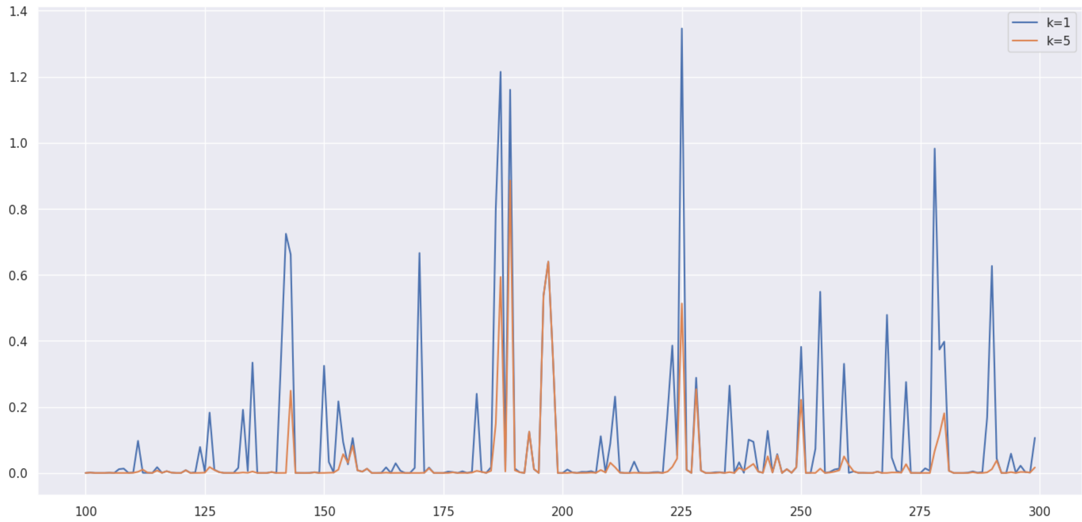

Figure 6: Syntactic surprisal at $k = 1$ (blue) and $k = 5$ (orange) for tokens in *The Little Prince*

2009); and RMS is an indicator of audio intensity, taken every 10ms. These predictors are known to affect speech comprehension, and are included as control variables to help isolate variance within the BOLD signal that is specific to syntactic processing.

Syntactic surprisal for each word was also aligned to the offset of each word in the audiobook. All predictors were convolved using SPM's canonical Hemodynamic Response Function (Friston et al., 2007).

For each participant, we compute how much the inclusion of the syntactic surprisal regressor increases the cross-validated r^2 over the baseline model, which includes only control variables. We repeat this process for syntactic surprisal at different levels of k separately in order to keep the number of parameters in each GLM model constant. Therefore, the r^2 *increase* scores represent the variance explained in each voxel by the addition of syntactic surprisal to the model as a predictor.

To compare models across two different levels of k and analyze their ability to explain the fMRI BOLD signal, we performed a paired t-test on individual r^2 *increase* maps to obtain z-maps. These z-maps show where syntactic surprisal from a model with a given number of paths explains the signal significantly better than does a model that incorporates a different number of paths (see Figures 7 and 8).

5 Results

Both English and Chinese paired t-tests on r^2 increase maps show differences in the superior temporal gyrus (STG). The Chinese results further show differences in middle temporal gyrus (MTG), while the English results show differences in the parietal lobe, including bilateral angular gyrus. The full list of statistically significant clusters corresponding to Figures 7 and 8 can be found in the Appendix. Although the paired t-tests were two-tailed, all of the resulting significant clusters show greater r^2 increase for syntactic surprisal at $k = 5$ than at $k = 1$. The r^2 increase maps for each model individually can be found in the Appendix.

6 Discussion

Multipath parsing effort, outside of the 1-best analysis, localizes to superior temporal regions of the brain bilaterally in both English and Chinese. This finding converges with Crabbé et al. (2019). Crabbé and colleagues parse phrase structure, rather than dependencies, employ a different complexity metric, and use a very large beam (400). Despite all these differences, their results implicate roughly the same brain regions.

