Beyond The Surface: An investigation into the role of hierarchical structure on human sentence processing using LLMs

by

Michael Wolfman

(Under the Direction of John Hale and Vera Lee-Schoenfeld)

Abstract

Interest in large language models (LLM) has exploded lately, driving speculation about their implications as models of human language processing. Language models have proven useful in isolating language-processing functions in the brain; however, debate continues regarding whether or not the best characterization of these language-processing functions employs hierarchical syntax. This study investigates this question by comparing two language models with the same underlying Transformer-XL architecture: one informed by hierarchical syntax (Transformer Grammar) and one not (Transformer-XL). Coupling these language models with human data previously collected via fMRI, results re-affirm the role of hierarchical structure in linguistic processing, implicating Broca's Area, the right Middle Temporal Gyrus, the left Temporal Pole, and the right Pre-Frontal Cortex in these aspects of processing.

Index words: [Computational Linguistics, Neurolinguistics, Cognitive Science, Syntax, Parsing, Transformers, fMRI]

Beyond The Surface: An investigation into the role of hierarchical structure on human sentence processing using LLMs

by

Michael Wolfman

B.A.Linguistics, University of Georgia, 2022

B.A. Classics, University of Georgia, 2022

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Arts

Athens, Georgia

2024

Beyond The Surface: An investigation into the role of hierarchical structure on human sentence processing using LLMs

by

Michael Wolfman

Major Professor:	John Hale
	Vera Lee-Schoenfeld
Committee:	Dustin Chacón
	Keith Langston

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2024

Dedication

To Carrie Stockard, Maria Kepler, and Mike Scirocco – who cultivated the love of language that ultimately brought me here.

Acknowledgments

First and foremost, I would like to thank Dr. Hale and Dr. Lee-Schoenfeld, not only for their guidance on this project, but for their mentorship throughout my undergraduate and graduate studies at the University of Georgia. Next, I would like to thank my committee members, Dr. Chacón and Dr. Langston, for their counsel on this project. Thank you as well to Jean Costa-Silva for his mentorship on writing a thesis and to Donnie Dunagan for sharing his invaluable knowledge with me throughout this project. Finally, I would like to thank Rose Sebaugh for the countless conversations she endured during which I was stressing about this project.

Contents

Acknowledgments **v**

List of Figures **viii**

List of Tables **x**

1 Introduction **1**

1.1 Background 1

1.2 Present Study 7

1.3 Contributions 8

1.4 Thesis Organization 9

2 Language as Hierarchical **10**

2.1 Hierarchical Structure 11

2.2 Chomskyan Syntax: An Overview 14

2.3 From Capacity to Processing 28

3 The Role of Hierarchical Structure in Processing and its Neurobiology **30**

3.1 Surprisal as a Linking Hypothesis 31

3.2 Investigating Hierarchy 32

4 Language as Sequential **35**

4.1 Rejection of Behaviorism and Middle-Ground Associationism 35

4.2 Late Assignment of Syntax Theory 38

4.3 Connectionism Revitalized 40

5 Sequential Processing and its Neurobiology 42

6 Language Models 45

6.1 Deep Learning 45

6.2 Backpropagation and Word Embeddings 46

6.3 Transformers 47

6.4 Language Models in the Present Study 51

7 Materials and Methods 55

7.1 Materials 55

7.2 fMRI Data 61

7.3 Methods 63

8 Results and Discussion 66

8.1 Results 66

8.2 Discussion 66

9 Conclusion 74

Appendices 76

A 76

Bibliography 78

List of Figures

1.1 A example of a neural net (Buckner & Garson, 2019) 4

2.1 Sequential, Finite State Language Device (Chomsky, 1957, pg. 19) 12

2.2 Phrase structure tree with PP attached to VP 13

2.3 Phrase structure tree with PP attached to NP 14

2.4 Phrase structure tree for *The dog barks* 18

2.5 Phrase structure tree for NP *The dog in the crate on the porch* 19

2.6 Y-Model of Syntax 20

2.7 Left-Headed X-bar Phrase Structure 24

2.8 Right-Headed X-bar Phrase Structure 25

2.9 Bare Phrase Structure for *On the back porch* 27

3.1 Surprisal of Prefix-String Probabilities 31

6.1 A example of a neural net (Buckner & Garson, 2019) 47

6.2 (Left) Scaled Dot-Product Attention. (Right) Multi-Head Attention consists of several attention layers running in parallel. (Vaswani et al., 2017, p. 4) 48

6.3 Example of a Transformer Block (Jurafsky & Martin, 2023, Ch. 10) 50

6.4 An example of a string x and tree y, which are modeled by an action sequence that generates a linearized tree representation of (x,y) (Choe & Charniak, 2016) 52

6.5 An example of the TG attention mask and operations (Sartran et al., 2022, Figure 2) . . 54

7.1 Surprisal of Prefix-String Probabilities 56

7.2 Illustration of equivalence in calculating surprisal via manual calculation and LogSumExp calculation 62

8.1 z-maps (with no thresholding) showing the r^2 increase for the syntax-less model compared to the base model. 67

8.2 z-maps (with no thresholding) showing the r^2 increase for the syntax-informed model compared to the base model 67

8.3 Results (z-valued) of paired T-test between syntax-informed and syntax-less cross-validated r^2, thresholded with an expected false discovery rate (FDR) < 0.05 and a cluster threshold of 50 voxels. 68

8.4 An example of indirect speech that is attested in the stimulus text (*The Little Prince*) used in the fMRI study. 69

A.1 TG Surprisal for each token in *The Little Prince* 77

A.2 TXL Surprisal for each token in *The Little Prince* 77

List of Tables

7.1 Log-probabilities from TXL for the first sentence of LPP 57

7.2 Log-probabilities from TG for part of the first sentence of LPP 59

7.3 GLM Model Predictors 65

8.1 Results of paired T-test between syntax-informed and syntax-less cross-validated r^2, thresholded with an expected false discovery rate (FDR) < 0.05 and a cluster threshold of 50 voxels. 67

Chapter 1

Introduction

1.1 Background

Human beings are able to produce and comprehend natural language extremely quickly. The average rate of speech in normal conversation is 120-150 words per minute (The National Center for Voice and Speech),[1] or 2-2.5 words per second. In a matter of milliseconds, a speaker of a given language is able to (subconsciously) select an idea, select the words to express that idea, compose those words into a sentence, and produce the sounds that encode those words. Similarly, comprehenders can hear sounds entering their ears, decode those sounds into units of speech, map those units onto words, and those words to meaning, all in a matter of a couple hundred milliseconds. How people are able to do this has driven work in cognitive science and other sub-disciplines of cognitive science. Three critical inter-related questions central to this research endeavor are:

1. What is the best way to conceptualize natural language?

2. How do humans comprehend natural language?

3. What neurobiological components subserve language comprehension?

This thesis will address these questions by reviewing different conceptualizations of language, investigating how these conceptualizations come to bear during language comprehension, and probing the regions of the brain that subserve this comprehension.

[1] This number is specifically the average rate of speech for English speakers in the United States.

To answer these questions requires situating language in the realm of cognition as whole. A major underlying assumption of cognitive science is that cognition is computational[2] (Gallistel & King, 2009; Newell, 1990; Pylyshyn, 1984). The precise kind of computation is still debated. One view – championed by Pylyshyn, Newell, Gallistel, King, and others – is based on symbols. Rather than assessing the causes of behavior as external to the mind, the causes of behavior are internal mental representations of these external realities, called symbols (Pylyshyn, 1984). They argue that direct physical signals are not sufficient information for the brain to act effectively, rather internal representations are needed. These representations act as mappings of physical reality to symbols in the brain (Gallistel & King, 2009). These symbols are the entities in cognition that carry information, information that is abstract but related to the observable world. Operations are then carried out over these symbols, manipulating or combining them to yield some further representation of this information. The formal operations that can be carried out over symbols are generally referred to as the syntax of that domain. Whether discussing physical movement, vision, or language, each domain has its own symbols and its own syntax that the brain uses to guide its actions (Gallistel & King, 2009). This view of cognition is broadly referred to as the Physical Symbol System Hypothesis (PSSH; Newell & Simon, 1976) and falls under the purview of the broader theory of cognitivism, which argues for thought as the controller of human behavior (Mandler, 2002; Newell, 1990).

Other schools of thought, such as Behaviorism and Connectionism, have rejected this symbol-centric account of human thinking. In its early form, behaviorism, a historically earlier school of thought than cognitivism, argued that human behavior and learning occurred due to stimulus-response pairs, in which a response is conditioned on a given stimulus. Learning, and in turn behavior, arise as an individual repeatedly experiences a given stimulus and its response. In contrast to the cognitivist view previously discussed, a behaviorist view of learning and human behavior places all emphasis on mind-external factors as the direct causes of individuals' behavior, rather than their internal mental state prompting the behavior (Moore, 2017; Skinner, 1963). Thinking then is a type of behavior that arises in response to other behavior, contrasting the cognitivist view that thought drives behavior.

Connectionism, like the PSSH and contrasting behaviorism, is a cognitivist theory that places the mind at the center of behavior; however, unlike a symbolic system, connectionism attempts to explain

[2] This is referred to as the Computational Theory of Mind (as developed early by Fodor, 1975; Putnam, 1967).

cognitive capabilities using artificial neural networks. This modeling is inspired by the neuron in the brain and how these neurons combine to carry out higher functions than each neuron could carry out on its own. In this view, researchers seek to better understand cognitive processes based on the behavior of artificial neural networks. These artificial neural networks form the foundation of modern deep learning models. This study employs two such models: Transformer-XL (Dai et al., 2019), which uses word-level sequential information in language modeling, and Transformer Grammar (Sartran et al., 2022), which utilizes abstract hierarchical structure in language modeling (see chapter 6 for more on the task of language modeling).

Understanding the core mechanisms of early artificial networks is critical in understanding how these language models can be used to characterize human language processing.[3] Artificial neural networks take an input. This input has certain activation values, corresponding to abstract feature(s) in the network, and this activation value feeds forward to all nodes in the first "hidden" layer to which the input node is connected. These activation values feed-forward through as many hidden layers as are present in the network, ending up in an activation function that maps the combination of activation values from all hidden nodes to a specific output (Buckner & Garson, 2019). The strength of connections between two nodes are referred to as "weights". The values of weights fluctuate from negative values, inhibiting a connection, to positive values, activating a connection. This simple neural network is called a feedforward network, in which the connections between each layer only operate moving forward (Elman, 1990). An illustration is given in Figure 1.1. Critically, the weights between units are adjusted during a neural network's training via a method called "backpropagation" (Rumelhart, Hinton, & Williams, 1986). During training, when the output of the neural network does not match the expected output, the "error" backwardly propagates through the hidden units, adjusting their weights so that the neural network is more likely to correctly identify the output next time (for more on backpropagation, see chapter 5).

There are a number of similarities and differences between the theories outlined above, such that they can be compared in multiple ways; however, the critical distinction for the present work is as follows: (i) the PSSH views cognitive processes as combinations and manipulations of symbols, (ii) connectionism

[3] The term linguistic or language "processing" can refer broadly to multiple mental processes associated with producing and understanding language. In this thesis, when the term processing is used, it is to refer specifically to the mental processes associated with language comprehension.