These results also cohere with proposals regarding the large-scale organization of language processing in the brain. On the Neuroanatomical Pathway model, for instance, "most basic syntactic processes" are handled by a ventral pathway that passes through STG regions identified here (Friederici, 2015). Matchin and Hickok (2020) similarly point to the ventral pathway as essential for "basic" syntactic processes. This heuristic notion of basic syntactic processing aligns well with Stanford Dependencies' avowed goal of providing a simple, surface-oriented notation for local gram-

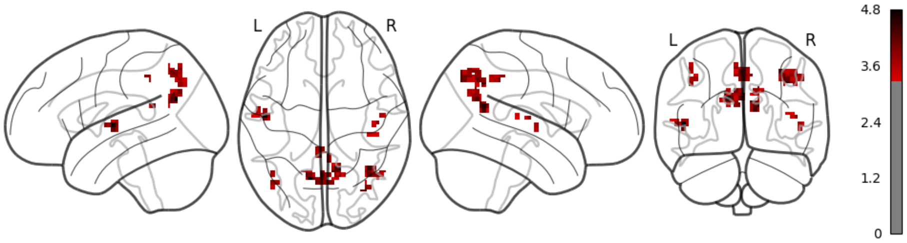

Figure 7: English Brain z-maps showing the significant clusters (p < .001 uncorrected; cluster threshold = 15 voxels) for the model comparison between syntactic surprisal at $k = 1$ vs $k = 5$. All significant clusters show greater r^2 increase for surprisal at $k = 5$ than for surprisal at $k = 1$.

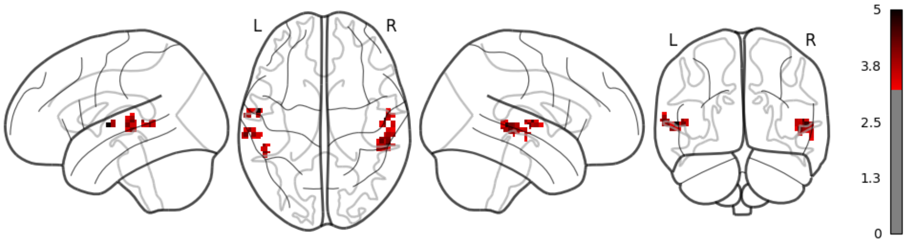

Figure 8: Chinese Brain z-maps showing the significant clusters (p < .001 uncorrected; cluster threshold = 15 voxels) for the model comparison between syntactic surprisal at $k = 1$ vs $k = 5$. All significant clusters show greater r^2 increase for surprisal at $k = 5$ than for surprisal at $k = 1$.

matical relations.

The present results do not identify a significant difference in r^2 increase in Broca's area (left IFG). Some increase in r^2 can be seen, however, in middle and IFG in the individual model maps shown in the Appendix. The absence of a significant difference, however makes sense in light of two considerations. The first is the idea that left IFG subserves reanalysis in the face of misinterpretation (Novick et al., 2005). The second is the literary style of the *The Little Prince*. The stimulus text is dissimilar to the sorts of garden-path materials psycholinguists have used (see e.g. Sedivy 2019, chapter 9 or Fernández and Cairns 2010, chapter 7). It seems likely that any garden-pathing was mild, remaining below the level of conscious awareness. This distinction between misinterpretations that rise to the level of conscious awareness and those that do not could help reconcile these results with earlier studies showing IFG activation in response to garden path stimuli (e.g. Mason et al., 2003; Rodd et al., 2010). For instance, Jäger et al. (2015) suggests that Chinese relative clauses involve as many as four temporary ambiguities. To our knowledge none of these reach the level of conscious awareness in everyday comprehension. Yet in a naturalistic fMRI study, Dunagan et al. (2022) observe activation in anterior and middle STG that is specific to Chinese object-extracted relative clauses. This contrast was not reliable in a statistical comparison between English and Chinese. Further work is needed to chart the space between unproblematic ambiguities (Lewis, 1993, §2.4.2) and conscious misinterpretations engendered by syntactic ambiguity. These STG results include activation in the primary auditory cortex, which could suggest, in line with the sensory hypothesis (Dikker et al., 2009), that expectations based on earlier syntactic processing first impact regions involved in low-level sensory processing. This possibility could be examined in a followup fMRI study with written or signed materials, along the lines of Henderson et al. (2016).