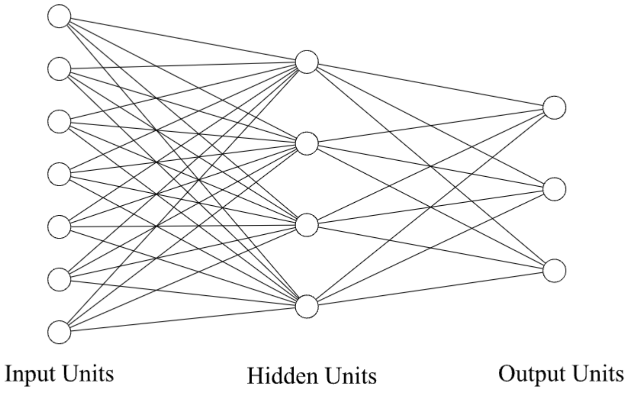


Figure 1.1: A example of a neural net (Buckner & Garson, 2019)

and behaviorism argue that cognitive processes occur as a result of learned associations.[4] In behaviorism, a stimulus and response are the learned associations, while associations in connectionism are represented via weights representing the strengths of connections between layers. In both cases, these associations are learned from a series of inputs that inform the strength of said associations, thus they view learning as an inductive process, in which exposure to information and the associations learned from this exposure result in knowledge.

Turning back to language, these theories draw distinct conclusions about the capacity for language (question 1 above). In a symbolic view, language, like other cognitive processes, are computational with a set of domain-specific physical symbols that are combined and manipulated to yield an output, such as the meaning of an utterance. In the associationist view (under which I argue connectionism and behaviorism fall), language is learned via repetition of patterns, of words co-occurring together in a given order. The implication of this view is that language is sequential, as opposed to hierarchical as discussed just below.[5]

[4] There are a variety of compromise positions between these associationist and symbolic views, notably the "Integrated Connectionist-Symbolist Architecture" as developed by Smolensky et al. (1993). This approach integrates connectionism and symbolism, arguing for a single computational system with two levels of computation. Patterns of activity distributed over connectionist units make up the lower level, while structures of symbols and their manipulation describe the same computational system at the higher level. They argue that this system allows for a "precise treatment of a complex set of interactions of semantic and syntactic constraints in a single language" (p. 381).

[5] It is important to note that associationist views of language necessarily imply that language is sequential, given that language is comprised of associations between words; however, there are also symbolic systems which treat language as sequential, such as those discussed in chapter 4.

The associations that underlie knowledge of language is acquired via surface-level statistical patterns: the ordering of the words and how they appear together.

In contrast to this sequential view, one of the most prominent symbolic views of language argues for a hierarchical system in which symbols combine with other symbols in a hierarchical manner to create larger constituents (Chomsky, 1956, 1957). The combinations of these symbols take place via explicit 'rules.' These rules correspond to operations over the symbols, which describe, among other things, how these symbols can combine. This set of rules and representations constitute the **syntax** of language. Under this view, syntax, then, is a critical aspect of the capacity for language.

This capacity for language – how language is learned and how this knowledge is stored – bears on how language is processed, implementing given language mechanisms to comprehend language as it is heard or seen (question 2 above). Thus, if the capacity for language is sequential, based on associations or statistical patterns, then too will these associations and statistical patterns play a role in comprehension. Similarly, if language is a hierarchical physical symbol system, like in the view of Chomsky (1957), then this hierarchical organization of symbols will play some role in comprehension.

While the capacity for language and comprehending language are largely abstract, the specific mechanisms that underlie them are thought to be rooted in specific neurobiological properties (Chomsky, 1957; Lennenberg, 1967). This idea has motivated an investigation of what the mapping of these linguistic abstractions to neural properties may look like (question 3 above).

One way to test both the nature of human language comprehension and the corresponding neurobiological properties is by leveraging computational modeling. This type of research seeks to relate language models to human behavior and neural signals (for a review of this type of computational modeling, see Hale et al., 2022). Language models are simply an assignment of probability to strings (Jelinek et al., 1992). The equation for modeling probability for a given word in a string, based on the preceding context, is:

$$P(W_n = w_n | W_1 = w_1, ...W_{n-1} = w_{n-1})$$

The above equation can be read as the probability that the nth word is w_n if the prior words in the string were w_1 through w_{n-1}. Surprisal (Hale, 2001; Levy, 2008) is a complexity metric that has been shown to correspond with processing natural language (see e.g., Boston et al., 2008; Hale et al., 2018; Henderson et al., 2016; Willems et al., 2016). Surprisal acts as a linking hypothesis that relates theorized mechanisms of linguistic processing to empirical data – such as EEG scalp voltages or fMRI Blood Oxygenated Level Dependent (BOLD) – to glean information about which regions of the brain subserve specific aspects of sentence processing (for details on how exactly surprisal is formulated in regard to language and how this formulation is implemented using TXL/TG, see chapter 3 and chapter 7 respectively).

If surprisal values from a given language model align well with the observed empirical measures, then "the observed data support the conjunction of [surprisal] and whatever defined the probabilities in the first place" (Hale, 2016, p. 398). The TXL language model used in the present study only considers prior words as the context for calculating the probability of word$_n$, whereas the TG language model used in the present study additionally incorporates abstract hierarchical structure into this context by including explicit symbolic linguistic representations in its probability distribution. TG-derived surprisal values then come from this probability distribution with explicit symbolic representations. In contrast, TXL-derived surprisal values come from a probability distribution that does not incorporate explicit hierarchical structure. If TG-derived surprisal values explain blood oxygen levels in the brain better than TXL-derived surprisal values, this would support the role of hierarchical structure during language comprehension.

Leveraging surprisal and computational modeling to isolate language processing functions in the brain has yielded much success (see e.g., Hale et al., 2022, for a recent review); however, whether or not language comprehension is best captured by a sequential view or hierarchical view remains in contention (e.g., Contreras Kallens et al., 2023; Fedorenko et al., 2020; Goldstein et al., 2022a; Tuckute et al., 2024). This thesis contributes to the debate at hand by investigating (i) how well hierarchical structure, instantiated via the novel Transformer Grammar (Sartran et al., 2022), and surface-level sequential word ordering, instantiated via a Transformer-XL (TXL), correlate with the BOLD signal collected during fMRI scans, and (ii) what specific regions of the brain are sensitive to to each model type (hierarchical vs. sequential). Specifically, this thesis argues that hierarchical syntax is, in fact, useful in characterizing language comprehension and that there are specific regions of the brain that subserve this hierarchical computation.

1.2 Present Study

1.2.1 Introduction

The present study investigates the role of hierarchical structure in incremental processing of natural language using a novel instantiation of a language model with syntactic knowledge, known as Transformer Grammars (TGs; Sartran et al., 2022). Surprisal (Hale, 2001; Levy, 2008) is a word-by-word complexity metric which can be calculated by language models, like TG, and which has previously been shown to correlate with human language processing difficulty (see e.g., Brennan et al., 2020; Hale et al., 2018; Henderson et al., 2016). This study compares syntactically-informed surprisal, derived from a TG, to syntax-less surprisal, derived from a TXL (Dai et al., 2019), in the context of fMRI data recorded as participants listen to a storybook. It is important to note here that the term 'syntax-less' refers to the fact that the TXL model is unconstrained in whether or not it learns some sort of syntactic representation(s) during its training. In contrast, the TG model has a compositional bias imposed that forces it to learn explicit symbolic hierarchical representations. In other words, 'having syntax' is a matter of degree: the syntax-less model *may* learn some implicit syntactic representation, while the syntax-informed model is forced to learn explicit symbolic hierarchy.

1.2.2 Materials

The fMRI data analyzed was the English component of the Little Prince Datasets (Li et al., 2022). Participants (N=49) were scanned while they engaged in the naturalistic task of listening to an audiobook recording of a children's story, The Little Prince. Word-by-word surprisal values were calculated using a TG model and TXL model, each trained on the BLLIP-LG data set, as prepared by Hu et al. (2020).

1.2.3 Methods

The neural correlates of the syntactically-informed surprisal metric were probed via an r^2 analysis (Crabbé et al., 2019) using generalized linear models (GLM). For more information on the analysis, see chapter 7.

1.2.4 Results

The results of the r^2 analysis are as follows: the syntax-informed model performed above-and-beyond the syntax-less model in goodness-of-fit (r^2 values) to the measured BOLD signal timecourse in the right medial temporal gyrus (rMTG; BA21), the pars opercularis (Broca's Area; BA44), the left temporal pole (BA38) and the right pre-frontal cortex (rPFC; BA10). There were no regions in which the syntax-less model outperformed the syntax-informed model.

1.2.5 Discussion

These results support two conclusions: 1) BA44 and BA38 play a major role in structure building, confirming numerous previous results (in line with large scale brain models of language, e.g., Friederici, 2017, Hagoort, 2013) and 2) right hemisphere brain regions are recruited during language comprehension, perhaps to assist in difficult complex structure building such as quotation. Taken together, these results affirm that syntactic structure is part of the best account of processing natural language, as well as the brain regions that subserve this aspect of processing. These findings fail to support a sequential view, suggesting contrariwise that language is organized with multiple levels of abstraction, including syntactic structure, which guide natural language comprehension and directly correspond to neurobiological properties.

1.3 Contributions

The contributions of this thesis are fourfold.

1) It re-affirms the role of hierarchical structure in the processing of natural language and the brain regions implicated in this processing. More specifically, it confirms the role of BA44 in structure building, and posits the recruitment of right homologues to left language network (e.g., BA22) in difficult syntactic processing.

2) It employs a more methodologically sound comparison of syntax-informed and syntax-less language models. In the past, comparisons between sequential and hierarchical models of language were carried out using models with different underlying architectures (e.g., Brennan and Hale (2019) compares a trigram model/simple recurrent neural network against a phrase-structure context-free grammar). In contrast,

the only difference between TG and TXL is the additional attention mask (a concept introduced in chapter 6) in TG. This additional attention mask biases the Transformer towards a more hierarchical mode of operation. In short, the architectures of TG and TXL only differ in whether or not 'syntax' is turned on. This comparison, then, offers more straightforward and convincing evidence for the role of syntax in language comprehension than previous studies.

3) It offers a possible explanation of the role of rMTG in processing, notably connecting ideas about 'Quotation' from the Philosophy of Language and semantics/pragmatics. Quotation results in complex metalinguistic representations that (as a form of indirect speech) might require additional work in the brain to process, perhaps done in rMTG and rPFC.

4) It provides a comparison illustrating the equivalence of two methods for calculating surprisal: one using conditional probabilities and one that uses LogSumExp, a helper function in Python (Virtanen et al., 2020), to calculate surprisal. The illustration of the equivalence of these two methods is a useful baseline for researchers, as language models increasingly output logarithmic probabilities rather than conditional probabilities.

1.4 Thesis Organization

This thesis is organized as follows: Chapter 2 discusses a capacity for language with hierarchical representation and Chapter 3 discusses how this capacity for language relates to comprehension and the brain. Chapter 4 and Chapter 5 discuss the same for a sequential view of language. Chapter 5 gives an overview of deep learning techniques in language modeling and the language models in the present study. Chapter 7 reviews the materials and methods used in the present study, including an illustration of two equivalent methods for deriving surprisal. Chapter 8 explores the results of the study and discusses possible explanations for said results. Finally, Chapter 9 concludes.

Chapter 2

Language as Hierarchical

If we conceptualize the human brain as an information-processing device (Marr, 1982), then natural language is one of the many types of information that the brain must process. Understanding natural language is a complex process involving multiple interconnected levels, such as sound, structure, and meaning. For example, when someone hears the sentence "I love dogs," they must process the incoming sounds, link the combinations of those sounds to individual phonemes, those phonemes to the morphemes *I*, *love*, *dog*, and *-s* (denoting plurality). Those morphemes combine to create words, which combine to create a meaningful English sentence. A natural follow-up question regards what the exact nature of these processes look like. Are meaningful sentences derived from the surface-level occurrence patterns of words? Or are they combined in some way different from their surface appearance? If so, what is the best characterization of those combinatory operations and their resulting representations?

The vein of research presented in this chapter, referred to as generative syntax, investigates questions like those posed above, with the goal of better understanding the human capacity for language. To this end, generative syntax treats an individual's capacity for language as a hierarchical symbolic system (Chomsky, 1957) and is primarily focused on explicating what this system looks like by defining the rules and representations of said system. The multiple iterations of generative syntax presented in this chapter each represent different instantiations of what the syntax of the language system could look like. The goal of this chapter is to provide the reader with an overview of the history of generative syntax, describing what the capacity for language could look like and, more specifically, what syntactic structure and operations could look like. Subsequent chapters turn to how these structures and operations may come into play during linguistic processing [1].

2.1 Hierarchical Structure

Remember that the system discussed here, being symbolic, assumes that cognition is computational in nature, meaning that "to think is to manipulate symbols in a particular manner" (Townsend & Bever, 2001, p. 1). These symbols are mental representations. Operations are then carried out over these symbols, manipulating or combining them to yield some result. These operations and the representations produced by applying them are the 'syntax' of that domain. The syntax of language then describes the structure and operations of language. More specifically, the syntax of language describes the structure and operations of the cognitive 'device' that is able to produce the set of grammatical sentences of a given language, and only those sentences (Chomsky, 1957, Ch. 3). In other words, the language 'device', which we will refer to as a **grammar**, is one that produces all grammatical sentences in a language, but does not produce any ungrammatical sentences.

A sentence is an utterance of finite[2] length composed of a sequence of representations of sound, called phonemes; however, attempting to model grammatical sentences based only on sequences of sounds would result in a grammar ill-fit for the task of describing something as complex as language. Rather, as hinted at above, the description of language recognizes "levels of representation," such that there are higher levels of representation above sound, like morphemes, the basic unit of meaning (Chomsky, 1957). Altogether then, language is a combination of structures at varying levels of representation. Generative syntax, and the current work, focus on the syntax of higher levels of representation, notably words and sentences.

Proponents of generative syntax argue that language is too complex a system to be described simply by the order in which these words appear, whether this system is symbolic or not. For example, Chomsky

[1] When referring to syntax, the terms 'rules' and 'operations' are used synonymously throughout this chapter to refer to the finite set of symbols and how they are manipulated to yield only the grammatical sentences in a language. Similarly, the terms 'structure' and 'representations' are used synonymously to refer to the output of the manipulations over symbols. An example of rules/operations is given in (3), and an example of structure/representation is given in Figure 2.3.

[2] Note that a sentence could, in theory, could be infinitely long; however, in reality they are finite due to memory and time constraints.

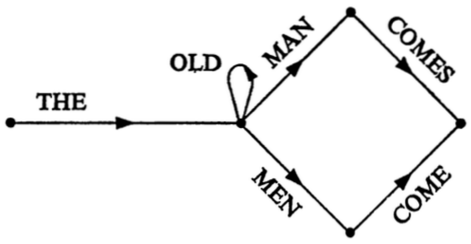


Figure 2.1: Sequential, Finite State Language Device (Chomsky, 1957, pg. 19)

(1957) argues against a sequential view of language in which "a speaker begins in the initial state, produces the first word of the sentence, thereby switching into a second state which limits the choice of the second word, third word, etc." (pg. 20), as illustrated by the machine in Figure 2.1. In this grammar, the ability to move from one state to another is indicated by an arrow from the prior state to the subsequent state. This grammar is able to produce the grammatical sentences *the old man comes*, *the old men come*, *the old old men come*, while not producing ungrammatical sentences like *the old man come* or *the old men comes*.

Chomsky argues that a sequential, finite grammar, such as this is able to capture some elements of language, but that it is unable to fully capture the facts and complexities of language. For example, consider the sentences below from Chomsky (1957, pg. 22).

1. If S_1, then S_2

2. Either S_3 or S_4

3. The man who said that S_5, is arriving today.

In (1), the word *or* cannot replace the word *then*, similarly in (2) the word *either* cannot be replaced by the word *then*. There is a dependency between the words before S_1 (*if*) and after (*then*) in (1), as well as before S_3 (*either*) and after (*or*) in (2). One can also combine these structures to yield something like (4), which nests (2) inside of (1).

1. If, either S_3 or S_4, then S_2

This nesting gives us a sequence of $a + b + S_3, + c + S_4 + d + S_2$, where there is a dependency between a and c, as well as between b and d. These dependencies are interrupted sequentially by the intervening S's.

A sequential, finite-state grammar is unable to capture these dependencies, as they are separated by the words making up the intervening sentences. Chomsky, from this, concludes that a sequential, finite-state grammar is ill-fit to describe human language, and in-turn proposes a finite hierarchical grammar instead. Despite Chomsky's criticisms, a number of theories push back against his reasoning against a sequential view of language as an oversimplification. These theories are reviewed in the next chapter.

Additional support for a hierarchical grammar comes from how it handles ambiguity in language. For example, the sentence *5he man saw a hiker with a camera*, can be read in one of two ways: (i) the man saw a hiker while looking through a camera or (ii) the man saw a hiker who was holding a camera. These alternate meanings of the sentence can be captured by the two different phrase structure trees in Figure 2.2 and Figure 2.3 respectively.[3]

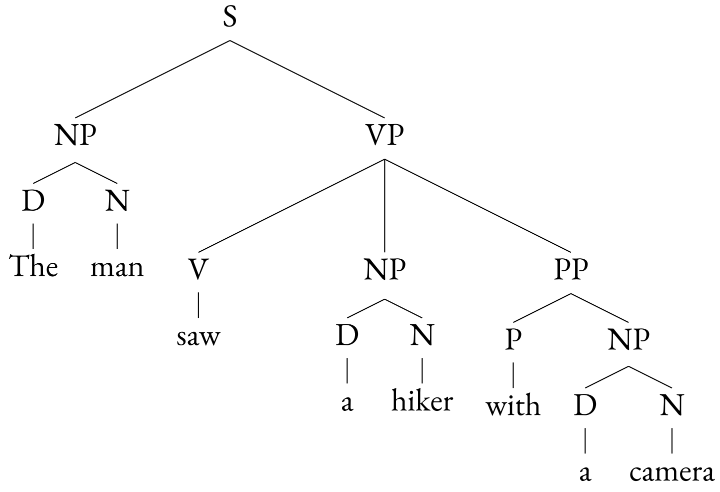


Figure 2.2: Phrase structure tree with PP attached to VP

In Figure 2.2, the prepositional phrase (PP) *with a camera* is attached to the verb phrase (VP), indicating that the PP describes the verb *saw*. This structural configuration gives the reading such that the man saw a hiker through his camera. In contrast, the prepositional phrase attached to the noun phrase (NP) *a hiker* in Figure 2.3. This attachment of the PP to the NP gives the reading such that the man saw a hiker who was carrying or holding a camera.

Phrase structure trees such as that in Figure 2.3 offer a 2-D graph-theoretic visualization for a possible mental representation of a sentence. These mental representations illustrate Chomsky's hierarchical view

[3] The examples here, and those throughout this section, abstract away from the representation of tense and auxiliaries as heads of their own phrases.

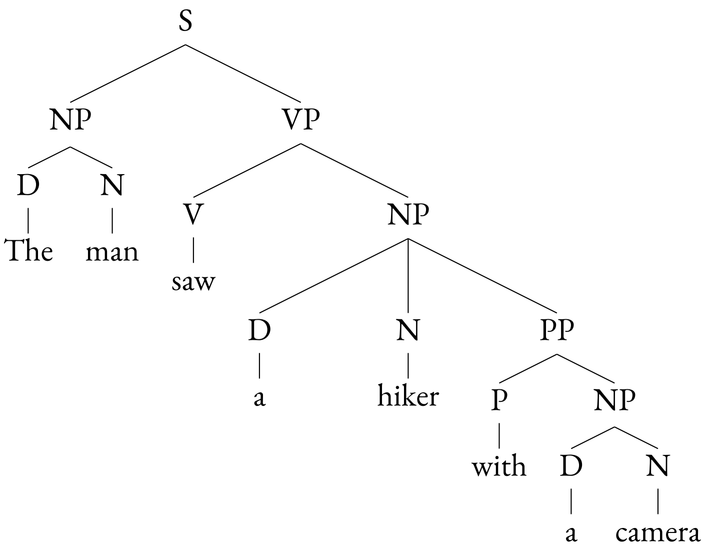


Figure 2.3: Phrase structure tree with PP attached to NP

of language, signifying the relationships between constituents that are not represented by a sequential, finite-state grammar, such as the different constituents to which the PP could attach in figures 2.2 and **??**. These phrase structure trees are produced according to the syntax of the grammar, which is a finite set of rules,[4] able to produce the infinite set of grammatical sentences in a language, while not producing any ungrammatical sentences in that language.

The rest of this chapter goes into more detail about the various instantiations of Chomskyan syntax, each building from the last in order to accurately define the rules that are present in the grammar and their resulting representations.

2.2 Chomskyan Syntax: An Overview

The roots of Generative Syntax lie in a prior intellectual movement broadly known as 'Structuralism,' the specific roots of which in modern-day linguistics began with the work of Ferdinand de Saussure in

[4] As discussed previously, when syntacticians refer to 'rules' they do not mean mean this in a *prescriptive* sense, such as "do not end a sentence with a preposition," but rather in a *descriptive* sense, seeking to describe the possible set of rules that define syntactic operations and, in turn, define syntactic structure. These rules describe how natural language *is* used, not how natural language *ought* to be used. This descriptive approach underlies the research program of linguistics as a whole, whether one is interested in the sounds of language, the meaning of language, or the structure of language.

structural linguistics (for more on the history of syntax, and linguistics as a whole, see Waugh et al., 2023, esp. Ch. 16 and Ch. 17). Saussure ([1916] 1959) thought about language beyond its use as a physical means of communication, developing four related ideas:

1. *Langue* 'language' and *parole* 'speech' are distinct but related entities. *Langue* refers to language as an abstract concept and *parole* refers to language as it is employed in daily life.
2. A 'sign' is the combination of a *signifié* 'signified' and a *signifiant* 'signifier'. *Signifié* is the abstract concept or idea being expressed by the *signifiant*, the sequence of sounds or visual image that is perceived.
3. Signifiers are *arbitrary*, due to the fact that concepts and ideas are expressed differently by different languages (e.g., English *dog* vs. Spanish *perro* 'dog')
4. The meaning of 'signs' is based on their relationships with other signs

Expansion of these ideas influenced structural linguistics and lie at the heart of modern-day linguistics: language is distinctly and relatedly a cognitive faculty and a physical means of communication, combinations of sounds represent an abstract idea or concept, the specific combination of sounds (i.e., a word) is arbitrary in representing an idea or concept, and meaning in language is dependent on the relationship between words and larger phrases. Structuralism moving forward largely centered on defining the primitives (i.e., the smallest units) of language. For example, Roman Jakobson (1952) and Zellig Harris (1951) famously contributed to studying phonology, laying the groundwork for the notion of 'linguistic features' as primitives that combine to form a phoneme.

Noam Chomsky, a student of Harris, was interested in syntax, a relatively understudied area in structural linguistics when compared to things like phonology (describing the sounds of language). Chomsky's goals at-large are to (i) explicate the precise structure of language as a hierarchical symbolic system, (ii) define the implications of that structure, and (iii) to do so with the understanding that syntax is grounded in human psychology and neurobiology. This program of research has primarily focused on two questions (Lasnik & Lohndal, 2013):

1. What is the accurate characterization of users of language? What is an accurate characterization of the capacity to have 'knowledge of language'

2. How does a user of language have this capacity? What elements of the capacity are innate (i.e., present in a language user before any exposure to language)? What elements of the capacity are acquired (i.e., developed from exposure to language(s))?

Chomsky's work since the 1950s has sought to elucidate answers to these questions. This work has had a large impact on developing the theory of syntax as part of larger linguistic theory. In pursuing the aforementioned goals and questions, there have been three broad iterations of Chomskyan Syntax, each one developing due to the shortcomings of the previous: (i) (Extended) Standard Theory (Chomsky, 1956, 1965, 1955/1975), (ii) Government and Binding (Chomsky, 1986), and (iii) Minimalism (Chomsky, 1995, 2001). The rest of this section outlines the core tenets of each of these theories, what phrase-structure looks like in each, and the shortcomings that ultimately led to the current iteration of Chomskyan syntax: Minimalism.[5].