7 Conclusion and Future Work

The main conclusion is that human parsing is multipath. This follows from observing greater r^2 increase for multipath surprisal than single-path surprisal in fMRI data. Even with a very bottom-up strategy, such as the arc-hybrid system used here, it seems that more than one path must be pursued in order to best-align with humans' word-by-word effort profile. This conclusion is consistent with disjunctive representations of choices such as PP attachment (Kitaev et al., 2022, §5.5).

The multipath interpretation that we offer here can be confirmed or refuted with other linked linguistic and neuro-cognitive databases – for instance, in different genres or languages. In addition, using brain data with a higher temporal resolution, such as MEG, may provide a benefit given the temporary nature of the syntactic ambiguities included in our model. Although we take the primary finding to be one of commonality across STG regions in both languages, English listeners did uniquely show an additional effect of multipath parsing in parietal regions. Activation in this area has been correlated with measures of human memory such as digit span (Meyer et al., 2012). Individuals' memory span may modulate the number of paths that they pursue during comprehension (Vos et al., 2001; Prat and Just, 2010). Future work should address the interaction between disambiguation and memory (e.g. Campanelli et al., 2018).

Limitations

The goal of correlating parser states with fMRI data is limited by the particularities of the parsing system – namely the arc-hybrid transition system and the limited size and content of the training data (which is dissimilar in genre to the audiobook text we study). In addition, the English results in particular are the result of careful model selection to limit "training away" the syntactic ambiguity we would like to measure. This may indicate that more formalized external limitations should be applied to modern high-performing parsers should they be used to model human sentence processing.

Ethics

Language models pose risks when used outside of their intended scope. We use BLOOM, which is publicly available under the Responsible AI License (RAIL) (Le Scao et al., 2022a). Our scientific enquiry falls within the intended use of public research on LLMs. We also use a publicly available fMRI dataset (Li et al., 2022), which is available under a CCO license. This dataset was anonymized to remove identifying facial features before publication.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number 1903783. We thank Laura Rimell and Michael Covington for valuable discussion, along with Christophe Pallier for developing the cross-validated r^2 code used here. We are grateful to many collaborators for assistance with the fMRI data collection. Among them are Shohini Bhattasali, Jixing Li, Wen-Ming Luh and Nathan Spreng.

References

Ina Bornkessel-Schlesewsky and Matthias Schlesewsky. 2015. The Argument Dependency Model. In (Hickok et al., 2015), chapter 30.

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.

Adrian Brasoveanu and Jakub Dotlačil. 2020. *Computational Cognitive Modeling and Linguistic Theory*. Language, Cognition, and Mind. Springer.

Jonathan R. Brennan, Chris Dyer, Adhiguna Kuncoro, and John T. Hale. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:1074–1079.

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 Lang.*, 157-158:81–94.

Joan Bresnan and Ronald M. Kaplan. 1982. Introduction: Grammars as mental representations of language. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages xvii,lii. MIT Press, Cambridge, MA.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Matthias Buch-Kromann. 2004. Optimality parsing and local cost functions in Discontinuous Grammar. *Electronic Notes in Theoretical Computer Science*, 53:163–179. Proceedings of the joint meeting of the 6th Conference on Formal Grammar and the 7th Conference on Mathematics of Language.

Jan Buys and Phil Blunsom. 2018. Neural syntactic generative models with exact marginalization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 942–952.

Luca Campanelli, Julie A. Van Dyke, and Klara Marton. 2018. The modulatory effect of expectations on memory retrieval during sentence comprehension. In Timothy T. Rogers, Marina Rau, Xiaojin Zhu, and Charles W. Kalish, editors, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 1434–1439. Cognitive Science Society, Austin, Texas.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Nature Communications Biology*, 5(1):134.

Robert W Cox. 1996. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173.