2.2.1 (Extended) Standard Theory

The Standard Theory of syntax began with two works from Chomsky, The Logical Structure of Linguistic Theory (1955/1975) and Syntactic Structures (1957), which has been partially discussed above already. In these works, Chomsky notes two critical elements of language that syntax must explain: structure and infinity. Structure, of course, refers to how language is organized. Infinity refers to the idea that infinite utterances can be constructed out of a finite amount of items. In other words, users of language are able to craft an infinite amount of sentences despite the fact that they are limited to a finite lexicon. One mechanism that allows for this infinity is 'recursion'. Recursion in syntax refers to a constituent's ability to contain another constituent of the same type. A simple example of this is given in (1). The NP contains two prepositional phrases (PPs) nested in each other, *in the crate* describes where the *dog* is and the PP *on the porch* describes where the *crate* is.

(1) The dog in the crate on the porch is barking

[5] It is important to note that this section presents a hindsight-20/20 type review of the history of this vein of syntactic research. Many of Chomsky's works cited below represent conclusions drawn from the synthesis of contributions made by numerous researchers. For a more thorough review of these individual contributions, see Lasnik and Lohndal, 2013 Furthermore, great debate has followed nearly every development reported below, and scholars remain divided on specifics regarding many of the core issues presented below.

In order to explicitly capture these features of language, Chomsky draws on mathematics to define a context-free phrase-structure grammar, thought to represent the mind-internal grammar of a language user. The grammar consists of:

(2) a. A designated start symbol: S

b. Rewrite rules: A symbol on the left, an arrow (expansion), one or more symbols on the right

Using this grammar, sentences are able to be derived. A derivation begins with the designated start symbol and expansions of re-write rules until all symbols have been advanced through, one at a time, until no symbols are left to be re-written and only lexical items remain. An example grammar is given in (3) and a derivation using this grammar is given in (4).

(3) Start symbol: S

Rewrite rules:

S → NP VP

NP → D N

VP → V

D → The

N → dog

V → barks

(4) Line 1: S

Line 2: NP VP

Line 3: D N VP

Line 4: The N VP

Line 5: The dog VP

Line 6: The dog V

Line 7: The dog barks

Constituency is captured by charting back through the derivation, drawing lines to connect each line of the derivation, yielding a PS tree, such as that in Figure 2.4. In the given PS tree, redundant nodes (e.g., the N present in both lines 3 and 4) from the derivation are reduced.

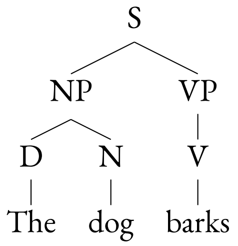


Figure 2.4: Phrase structure tree for *The dog barks*

These PS trees capture the (hierarchical) structure of language through the use of non-terminal (unpronounced) nodes. But, what if someone wanted to say something like *I think that the dog is barking*? In Standard Theory, the recursive nature of embedding two clauses (i.e., two S's) is captured in a special operation called 'General Transformation'. This operation takes multiple trees and combines them into one.

In addition to the general transformation, Chomsky proposes another operation, singulary transformations. Singulary transformations are applied to derivations to yield surface-level representations of language. An example of a singulary transformation is Subject-Auxiliary Inversion, which refers to the fact that when, in English, someone wants to ask a yes/no question, the auxiliary verb appears at the front of the sentence, like in (5).[6]

(5) The dog has barked today → Has the dog barked today?

Transformations such as the one above capture the connectedness of the declarative statement and question by deriving one from the other. Both examples in (5) contain all of the same meaningful elements, the order of their presentation is just shifted to seek out missing information (whether the dog has barked today).

[6] While the example above only includes one transformation, many sentences undergo a series of transformations to get from their base form to their surface form. The order in which these transformations occur are specified, comprising an ordered list of a vast amount of transformations. This is necessitated by the fact that improper rule ordering would result in ungrammatical sentences. These ordered transformations are applied cyclically (Chomsky, 1965), beginning with the most deeply embedded clause and ending with the matrix (i.e., highest-level/main) clause.

In this conception of syntax, the representation from applying only PS rules is referred to as the Deep Structure (DS) and the representation from applying transformations to the DS-representation is referred to as Surface Structure (SS). Semantic interpretation, the meaning of an utterance, is derived from the DS, while surface structure maps to the phonetic interpretation (i.e., how the utterance needs to be pronounced for information-structural reasons). This system captures one of the foundational ideas in the overarching conceptualization of language described throughout this thesis: sentences have multiples levels of representation with abstract structure.

In the name of increased simplicity and explanatory adequacy, Chomsky (1965) proposes updates to the Standard Theory, replacing generalized transformations with recursion-internal PS rules, reducing the amount of possible operations. This allows for rules such as that in (6), which yield an NP like that in Figure 2.5.

(6) PP $\rightarrow$ P NP
NP $\rightarrow$ D N PP

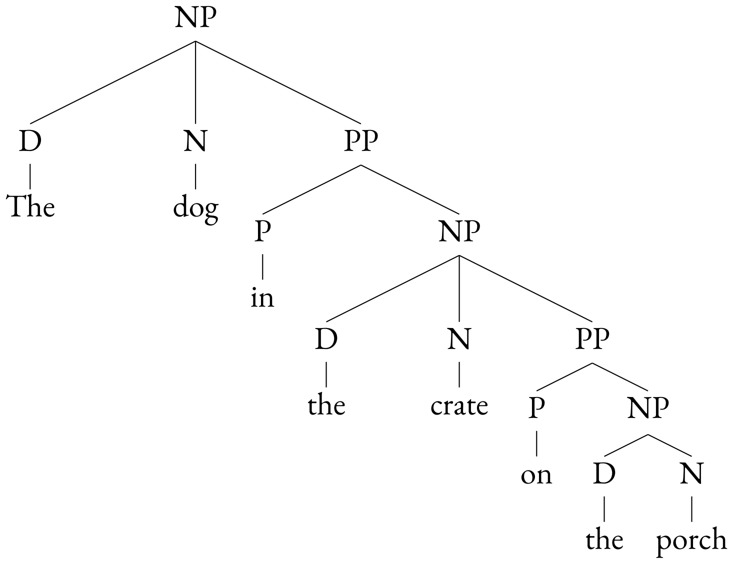


Figure 2.5: Phrase structure tree for NP *The dog in the crate on the porch*

This model proposed by Chomsky (1965) faced some major critiques, particularly regarding the notion that DS strictly dictates the meaning of an utterance (see e.g., Lakoff, 1971). The Extended Standard Theory was born from trying to answer this critique. In this version of the theory, only grammatical relations (e.g., the subject of the sentence) are determined at DS, while SS contributes the remaining semantic information (e.g., scope). The role of DS in determining semantic meaning was further reduced with the introduction of Trace Theory (see e.g., Fiengo, 1977)), which stipulates that whenever an NP undergoes movement, it leaves behind an unpronounced trace in the location from which it moved. This addition essentially meant that all semantic information (including grammatical relations) is determined at SS. More specifically, the representation that maps onto meaning is referred to as Logical Form (LF) and the representation that maps onto sound is referred to as Phonetic Form (PF). This shift of meaning determination in DS and SS to only in SS resulted in the Y-model, shown in Figure 2.6

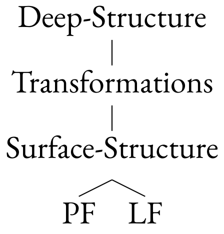


Figure 2.6: Y-Model of Syntax

Briefly returning to the two critical questions that drive the research described here: (i) What does the capacity for having 'knowledge of language' look like and (ii) How is it acquired, the Extended Standard Theory makes claims about both. The capacity for language looks like the Y-model described above. The Deep Structure is generated from the lexical items in an individual's lexicon and are combined according to the phrase structure rules of that individual's grammar. Transformations are then applied to the DS, to get SS. The surface structure is sent to two interfaces one yielding LF, the representation corresponding to a sentence's meaning, and and the other yielding PF, the representation corresponding to the pronunciation of the sentence.

Regarding how the capacity for language comes to be in an individual, children are believed to be endowed with knowledge of the specific primitives at each level of representation. This innate knowledge is then augmented by exposure to language to create a full-fledged grammar. In other words, children are born with an innate knowledge of the levels of abstract linguistic representation (often referred to as universal grammar). Real-world experience with language then informs the properties at each specific level, resulting in the fully-developed ability to comprehend and produce language.

Keeping in mind the connection between the capacity for language and the processing of linguistic input, the description of grammar above should bear on how language is processed. This claim is referred to by Chomsky 1965 as the Competence Hypothesis. More specifically, the hierarchical nature of language in which larger structure is built through the composition of smaller units should play some role in processing. Positive evidence for hierarchy playing a role in processing could come in the form of a computational model trained on data with hierarchical structure outperforming a computational model trained only on sequential data in modeling empirical data reflecting human processing of natural language. Previous studies (e.g., Brennan et al., 2016; Hale et al., 2018), which are further discussed in 5, provide this evidence, though not without methodological criticism (see e.g., van der Burght et al., 2023, which discusses how the variety of theoretical assumptions and implementations of language used in cognitive neuroscience has contributed to a lack of consensus on the best characterization of language). The study presented here ultimately provides additional positive evidence for the role of hierarchical structure in linguistic processing, addressing methodological concerns levied against previous studies.

2.2.2 Government and Binding/Principles and Parameters

The (Extended) Standard Theory successfully explained many phenomena of natural language; however, the construction-specific nature of PS rules and transformations resulted in many ad-hoc solutions and an overall convoluted system made up of numerous and intricate PS rules and transformations. This style of phrase structure was abandoned in favor the X-bar Schema, which paired nicely with the developing Government and Binding (GB) Theory and Principle and Parameters (P&P) Framework (each of which is synthesized in Chomsky, 1986, 1981b, respectively). GB and P&P are worth understanding, as GB, in particular, solidifies many of the developments from the previous decade using the Extended Standard Theory. These developments profoundly influenced the Penn Treebank annotation style (Marcus et al., 1993), which features in the data on which the syntax-knowledgeable language model, TG, is trained.

Principles and Parameters

The Principles and Parameters Framework (P&P) seeks to answer questions (i) and especially (ii) above by specifying that language consists of principles, referring to rules/grammar, and parameters, switches that set grammar-specific 'settings.' The knowledge of principles and parameters are thought to be endowed in a human upon birth; this is the 'innate' part of language. Parameters are then 'set' based on exposure to linguistic input, one aspect of the part of language 'acquired' through exposure.

Principles specifically refer to elements of language that are invariant across the world's languages, while parameters refer to the specific instantiation of that element in a given language. For example, a theorized principle is that all languages must have a subject, and a related-parameter is whether or not that subject must be overtly pronounced. In English, the subject of a sentence must always be pronounced; however, in Italian (and a number of other languages), the subject can be unpronounced (a phenomenon referred to as 'pro-drop'). The pro-drop parameter is 'switched off' in speakers with a grammar of English and 'switched on' for speakers with a grammar of Italian. These switches are made as children acquire their language through exposure to said language.

Proponents of this view argue that the simplicity of this system explains how a child can acquire language from incomplete and imperfect input. When acquiring language, no child hears every word nor every possible utterance in a language. Moreover, the language that they do hear is often imperfect: filled with pauses, tripping over words, mispronunciation, etc. How then do children successfully develop a complete grammar? The simplicity of P&P shines here. The innateness of language—the knowledge of principles and existence of parameters—offers the foundation, which then is augmented by linguistic exposure, rather than a system in which all acquisition of language comes only from linguistic exposure.

Government and Binding

Government and Binding (GB) Theory (Chomsky, 1986) is a specific implementation of P&P, specifying how this framework manifests in the grammar itself. Put differently, GB concretizes how principles and parameters could operate in the grammar. Government refers to abstract relations in syntax (e.g., case assignment from a verbal head to a nominal phrase), while binding concerns the relationships between referents and the antecedents to which they refer.

A simple example of government is the fact that the subject of a sentence in English always receives the nominative (NOM) case, while the object receives accusative (ACC) case. These cases are abstract; however, they manifest morphologically on (some) personal pronouns in English. Example (7a) shows a grammatical English sentence with correct case assignment, while example (7b) illustrates ungrammatical (as indicated by the *) case assignment. In GB, the ungrammatical (7b) is ruled out by the fact that the governor of the subject position (*She* / **Her*) assigns nominative case and the verb (*love*), which is the governor of the object position (*her* / **she*, assigns accusative case.

(7) a. She$_{\text{NOM}}$ loves him$_{\text{ACC}}$.

b. * Her$_{\text{ACC}}$ loves he$_{\text{NOM}}$.

Binding describes the relationship between pronouns and their (possible) antecedents, postulating conditions on possible interpretations of co-referentiality between a pronoun and its antecedent. For example, in (8a) below, the only possible reading is that *himself* refers to *John*, as indicated by subscript i; however, in (8b), there is not an available interpretation such that *him* refers to *John*; *him* must refer to an additional person. Contrast this with the examples in (9), in which *he* is able to refer either to *John* (9a) or to another person (Fred in 9b), indicated by subscript j.

(8) a. John$_i$ loves himself$_i$.

b. * John$_i$ loved him$_i$.

(9) a. John$_i$ told me that he$_i$ is tired.

b. Situation: John$_i$ and I had been talking about Fred$_j$
John$_i$ told me that he$_j$ is tired.

This pattern of grammaticality is captured by two of the three binding principles laid out in GB (Chomsky, 1986). Example (8a) adheres to Principle A, which states that "an anaphor (*himself*) must be bound in its binding domain", while (8b) violates condition B, which stipulates that "a pronoun (*him* above) must be free in its binding domain." Principle A is satisfied in (8a) because *himself* is bound by the co-referential (represented by subscript i) noun *John* in its clause, which is its binding domain. In contrast, Principle B is violated in (8b) because the pronoun *him* is ungrammatically bound by the co-referential noun *John* in its binding domain (the clause). Note that example (8b) would be acceptable if *him* referred to someone other than *John*, because then it would be free in its binding domain.

The grammaticality of both (9a) and (9b) results from both examples adhering to Principle B. The pronoun *he* is free whether it refers to *John* or *Fred*. While both the noun *John* and the pronoun *he* are present in the same sentence, *John* appears in the matrix clause and *he* appears in the embedded clause. The appearance in separate clauses means that the pronoun remains free in its binding domain, satisfying Principle B. In (9b), the noun *Fred* does not appear in either clause, so the pronoun *he* is free in satisfaction of Principle B. These examples highlight the mechanisms that linguists working in GB posited to account for the intricacies of human language.

X-bar Phrase Structure

X-bar Phrase structure (Chomsky, 1970a) is the primary implementation of phrase structure that accompanied GB and P&P. It is argued that X-bar solves many of the ad-hoc problems that arose in the PS of the Extended Standard Theory described above. The general schema for X-bar is given in Figure 2.7. In X-bar, rather than having PS rules specific to each phrase type, the head (i.e., the lexical item), called the minimal projection, projects at all levels of syntactic structure. The X, then, is a variable that ranges across the different lexical heads (e.g., V, N, A, P), which then projects up to each level of the syntactic structure: the bar levels (X') up to the phrase level (the maximal projection; XP). The implications for this are that syntactic requirements of a lexical item percolate up to all levels of syntactic structure. This allows the syntax to capture all requirements of the head while simplifying the previous, less streamlined PS rules system.

X-bar further captures a number of grammatical relationships. The bar levels allow for a clear distinction between complements, constituents selected by the head, and adjuncts, information describing the head in some way. In Figure 2.7, the ZP is the complement, sister node to the head, while the YP is an adjunct, located further away from the head. This placement of the two types signifies the more close-knit relationship of a head and its complement, as well as the looser relationship between said head and its adjunct. The spec(ifier) of a phrase is a mixed set of phrases that in some way narrow the purview of the phrase (e.g., a phrase as a specifier of another phrase).

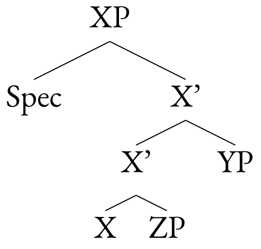


Figure 2.7: Left-Headed X-bar Phrase Structure

X-bar also applies the conceptual ideas of P&P. For example, the headedness of a language — the order in which heads and complements appear — is a principle, that is then parameterized to be left or right headed. Figure 2.7 is an example of what a left-headed language, such as English, would look like. Figure 2.8 is an

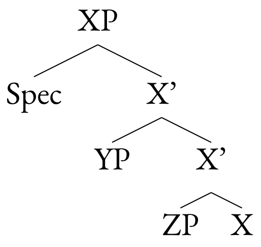


Figure 2.8: Right-Headed X-bar Phrase Structure

example of what a right-headed language, such as Japanese, would look like. This example illustrates one way that the overarching principles and parameters agenda worked out in terms of particular analyses.

Let's return again to the two critical questions described above: (i) What does the capacity for having 'knowledge of language' look like and (ii) How is it acquired. In the theories described here, the capacity for language looks like the Y-model described above, without the very involved and seemingly unconstrained transformations allowed by Standard Theory. The Deep Structure is generated from the lexical items in an individual's lexicon. These lexical items then project their requirements up through the structure, adhering to the rules of the grammar, resulting in SS.[7] X-bar offers a plausible schema for what this abstraction looks like.

Regarding how the capacity for language comes to be in an individual, children are thought to be endowed with knowledge of the existence of linguistic principles and parameters. These parameters are then set in the grammar based on exposure to linguistic input. Overall, the system described here offers a further simplified, more plausible instantiation of the innate faculty of language.

Again, keeping in mind the connection between the capacity for language and the processing of linguistic input, the description of grammar above should bear on how language is processed. If the advancements discussed here do in fact offer a better description of the grammar, then too should these advancements bear on processing. There is some evidence that X-bar representations, derived from a Minimalist parser (Stabler, 2013), better explain fMRI data collected during naturalistic language comprehension than the more basic phrase structure (Marcus et al., 1993) akin to that in the Extended Standard Theory. Stanojević et al. (2023) present another study which incorporates more complex structure. It uses a rich notion of

[7] In GB, there is a operation called move-α, which enables constituents (phrases or heads) to move in order to fulfill various requirements of the grammar, such as acquiring case. These movements are part of the syntax, yielding SS from DS.

hierarchical structure able to address, in a different way, the same discontinuity phenomena that occupied linguists operating in GB. Despite the success of richer hierarchical representations such as these, most studies investigating the role of hierarchical structure in processing use a simpler instantiation of syntax due to its availability and scalability. The present study too uses a simpler phrase structure, further discussed below.

2.2.3 Minimalism

Minimalism (Chomsky, 1995, 2001) remains in the framework of P&P, but further simplifies the operations of the grammar, reducing the set of basic operations to Move, Merge, and Agree (other operations have been proposed by scholars such as Label and Adjunction (see e.g., Hornstein & Nunes, 2008)). Minimalism eliminates many ideas of the previous theories, including DS and SS. Instead, narrow syntax is the locus of computation where syntactic operations take place. These computations occur cyclically in a bottom-up fashion. More specifically, a derivation in Minimalism occurs phase-by-phase, where a phase is a syntactic domain. A finite clause is one example of a phase. The phasehood of finite clauses is motivated by the presence of propositional content, tense, and movement (e.g., the yes/no question formation discussed earlier) within that domain (Chomsky, 2001); however, debate over how best to define a phase remains lively (see e.g., Uriagereka, 2011, for discussion of phases).

The prevailing view of phases is that when all computations are done, a phase is spelled-out, meaning that the syntactic representation is sent to the interface with LF for semantic interpretation and PF for phonetic interpretation (Hornstein et al., 2005). This directly contrasts the last two iterations of Chomskyan syntax, in which the syntactic representation of the *entire* utterance was sent to the LF/PF interfaces only after all computation is complete.

Minimalism dispenses with the distinction between deep structure and surface structure, returning instead to one Generalized Transformation called "Merge" (though in-practice, X-bar is often still used for representatives purposes in the Minimalist framework). Merge is said to be driven by lexical features such as two N features on a verb like *love*, which forces the verb to merge with two nouns, yielding something like *Mary loves John*.[8] This alternate phrase-structure is called Bare Phrase Structure, an example of which

[8] There are additional elements of GB that have been removed. These additional innovations are not discussed here for space reasons. The textbook by David Adger (2003) provides a strong introduction to Minimalism.

can be seen in Figure 2.9. In said example, the N *porch* merges first with the A *back*, the combined N *back porch* then merges with the D *the*. Finally the N *the back porch* mergers with the P *on*.

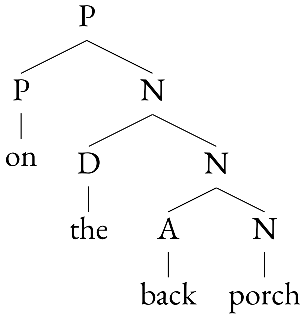


Figure 2.9: Bare Phrase Structure for *On the back porch*

This overview of Minimalism is extremely brief; however, it highlights some of the major changes in the shift from GB to Minimalism. Overall, it seeks to further simplify the computational system of language by reducing operations and attempting to create generalizations that offer better explanatory adequacy of natural language. The adoption of Minimalism has faced much scrutiny from scholars (see for example Lappin & Johnson, 2000a, one article in a larger debate on the merits of Minimalism); however, it is currently the leading research program in generative syntax.

Let's return one final time to the two critical questions described above: 1) What does the capacity for having 'knowledge of language' look like and 2) How is it acquired. In Minimalism, the capacity for language removes the notions of DS/SS, arguing for cyclic spell-out instead, in which syntactic representations, built from a reduced set of operations over lexical items, are sent to the LF/PF interfaces for semantic and phonetic interpretation respectively.

Regarding how the capacity for language comes to be in an individual, Minimalism grows from the P&P framework. Children are thought to be endowed with knowledge of linguistic principles and parameters. These parameters are then set in the grammar based on exposure to linguistic input. Minimalism seeks to understand the extent to which principles and parameters result from the Minimalist model as the optimal view of the human capacity for language.

2.3 From Capacity to Processing

The previous section offers an overview of the Chomskyan tradition in generative syntax, specifically in defining the capacity for language. The core ideas presented above, summarized as: language is a cognitive faculty with computational mechanisms rooted in neurobiology, directly influence this entire thesis; however, the specific phrase structure used in GB and Minimalism do not feature much further on the project. Rather, the hierarchical nature of language, represented in all theories discussed above, is the critical component.

This hierarchical structure can be related to linguistic processing by construing the parsing of linguistic input during comprehension as traversing through the hierarchical tree (tree traversal). Different annotation schema construe tree traversal differently, resulting in what we will refer to as a 'linearized tree.'[9] The two particular annotation schema relevant here are the Penn Treebank (PTB; Marcus et al., 1993) style and Choe-Charniak (2016) style. Both of these schema treat parsing as a top-down, left-to-right process, similar to depth-first search (see e.g., Grune & Jacobs, 2008, §3.5.2). Constituency is captured by matching brackets, where the opening bracket posits a node and the closing bracket indicates that computation for that node is complete. This matching-bracket style is comparable to the index/outdex notation of Graf 2017's Minimalist parser, in which the index marks the time point when a node is posited and the outdex marks the time point at which computation for that node is complete. Both of these styles capture hierarchical structure and relate it to the temporal aspect of comprehending language (i.e., hearing or reading linguistic input in real time).