Benoit Crabbé, Murielle Fabre, and Christophe Pallier. 2019. Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1150–1160, Hong Kong, China. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Suzanne Dikker, Hugh Rabagliati, and Liina Pylkkänen. 2009. Sensitivity to syntax in visual cortex. *Cognition*, 110(3):293–321.

Donald Dunagan, Miloš Stanojević, Maximin Coavoux, Shulin Zhang, Shohini Bhattasali, Jixing Li, Jonathan Brennan, and John Hale. 2022. Neural correlates of object-extracted relative clause processing across English and Chinese. *Neurobiology of Language*, pages 1–43.

Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. 2022. Probing for incremental parse states in autoregressive language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2801–2813, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eva M. Fernández and Helen Smith Cairns. 2010. *Fundamentals of Psycholinguistics*. Wiley-Blackwell.

Janet Dean Fodor and Fernanda Ferreira, editors. 1998. *Reanalysis in sentence processing*, volume 21 of *Studies in Theoretical Psycholinguistics*. Kluwer, Dordrecht.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: a new two-stage parsing model. *Cognition*, 6:291–325.

Angela D. Friederici. 2015. The neuroanatomical pathway model of language. In (Hickok et al., 2015), chapter 29.

Karl J Friston, John Ashburner, Stefan J Kiebel, T Nichols, and William Penny. 2007. *Statistical Parametric Mapping*. Academic Press.

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

Edward Gibson and Neal J. Pearlmutter. 2000. Distinguishing serial and parallel parsing. *Journal of Psycholinguistic Research*, 29(2).

Paul G. Gorrell. 1987. *Studies of human syntactic processing: Ranked-parallel versus serial models*. Ph.D. thesis, University of Connecticut.

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

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

John T Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the North American chapter of the association for computational linguistics*.

John T Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R Brennan. 2022. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8:427–446.

John M Henderson, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, 132:293–300.

Gregory Hickok, Steven L. Small, and Steven L. Small. 2015. *Neurobiology of Language*. Elsevier, San Diego.

Jens-Max Hopf, Markus Bader, Michael Meng, and Josef Bayer. 2003. Is human sentence parsing serial or parallel?: Evidence from event-related brain potentials. *Cognitive Brain Research*, 15(2):165–177.

Lena Jäger, Zhong Chen, Qiang Li, Chien-Jer Charles Lin, and Shravan Vasishth. 2015. The subject-relative advantage in Chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, 79:97–120.

Daniel Jurafsky. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, 20(2):137–194.

Nikita Kitaev, Thomas Lu, and Dan Klein. 2022. Learned incremental representations for parsing. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3086–3095, Dublin, Ireland. Association for Computational Linguistics.

Marco Kuhlmann, Carlos Gómez-Rodríguez, and Giorgio Satta. 2011. Dynamic programming algorithms for transition -based dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 673–682, Portland, Oregon, USA. Association for Computational Linguistics.

Prantik Kundu, Souheil J Inati, Jennifer W Evans, Wen-Ming Luh, and Peter A Bandettini. 2012. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *Neuroimage*, 60(3):1759–1770.

Howard S. Kurtzman. 1984. Ambiguity resolution in the human syntactic parser: An experimental study. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 481–485, Stanford, California, USA. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022a. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint 2211.05100*.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022b. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Richard L. Lewis. 1992. Recent Developments in the NL-Soar Garden Path Theory. Technical Report CMU-CS-92-141, Carnegie Mellon University.

Richard L. Lewis. 1993. *An Architecturally-based Theory of Human Sentence Comprehension*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Richard L. Lewis. 2000a. Falsifying serial and parallel parsing models: empirical conundrums and an overlooked paradigm. *Journal of Psycholinguistic Research*, 29(2).

Richard L. Lewis. 2000b. Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. In Matthew W. Crocker, Martin Pickering, and Charles Clifton, Jr., editors, *Architectures and mechanisms for language processing*. Cambridge University Press.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific data*, 9(1):530.

Jixing Li and John Hale. 2019. Grammatical predictors for fMRI timecourses. In Robert C. Berwick and Edward P. Stabler, editors, *Minimalist Parsing*. Oxford University Press.