Consider the syntax tree for the sentence *The dog barks* in Figure 2.4. The PTB linearized tree is given in (10a) and the Choe-Charniak tree is given in (10b). According to this tree traversal view, when a comprehender reads the sentence *The dog barks*, the mental computation that takes place to process the sentence while being read is like that of traversing the hierarchical tree, including both the pronounced words and the unpronounced hierarchical structure.

(10) a. (S (NP (D the) (N dog)) (VP (V barks)))

b. (S (NP the dog NP) (VP barks VP) S)

[9] The term linearized tree is synonymous with what linguists call 'labelled bracketing'.

In the PTB annotation style, all words are assigned a part-of-speech (POS) tag and are labelled for their phrase-type. The brackets capture hierarchy by illustrating that the D and N are subconstituents of the NP, the V is a subconstituent of the VP, and both the NP and VP (and all sub-elements) are subconstituents of S. The Choe-Charniak parsing takes the PTB trees, strips them of their parts of speech, and composes them into just their phrasal representation. The opening (S, (NP, and (VP posit those nodes and the NP), VP), and S) mark the end of computation for those composed phrases.

The BLLIP-LG Dataset (Charniak et al., 2000), as prepared by Hu et al. (2020), acts as the training, validation, and testing data for the Transformer Grammar (TG; Sartran et al., 2022), the syntax-knowledgeable language model from which surprisal values are derived in this project. That data set is annotated according to the PTB annotation guidelines. As part of the pre-processing of the data for the TG, these PTB style trees are converted to the Choe-Charniak style before being used for training, validation, and testing. An un-annotated version (i.e., words only) of the same data set acts as the training, validation, and testing data for the Transformer-XL (TXL; Dai et al., 2019), the syntax-less language model.

These two language models, the TG and TXL, incorporate syntax differently. As just discussed thoroughly, there are many different properties of syntactic representations – features, constituency, movement, traces, co-indexation, etc. TG incorporates syntax into the model by explicitly encoding constituency when modeling language. In contrast, the TXL model does not incorporate any sort of explicit encoding of syntax; however, it is possible that some sort of syntactic structure will emerge and be learned by the TXL model.

If the hypotheses reviewed throughout this chapter represent the best characterization of language, and the grammar of humans is symbolic and hierarchical in nature, with observable correlates in neurobiology, then the TG should better model human sentence processing than the TXL in certain regions of the brain. In contrast, if the hypotheses reviewed later in chapter 4 represent the best characterization of language, and a sequential view of language better describes the language system, then the TXL model should outperform TGs in certain regions of the brain. For more on this, see chapter 7.

Chapter 3

The Role of Hierarchical Structure in Processing and its Neurobiology

A number of things must be described in order to characterize the role of hierarchical structure in linguistic processing and the brain regions that subserve this computation. Given that we cannot simply look inside the brain to directly understand how processing works, what human responses (psychometric data) to language comprehension (reading written language, seeing signed language, or listening to speech) reflect processing? How is that psychometric data linked to theorized cognitive processes, such as hierarchical computation or sequential pattern recognition? How can those cognitive processes be related to the brain regions that carry them out?

Psychometric[1] data such as reading time, eye movements during reading, scalp voltages, and BOLD signal are all indicative of human effort during linguistic processing. More effort or difficulty during processing is reflected in this psychometric data. For example, longer reading times indicate greater processing difficulty, which could be explained by the mind-eye assumption, which postulates that the human eye remains fixated on a word as long as it takes to process (Just & Carpenter, 1980).

[1] The term psychometric here refers broadly to behavioral or neural responses elicited during language comprehension.

3.1 Surprisal as a Linking Hypothesis

One way to relate this psychometric data and theories of linguistic processing borrows from information theory, using the metric Surprisal (Shannon, 1948). Generally, the surprisal for an outcome y is equivalent to the below equation, where y stands for Y=y (i.e. the observed outcome of the random variable Y):

$$log_2(\frac{1}{P(y)})$$

Surprisal, applied to language (Hale, 2001; Levy, 2008), quantifies the transition from one word to the next in the incoming linguistic input. Following the formulation provided by Hale (2016), surprisal for the next word in a string is equivalent to the log of the ratio of the 'prefix probability' of that next token and the 'prefix probability' of the preceding substring. The prefix probability is the total probability mass assigned by the substring based on the probability distribution derived from the grammar. In other words, prefix probability is the conditional probability of the entire string up until that input, obtained by multiplying the conditional probabilities of each input so far. The transition probability to the next word, then, is the ratio of these two prefix probabilities. This transition probability is the outcome, y, in the surprisal equation above. This formulation is given in Figure 7.1. The prefix probability for the preceding substring is written as $\sum$ before and the prefix probability for the string including the next word is written as $\sum$ after.

Let y be the ratio of prefix probabilities

$$
\begin{aligned}
&log_2(\frac{1}{P(y)})\\
&= log_2(\frac{1}{\frac{\sum \text{after}}{\sum \text{before}}})\\
&= log_2(\frac{\sum \text{before}}{\sum \text{after}})\\
&= -log_2(\frac{\sum \text{after}}{\sum \text{before}})
\end{aligned}
$$

Figure 3.1: Surprisal of Prefix-String Probabilities

Surprisal acts as a linking hypothesis between language models and empirical data observable during language processing (Hale, 2001, 2016). The formulation of word surprisal in terms of prefix probabilities (Figure 7.1) means that any language model can derive predictions about cognitive load associated with incoming input. Comprehending unexpected input should correspond to higher surprisal values, which in turn should correlate with empirical data that reflects greater processing difficulty (e.g., higher reading time in a reading time study or greater BOLD activation in fMRI). This relationship has proven robust, yielding positive results across a number of methodologies such as reading time studies (e.g., Pimentel et al., 2023; Roark et al., 2009), eye-tracking (e.g., Boston et al., 2008, 2011), EEG/MEG (e.g., Brennan & Hale, 2019; Heilbron et al., 2022), and fMRI (e.g., Hale et al., 2015; Willems et al., 2016).

These surprisal values are derived according to theorized comprehension mechanisms, implementing whatever capacity for language is instilled in the language model (whether hierarchical or sequential). If surprisal predicts this psychometric data, then inferences can be made about about how hierarchical structure or sequential word ordering impact processing and from where in the brain this impact is driven. If language models that incorporate hierarchical structure predict psychometric data above-and-beyond language models with only a sequential view of language, this supports the existence of hierarchical structure and its role in linguistic processing. If sequential-only models better predict psychometric data, then this supports the notion that only surface-level ordering lexical items play a role in processing.

3.2 Investigating Hierarchy

To test the role of hierarchical syntax during sentence comprehension, Brennan and Hale (2019) investigate whether linguistic processing reflects hierarchy and not just linear order in everyday-language tasks such as listening to an audio book. They compared surprisal values from two sequential models — a trigram model and 3-layer recurrent neural network (RNN) — and a hierarchical model, a probabilistic context-free grammar (CFG). Scalp voltages were recorded using EEG while participants listened to an audiobook recording of the first chapter of *Alice in Wonderland*. They found that linear regression models with surprisal from each language model were a better fit than a baseline with only variables of non-interest. They also found that the sequential models were a better fit than the CFG in some anterior right portions of the scalp, while finding that the CFG explained the EEG-recorded scalp voltages above-and-beyond sequential models in left anterior and certain right anterior portions of the scalp. Using the same data but a different implementation of a probabilistic phrase structure grammar, Hale et al. (2018) find that surprisal from the hierarchical model correlates significantly with voltages in the anterior portion of the scalp during an early time window (200-400 ms post onset of stimulus).

Using both EEG and MEG, Heilbron et al. (2022) find that surprisal modulates evoked responses at multiple levels of language — phonemic, syntactic, and semantic while participants listened to an audiobook; however, their 'syntactic level' is not hierarchical, but rather sequential, corresponding only to part-of-speech. Surprisal derived from part-of-speech correlated with bilateral temporal regions, while semantic correlation spread across a wider set of bilateral cortical regions. These findings highlight that linguistic processing occurs at multiple levels of representation and that the processing of these representations correspond to different brain regions.

A number of fMRI studies have also investigated the role of hierarchy in language processing. Hale et al. (2015), Brennan et al. (2016), Brennan et al. (2020), and Shain et al. (2020) all explore the role of hierarchical structure in language processing by comparing surprisal from syntax-less language models and surprisal from those with syntactic knowledge. Hale et al. (2015) compares multiple models of syntax with varying degrees of complexity in modeling said syntax. They find that Penn treebank style phrase structure (Marcus et al., 1993) improves a regression above-and-beyond n-gram models and an X-bar style minimalist grammar (Stabler, 2013) further improves on the Penn-style phrase structure. These results hold in anterior temporal lobe (aTL), an area implicated in syntactic processing (Friederici & Gierhan, 2013). Brennan et al. (2016) and Brennan et al. (2020) report similar findings that surprisal from structured language models is a significant predictor of fMRI timecourses in left anterior and posterior temporal lobe. These studies, as well as that conducted by Hale et al. (2015), report that sequential models successfully predict the fMRI timecourses in these regions as well as broader regions spread across the language network, including inferior frontal gyrus (IFG). Shain et al. (2020) finds that surprisal from both a 5-gram model and a PCFG predict fMRI timecourses in the language network. Each model modulates the BOLD signal in the language network independently of the other, indicating that the human language processor is sensitive not only to structure but also independently, and in-parallel, to surface-level co-occurrence patterns of words themselves.

Taken together, the studies above affirm that hierarchical structure plays a role in human sentence processing and that these mechanisms occur in specific brain regions. There are some differences across studies in the specific brain regions implicated in these processes; however, these could be due primarily to methodological differences in the experiments, language models used, and tasks carried out. While these small differences exist, there is widespread evidence implicating Broca's area in hierarchical syntactic processing (e.g., Brennan & Hale, 2019; Hale et al., 2018; Shain et al., 2020), as well as other regions in the temporal and frontal lobes (Brennan & Hale, 2019; Brennan et al., 2020; Hale et al., 2015). Largely, the brain regions that subserve language, referred to as the language network (Friederici & Gierhan, 2013), are situated in the left hemisphere; however, a number of studies have shown that additional right hemisphere homologues to left language network are implicated in linguistic processing (e.g., Crabbé et al., 2019; Dunagan et al., 2023; Heilbron et al., 2022; Wehbe et al., 2014).

The studies above also implicate a number of brain regions in sequential lexical processing. This points to a system in which certain brain regions correspond to lexico-semantic information about the words and the order they appear in, while other regions are sensitive to other levels of representation, such as hierarchical syntax. In any view of language, a system with multiple levels of representation, which involve multiple brain regions, should not be surprising. The evidence above supports that (at least) one of these levels incorporates hierarchical structure in processing and involves specific brain regions. The next chapter goes into more detail regarding the sequential-only view of language, before chapter 5 relates sequential-only views directly to theories of processing that do, in fact, argue against any dissociable impact of syntax in processing.

Chapter 4

Language as Sequential

Chapter 1 introduced a number of views attempting to explain human cognitive abilities and behavior, each of which make predictions about what the cognitive system underlying human language looks like and how it works. There are a number of ways that these theories could be divided. The two distinctions relevant here are (i) symbolic vs. non-symbolic approaches to cognition/behavior and (ii) hierarchical vs. sequential models of human language. The previous chapter reviewed a symbolic and hierarchical model of the capacity for human language. The present chapter discusses various sequential views of language, beginning with strict behaviorism (Moore, 2017; Skinner, 1963), then middle-ground associationism (Wilson, 1980), a two-system theory of comprehension (Townsend & Bever, 2001), and finally the recent reemergence of connectionism (e.g., Goldstein et al., 2022a).

4.1 Rejection of Behaviorism and Middle-Ground Associationism

Strict behaviorism argues that human behavior and learning occur due to stimulus-response pairs, in which a response is conditioned on a given stimulus. Learning, and in turn behavior, arise as an individual repeatedly experiences a given stimulus and its response. This view contrasts the cognitivist view discussed in the last chapter, in which language, among other cognitive faculties, employ computational manipulation of symbols. A behaviorist view of learning and human behavior places all emphasis on a mind-external stimulus as the direct cause of the behavior of individuals, rather than their internal mental state prompting the behavior (Moore, 2017; Skinner, 1963). Thinking then is a type of behavior that arises in response to other behavior, contrasting the cognitivist view that thought drives behavior. The assumptions of strict behaviorism are summarized by Wilson (1980, pg. 14), modified from Anderson and Bower (1973):

1. Assumption of Observationally Based Concepts
   The only conceptual elements required in a psychological explanation can be put into direct one-to-one correspondence with observable events, or are direct derivatives of such events. The observable events are stimuli, responses or reinforcements. The derived events include mediating responses, sensations or reinforcement contingencies.

2. Assumption of Association by Contiguity
   The elements of 1 above become connected or associated only if they occur (or are activated) in close temporal contiguity, with the association being from the prior to the subsequent element.

3. Linear Sequence of Activated Associations
   When the associations of 2 are activated, each prior element activates but one subsequent element.

4. Completeness Assumption
   All observable behavior is generated by associative connections formed as in 2 and activated as in 3.

In the behaviorist view, there is no symbolic structure to language, nor any hidden representations, but rather it is learned from matching observable stimuli (speech, writing, signs) to responses. Associations, such as the noun appearing before a verb being the doer of that verb, form as these stimuli-response pairs occur in close temporal proximity repeatedly. These associations are then activated, sequentially and one at a time by the prior element. In language, this looks very similar to the finite state sequential model criticized by Chomsky (1957) and discussed in chapter 2. The conclusion then is that language, and all other observable behavior, is generated by these sequentially-activated associations.

This strict view of behaviorism was widely critiqued, especially by cognitivists like Chomsky, for its ignorance of the complexity of internal processes such as thought (Chomsky, 1980; Wilson, 1980). These critiques led to the development of theories of behaviorism and associationism that are more nuanced, better taking into account mind-internal processes. One of these views was developed by Kellogg Wilson (1980), which we refer to as 'middle-ground associationism'. This middle-ground associationism critiques the strict behaviorist view along the same lines; however, it levies major critiques against Chomsky's mentalist view of language. In particular, Wilson argues that Chomsky's attempts to define language without regard for additional psychological processes is, too, an oversimplification. Moreover, Wilson (1980, Ch. 3) picks apart Chomsky's argument that sequential models cannot capture the fact that a sentence in language can be of infinite length (Chomsky, 1957). He argues that, while in theory, language can be of an unbounded length, in reality it is limited by other psychological factors such as working memory. To only describe the ideal adult language user with no regard for other psychological factors will fail to adequately describe human psychology, including language. Such weakening of Chomsky's idealization has been pursued in the literature on finite state approximation of grammar (see e.g., Langendoen, 1975; Pereira & Wright, 1991).

It is true that Chomsky's view does not take into account other psychological factors; however, the enterprise of generative syntax is not concerned with these other factors, rather:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its (the speech community's) language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance. Chomsky (1965, pg. 3).

To put this in the terms used throughout this thesis, Chomsky is concerned with the capacity for language and leaves the question about how this capacity is utilized in processing, modulated by other psychological factors, to others. This is where the psycholinguistic work investigating linguistic processing comes into play. In this sense, Wilson's critique of Chomsky's enterprise is a valid one; however, it does not eliminate the utility of the research program. Instead, multiple avenues of research in formal linguistics, psycholinguistics, and neurolinguistics comes together to create a full picture of language: its capacity, processing, and neural correlates.

Middle-ground associationism focuses primarily on the capacity for language as a part of behavior. There are three core tenets to this view, summarized from Wilson (1980, pg. 295):

1. Internal representations exist for external stimuli and physical responses. In early development, associations are formed between those representations as one is exposed to them repeatedly.

2. After these associations are built, they can act as building blocks in further associations. Behavior in early development, then, is similar to the stimulus-response pairs from behaviorism; however, later development does not. Rather, later development employs increasingly complex associations.

3. Humans are able to develop propositional concepts, which are integrated into longer sequences via association. Problem solving then occurs by employing the longer sequences (e.g., plans) to guide behavior.

The tenets described above define a system in which early development looks similar to behaviorism, but that internal representations become more complex through associative linking. Eventually, these internal representations are employed in longer sequences to engage in higher-level thinking, such as problem solving or language. The major development from behaviorism is the acknowledgement of some sort of internal representation, as well as increasing complexity of associations, rather than only simple stimulus-response pairs.

This view of behavior does, in fact, fall between early behaviorism and cognitivism; however, critically, it still argues for a sequential view of language. The associative structures employed in planning complex actions remain associated sequentially. Additionally, this system argues that internal representations are crucially grounded in particular observable words, unlike the phrase structure trees described in chapter 2 that had internal nodes such as NP and VP.

In summary, middle-ground associationism presents a sequential view of language that takes ideas from both behaviorism and cognitivist views of language, such as Chomsky's.

4.2 Late Assignment of Syntax Theory

Going a step further in describing the relationship between the surface of language and possible structure is Late Assignment of Syntactic Theory (LAST; Bever, 1970; Townsend & Bever, 2001). This view of sentence comprehension combines the associationist habit-based approach with the cognitivist rule-based approach. In this theory, the sentence is the primary level of interpretation. Each sentence is interpreted twice: once according to templates that are based on habits and once according to the full syntax system. These templates, also referred to as pseudosyntax, form an initial interpretation of the relationship between meaning and form in the sentence. This initial interpretation then acts as the input for "synthetic parsing" in which another interpretation is developed according to the full system of abstract grammatical rules. This full syntactic representation is then compared to the initial interpretation. If these interpretations match, the meaning is stored and full syntactic information is no longer readily available.

The pseudosyntax, or templates, in LAST are established based on repeated exposure to similar sequential structures in language, and the sentence is the level at which associative patterns are formed. A prime example of this is the Noun-Verb-Noun (NVN) template for transitive verbs and its variants: NV for intransitives and NVNN for ditransitives. Townsend and Bever argue that difficulty in linguistic processing arises when the pseudosyntactic parse of a sentence and the full syntactic parse are mismatched, because this indicates that the template applied to the actual input does not match the full structure of that input, and therefore additional processing cost is incurred updating the parse. For example, LAST explains the famous processing difficulty incurred by garden path sentences, such as:

(11) The horse raced past the barn fell.

Given the N-V template usually assigns the semantic roles of "agent-action" the initial parse is something like:

(12) The horse raced past the barn. Fell.

However, after the full syntactic parse properly assigns "raced past the barn" the structure of a reduced relative clause, more similar to the structure of a full relative clause, such as:

(13) The horse which was raced past the barn (by the jockey) fell.

The salient factors of the initial parse – notably the first N and a verb that can act intransitively, which would assign the N-V and agent-action template, an ultimately incorrect parse – induce the garden path effect and its incurred processing cost. This example highlights how LAST uses both sequential and hierarchical information in processing language.

To summarize, LAST (Bever, 1970; Townsend & Bever, 2001) posits a capacity for language that could be similar to that defined by Chomsky and others, in that it is symbolic and involves hierarchical computation; however, in terms of how this plays out in linguistic processing, it argues for a sequential-first view. Following the sequential parse, a hierarchical parse occurs too, but the syntactic information only becomes available if the pseudosyntactic parse and full syntactic parse deviate from each other (For a critical assessment, see Phillips, 2013). The view of linguistic processing tested by the study at hand posits that hierarchical syntactic information is available and relevant upon the initial parse; however, the results presented in later chapters *could* be compatible with this theory.

4.3 Connectionism Revitalized

Connectionism, briefly discussed in chapter 1, is a non-symbolic cognitivist line of research that attempts to explain cognitive capabilities using artificial neural networks (see e.g., Clark, 2001, Ch. 4). This modeling is inspired by the neuron in the brain and how these neurons combine to carry out higher functions than each neuron could do on its own. In this view, researchers seek to better understand the nature of the mind and its relation to the brain based on the behavior of artificial neural networks (Rumelhart, Hinton, McClelland, et al., 1986). At its core, connectionism asserts that a distributed network of brain regions carry out large quantities of cognitive operations simultaneously (Buckner & Garson, 2019).

Historically, modeling language using neural networks was not widely adopted due to limitations in the technology, resulting in simple networks with relatively few hidden layers not suited to explaining the complexity of human language. Recent advancements in deep learning and the advent of large language models (LLMs) have revitalized this connectionist approach. Notably, LLMs have been able to generate large swaths of human-like language with minimal mistakes (Goodkind & Bicknell, 2018; Wilcox et al., 2020). For example, ChatGPT is a Generative Pre-Trained LLM (Radford et al., 2019; OpenAI, 2023) fine-tuned to act as a conversational agent that has experienced large-scale success in producing and responding to human language. This widespread success of LLMs in producing and responding to human language has prompted a resurgence in treating language models as a "biologically feasible computational framework for studying the neural basis of language" (Goldstein et al., 2022a).

A number of scholars argue for LLMs as suitable cognitive models of how the brain encodes and decodes human language (see e.g., Contreras Kallens et al., 2023; Goldstein et al., 2022a; Tuckute et al., 2024). This view argues that there are a number of shared principles between how humans and LLMs process language. This processing is non-symbolic and sequential, based on words and the order in which they appear. In the connectionist view of language, representational constraints such as phrase structure rules are excluded in favor of overt associations between words, represented by patterns of activation (Bates et al., 1996).

Chapter 5

Sequential Processing and its Neurobiology

The human-like performance of LLMs (e.g., Goodkind & Bicknell, 2018; Wilcox et al., 2020) on a number of linguistic tasks has prompted questions about how best to characterize human language processing. These models employ a statistical approach to language learning, trained on massive bodies of text. Their success, trained only on real life linguistic data, have led some scholars to conclude that "LLMs have already demonstrated that human-like grammatical language can be acquired without the need for a built-in grammar" (Contreras Kallens et al., 2023, p. 1), calling into question any sort of explicit symbolic grammar in humans, including one that employs hierarchical structure.

Using fMRI, Fedorenko et al. (2020) probe which brains regions correspond to manipulations of different levels of linguistic representation: lexico-semantic vs. morpho-syntactic. Across the three experiments conducted, they found that regions across the Inferior, Temporal, and Angular Gyri were equally or more sensitive to lexico-semantic information than morpho-syntactic information. No regions of the brain were more sensitive specifically to syntactic processes. They argue that these results support a language system which is driven by meaning, as derived from words and the order in which they appear, with minimal to no dissociable effect of combinatorial or syntactic processes.

This view is argued for a step further, bringing in LLMs and their successful linguistic performance specifically, by Contreras Kallens et al. (2023). These authors argue (p.4) that LLMs "can be viewed as working models of the potential of pure statistical learning of grammar based on prediction, memorization, generalization, and abstraction". This view directly calls into the question the relevance of grammar, pushing for a view in which there is no grammar and the language system is pure statistical learning.[1]

Goldstein et al. (2022a) further explicate this view by arguing that LLMs are excellent models of human language due to shared principles in how each processes language. The similarity between how humans and LLMs process language is summarized in three shared computation principles: "(i) both are engaged in continuous context-dependent next-word prediction before word onset; (ii) both match pre-onset predictions to the incoming word to induce post-onset surprise (that is, prediction-error signals); (iii) both represent words using contextual embeddings" (Goldstein et al., 2022a, p. 369). Through a series of experiments probing language models, they find that human and LLM performance on a next-word prediction task are near identical, that the success of predictions in both humans and LLMs results in surprise after the onset of the predicted word, and that contextual embeddings better predict neural responses than static embeddings of words. They argue from these findings indicate that LLMs and humans employ similar language systems. These systems are sequential and based on statistical learning through exposure.