Alessandro Lopopolo, Antal van den Bosch, Karl-Magnus Petersson, and Roel M. Willems. 2021. Distinguishing syntactic operations in the brain: Dependency and phrase-structure parsing. *Neurobiology of Language*, 2(1):152–175.

Mitchell P. Marcus. 1980. *A theory of syntactic recognition for natural language*. MIT Press.

David Marr. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company.

Robert A Mason, Marcel Adam Just, Timothy A Keller, and Patricia A Carpenter. 2003. Ambiguity in the brain: what brain imaging reveals about the processing of syntactically ambiguous sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1319.

William Matchin and Gregory Hickok. 2020. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498.

Lars Meyer, Jonas Obleser, Alfred Anwander, and Angela D Friederici. 2012. Linking ordering in broca's area to storage in left temporo-parietal regions: the case of sentence processing. *Neuroimage*, 62(3):1987–1998.

Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the workshop on incremental parsing: Bringing engineering and cognition together*, pages 50–57.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Jared M. Novick, John C. Trueswell, and Sharon L. Thompson-Schill. 2005. Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective & Behavioral Neuroscience*, 5(3):263–281.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Subba Reddy Oota, Mounika Marreddy, Manish Gupta, and Raju Bapi. 2023. How does the brain process syntactic structure while listening? In *Findings of* *the Association for Computational Linguistics: ACL 2023*, pages 6624–6647, Toronto, Canada. Association for Computational Linguistics.

Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, and Christophe Pallier. 2023. Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax and context. *arXiv preprint arXiv:2302.14389*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Chantel S. Prat and Marcel Adam Just. 2010. Exploring the Neural Dynamics Underpinning Individual Differences in Sentence Comprehension. *Cerebral Cortex*, 21(8):1747–1760.

Aniketh Janardhan Reddy and Leila Wehbe. 2021. Can fmri reveal the representation of syntactic structure in the brain? In *Advances in Neural Information Processing Systems*, volume 34, pages 9843–9856. Curran Associates, Inc.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

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. Association for Computational Linguistics.

Jennifer M Rodd, Olivia A Longe, Billi Randall, and Lorraine K Tyler. 2010. The functional organisation of the fronto-temporal language system: evidence from syntactic and semantic ambiguity. *Neuropsychologia*, 48(5):1324–1335.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Julie Sedivy. 2019. *Language in Mind: An Introduction to Psycholinguistics*, second edition. Oxford University Press.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

Patrick Sturt. 1996. Monotonic Syntactic Processing: A Cross-linguistic Study of Attachment and Reanalysis. *Language and Cognitive Processes*, 11(5):449–494.

Sandra H. Vos, Thomas C. Gunter, Herbert Schriefers, and Angela D. Friederici. 2001. Syntactic parsing and working memory: The effects of syntactic complexity, reading span, and concurrent load. *Language and Cognitive Processes*, 16(1):65–103.

A Appendix

A.1 Generative Models

As noted in Section 3.2, we use BLOOM's tokenizer to encode all tokens. Since some of these tokens are not found in the training data, and thus are unknown to the parser's next word prediction classifier, this prevents our parser from being strictly generative.

As a comparison, we train additional models which are truly generative by replacing tokens not seen in the training data with unknown tokens in the encoder input, in the same way as the word prediction classes are defined. We report the accuracy for this *generative* approach alongside the previously defined *full input* model in Table 3. This distinction is also explained in Figure 9 for an example prefix string.

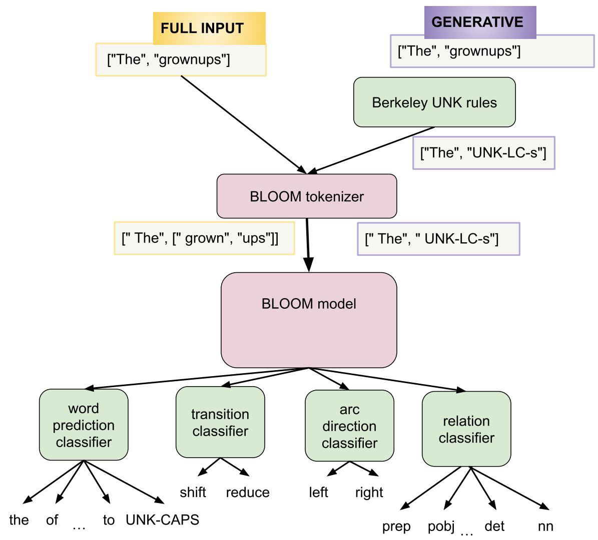

Figure 9: Parser input is pre-processed according to the given method: *Full Input* indicates no pre-processing apart from adding a leading space to all pretokenized text from the corpus. *Generative* indicates tokens are preprocessed by replacing tokens not seen more than once in the training data with unknown tokens according to the rules of the Berkeley Parser. The BLOOM model and the system of classifiers used to output probabilities is the same for both models. Note that the word prediction classifier predicts over the size of the training data vocabulary + Berkeley unknown tokens for both methods.

Corpus	Model	Dev		Test	
		LAS	UAS	LAS	UAS
PTB-3	Buys and Blunsom (2018)	88.66	91.19	88.54	91.01
PTB-3	English BLOOM Generative	89.27	92.00	90.53	92.68
PTB-3	English BLOOM Full Input (brain analysis)	90.26	92.71	90.32	92.62
CTB-7	Chinese BLOOM Generative	73.12	80.60	74.47	81.60
CTB-7	Chinese BLOOM Full Input (brain analysis)	77.07	83.65	74.71	81.66

Table 3: Labeled attachment score (LAS) and unlabeled attachment score (UAS) for the English PTB corpus and Chinese CTB corpus

A.2 Parser Training

The output representation from the BLOOM model has a dropout of 0.5 applied, and is then fed into a single layer feedforward neural network.

During training, gradient norms are clipped to 5.0, and the initial learning rate is 1.0, with a decay factor of 1.7 applied every epoch after the initial 6 epochs. English and Chinese models train for approximately 5.5 and 4.5 minutes, respectively, per epoch on an A100 GPU.

A.3 Intermediate Results

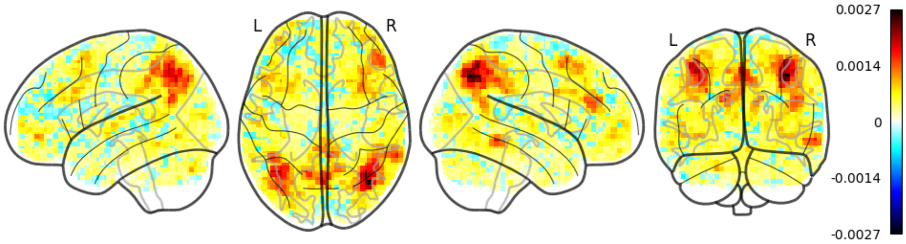

Figure 10: English Brain z-maps showing the r^2 increase for the model including syntactic surprisal at $k = 5$ compared to the model including only the following regressors: word rate, fundamental frequency (f0), word frequency, and root mean square intensity (RMS).

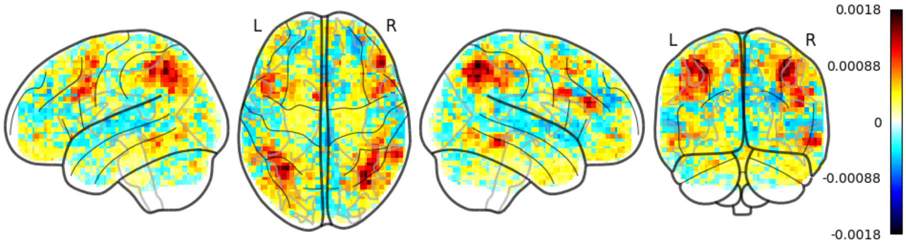

Figure 11: English Brain z-maps showing the r^2 increase for the model including syntactic surprisal at $k = 1$ compared to the model including only the following regressors: word rate, fundamental frequency (f0), word frequency, and root mean square intensity (RMS).

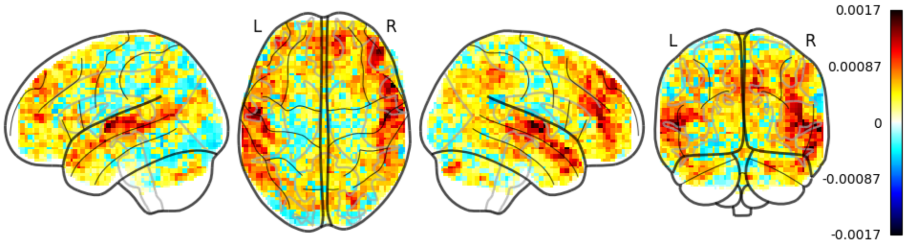

Figure 12: Chinese Brain z-maps showing the r^2 increase for the model including syntactic surprisal at $k = 5$ compared to the model including only the following regressors: word rate, fundamental frequency (f0), word frequency, and root mean square intensity (RMS).

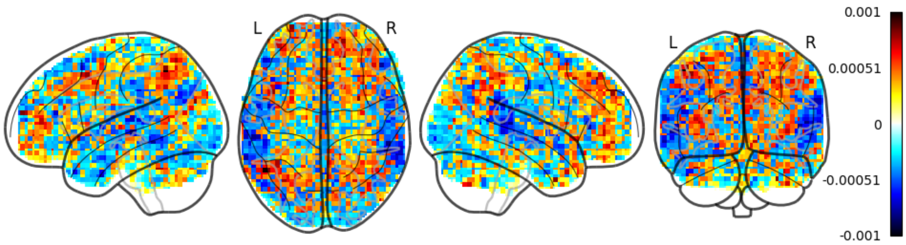

Figure 13: Chinese Brain z-maps showing the r^2 increase for the model including syntactic surprisal at $k = 1$ compared to the model including only the following regressors: word rate, fundamental frequency (f0), word frequency, and root mean square intensity (RMS).

A.4 Results

Tables 4 and 5 show the full list of significant clusters (p < .001 uncorrected; cluster threshold = 15 voxels) for the model comparison between syntactic surprisal at $k = 1$ vs $k = 5$.

Region	X	Y	Z	Peak Stat	Cluster Size (mm3)
Left-SecVisual(18), Right/Left DorsalPCC(31)	-12.0	-64.0	24.0	4.78	968
Right-VisMotor(7)	0.0	-64.0	48.0	4.02	856
Right-AngGyrus (39)	38.0	-54.0	38.0	4.42	800
Left-PreMot + SuppMot(6) / Left-PrimAuditory (41)	-48.0	-10.0	4.0	4.68	568
Right-VentPostCing(23)	14.0	-54.0	20.0	4.57	488
Right-AngGyrus (39)	36.0	-74.0	40.0	4.27	360
Left-AngGyrus (39)	-44.0	-68.0	36.0	3.77	352
Right-DorsalPCC (31)	14.0	-62.0	34.0	4.03	216
Right-DorsalPCC (31)	4.0	-50.0	44.0	3.65	192
Right-PrimAuditory (41)	48.0	-12.0	2.0	3.69	160
Right-PrimAuditory (41)	38.0	-26.0	12.0	3.39	152

Table 4: English significant clusters (p < .001 uncorrected; cluster threshold = 15 voxels) for the model comparison between syntactic surprisal at $k = 1$ vs $k = 5$

Region	X	Y	Z	Peak Stat	Cluster Size (mm3)
Right-MedTempGyrus (21) / Right-SupTempGyrus (22)	56.0	-36.0	4.0	4.68	1336
Left-SupTempGyrus (22)	-64.0	-26.0	6.0	4.32	392
Right-PrimAuditory (41)	52.0	-14.0	6.0	4.59	384
Left-PrimAuditory(41)	-56.0	-6.0	4.0	5.04	384
Left-MedTempGyrus (21)	-48.0	-42.0	4.0	4.25	320
Left-SupTempGyrus (22)	-58.0	-26.0	0.0	4.06	288

Table 5: Chinese significant clusters (p < .001 uncorrected; cluster threshold = 15 voxels) for the model comparison between syntactic surprisal at $k = 1$ vs $k = 5$