The shared computational principles offered here can co-exist with a hierarchical view of language. All of the studies reported on in chapter 3 acknowledge the role of prediction, as well as the role of word-level information in prediction, during linguistic processing. The role of prediction is undeniable in human processing; however, the information used to make predictions can exist at multiple levels of representation, including both at the word level and the syntactic level, as previously shown (e.g., Heilbron et al., 2022; Shain et al., 2020).

Another recent study by Tuckute et al. (2024) argues that LLMs are able to mimic human language and govern neural activity. They train an LLM to generate predictions about the magnitude of BOLD response using last-token sentence embeddings from GPT2-XL. The BOLD responses used as training data for the model were taken while participants listened to 'baseline' sentences (n = 10 scanning sessions). These sentences generated a baseline amount of neural activity. The evaluative data for the model (n = 9 scanning sessions), collected from another set of participants, included 'drive' and 'suppress' sentences. The drive sentences included 250 sentences selected to elicit maximally strong activity in the language

[1] Statistical learning (e.g., Charniak, 1996; Newport, 2016) refers to using the statistical information from the distribution of linguistic elements that one encounters during language learning to draw conclusions about how those elements can be used (e.g., categories of words, contexts in which words are used, ordering of these categories, etc.).

network (e.g., Turin loves me not, nor will). The suppress sentences included 250 sentences that elicit minimal activity in the language network (e.g., We were sitting on the couch). The results of feeding the evaluative sentences into the LLM were that BOLD signal for drive sentences was 85.7% higher than baseline and suppress sentences were 97.5% lower than baseline. The authors conclude that this indicates that LLMs not only mimic human language abilities but that they drive and suppress neural signal in high level cortical areas.

This study by Tuckute et al. (2024) sets up a very interesting follow-up study, in which one could compare two LLMs: one that is sequential only like GPT2-XL and one that is hierarchical like TG. One could then compare the outcomes of the two models to explore the role of hierarchical syntax in driving or suppressing neural responses. If there are areas of the brain in which the hierarchical representations better modulate the neural response, it would again indicate that these syntactic processes take place.

The studies presented above fail to find evidence for hierarchical structure. Rather, they argue that language is the result of generalizing from statistical patterns of language. This exact stance has been repeatedly denied by Chomsky, arguing that the internalized grammar in the adult state is "in no sense an inductive generalization from [primary linguistic] data" (Chomsky, 1965, p. 33). Nonetheless, in this view, linguistic processing too is based only on sequential patterns of surface-level word ordering rather than any sort of hierarchical structure. In this view based on the evidence presented above, linguistic processing is primarily driven by lexico-semantic information and this occurs across the language network.

Evidence in chapter 3 supported the role of hierarchical structure in linguistic processing, while the evidence here is consistent with the hierarchy-free theories discussed earlier in chapter 4. The study presented here contributes to this debate by comparing two language models, one with hierarchical syntax and one with out. This comparison between TG and TXL provides an especially clean comparison because the architecture of the two language models differ only in whether or not an attention mask that is sensitive to hierarchical syntax is turned on.

Chapter 6

Language Models

Remember that language models are used as approximations of human linguistic behavior. To re-iterate, language models are defined as models that, given a lexicon, assign a probability to every possible string of words from said lexicon (Jelinek et al., 1992). The equation for modeling probability for a given word in a string, based on the preceding context, is:

$$P(W_n = w_n | W_1 = w_1, ...W_{n-1} = w_{n-1})$$

The above equation can be read as the probability that the nth word is w_n if the prior words in the string were w_1 through w_{n-1}. Simply put then, a language model is "an assignment of probability to each string in this set" (Hale, 2016, p. 398). How these probabilities are defined differ across language models.

This study employs two language models, the TG, which operates with hidden hierarchical representations, and the Transformer-XL, which operates with sequential surface-level representations. The only difference between these two models is an additional attention mask in the TG, which encourages the model to use hierarchical representations in its processing of language. In contrast, TXL only uses sequential word-level information in its processing. These differences are further discussed below.

6.1 Deep Learning

Deep learning refers to artificial neural networks that use multiple layers to learn representations from massive amounts of training data (Jurafsky & Martin, 2023). These models do not begin with any sort of representations, but rather learn representations from their data; however, they can be imbued with inductive biases that constrain this learning to certain types of representation. In fact, TGs have an inductive compositional bias, which encourage them to learn compositional representations. This inductive bias is further discussed below.

6.2 Backpropagation and Word Embeddings

The method through which artificial models learn these representations is called 'backpropagation' (Rumelhart, Hinton, & Williams, 1986). Backpropagation was previously mentioned in relation to connectionism, whose early instantiations provide simple deep learning model examples. Consider again Figure 6.1, which is a neural network from which more modern deep learning techniques, such as transformers (Vaswani et al., 2017), take inspiration.

These artificial neural networks take an input. This input has certain activation values, corresponding to abstract feature(s) in the network, and this activation value feeds forward to all nodes in the first "hidden" layer to which the input node is connected. These activation values feed-forward through as many hidden layers as are present in the network, ending up in an activation function that maps the combination of activation values from all hidden nodes to a specific output (Buckner & Garson, 2019). The strength of connections between two nodes are referred to as "weights". The values of weights fluctuate from negative values, inhibiting a connection, to positive values, activating a connection. This simple neural network is called a feedforward network, in which the connections between each layer only operate moving forward (Elman, 1990). Critically, the weights between units are adjusted during a neural network's training via a method called "backpropagation." During training, when the output of the neural network does not match the expected output, the "error" backwardly propagates through the hidden units, adjusting their weights so that the neural network is more likely to correctly identify the output next time.

This type of learning is referred to as gradient learning. Generalizing from the nodes in neural networks specifically, gradient learning minimizes error by adjusting weights according to the partial derivative of the function that calculates said error associated with each parameter in the model.

One way to use artificial neural networks in the study of language is to use them in language models. A language model assigns a probability to every string of words from the lexicon (Jurafsky & Martin, 2023)

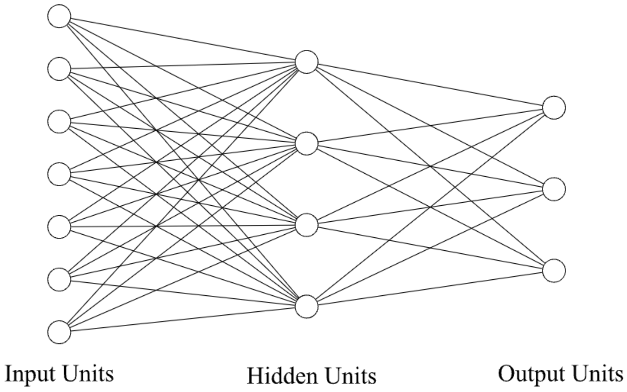


Figure 6.1: A example of a neural net (Buckner & Garson, 2019)

by assigning a probability to each incoming word based on the previous context. Rather than representing these words as the words themselves, they are represented by a vector of numbers called a ‘word embedding’ (see e.g., Pilehvar & Camacho-Collados, 2020). These vectors represent the generalization of the idea of decompositional semantics. In decompositional semantics (see e.g., Katz & Fodor, 1963), each number in the word embedding vector is thought to represent some quality about the word. For example, the word *king* might have a number in its word embedding that represents the quality of *royalty* and another representing the quality for *man*. If each component were symbolically interpretable, as in the given example, they might be viewed as features or properties; however, there is no requirement that they have such an interpretation. In practice it is often impossible to make sense of what a component "means." Despite the uninterpretability of specific components, these word embeddings are relevant throughout deep learning architectures in language modeling; notably they act as the input for the model.

6.3 Transformers

Transformers (Vaswani et al., 2017) are a recent deep learning architecture used heavily in language modeling among other machine learning tasks. Transformers use both techniques mentioned above, word embeddings as well as end-to-end training via backpropagation. The transformer also introduces the ‘transformer block’, which is a set of layers carrying out computation in the model. The transformer block is the backbone of the transformer architecture. At the core of the transformer block is ‘self-attention’.

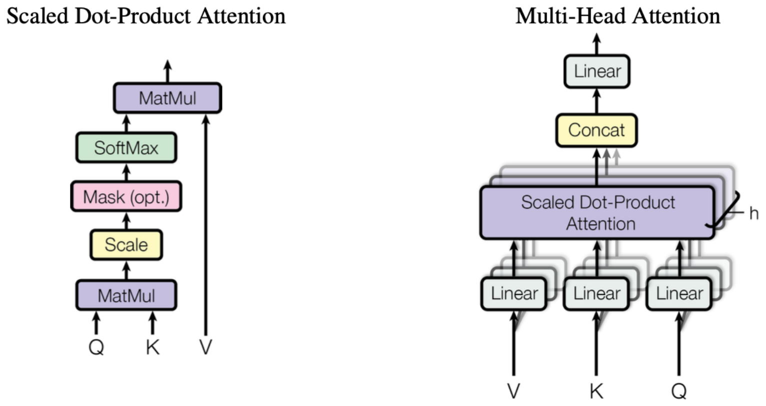


Figure 6.2: (Left) Scaled Dot-Product Attention. (Right) Multi-Head Attention consists of several attention layers running in parallel. (Vaswani et al., 2017, p. 4)

6.3.1 Self-Attention

Modeling natural language must account for the fact that words are related over arbitrarily long distances. One recent innovation in capturing this distance is (multi-head) 'self-attention' (Vaswani et al., 2017). Self-attention relates "the different positions of a single sequence in order to compute a representation of the sequence" (Vaswani et al., 2017, p. 2), enhancing the information about the current input embedding by incorporating information about its context. In other words, a model uses self-attention to determine the importance of different tokens in a sequence when making predictions.

The attention mechanism maps a query and a key-value pair to an output; the query, key, value, and output are all vectors. The particular attention mechanism implemented by (Vaswani et al., 2017) is 'Scaled Dot-Product Attention.' In this version of attention, the input is made up of queries, keys, and values, all of dimension d. In order to obtain the weights on the values, the dot product of the query with all keys is computed, scaled by dividing by $\sqrt{d_k}$, and entered into a softmax function[1]. The result of the softmax function is the attention score. The input is packaged into the matrix Q, keys into matrix K, and values into matrix V, computing the matrix of outputs as:

[1] The scaling is done to avoid numerical instability due to attention weights growing too large or shrinking too small.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6.1}$$

Rather than performing attention with a single d_{model}-dimensional queries, keys, and values, attention can be computed using multiple attention heads. The inputs are linearly projected h times with different, learned linear projections. Attention is then computed in-parallel, resulting in d_v-dimensional output values. These values are then concatenated and projected again, yielding the final values as shown in Figure 6.2. This multi-head attention enables the model to attend to more information from additional representational spaces at different positions than a single attention head. Multi-head attention looks as follows:

$$\begin{gathered}\text{MultiHead}(Q, K, V) = \text{Concat}(head_1 ... head_h)W^O \\ \text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{gathered} \tag{6.2}$$

Where the projections are the parameter matrices in equation 1.3.

$$\begin{aligned} W_i^Q &\in R^{d_{model} \times d_k} \\ W_i^K &\in R^{d_{model} \times d_k} \\ W_i^V &\in R^{d_{model} \times d_v} \\ W^O &\in R^{h_{dv} \times d_{model}} \end{aligned} \tag{6.3}$$

6.3.2 Transformer Blocks

Self-attention scores are the critical component in transformer blocks, shown in Figure 6.3. Looking at the transformer block, if we were to zoom in on the self-attention layer, it would look like the right image in Figure 6.2. The residual connections in the transformer block allow for information to be passed from a lower layer to a higher layer without going through any intermediate layers (Jurafsky & Martin, 2023). The layer normalization is similar to z-scoring in statistics, and is used to keep the values in the hidden layers within a certain range to maximize gradient learning during training.

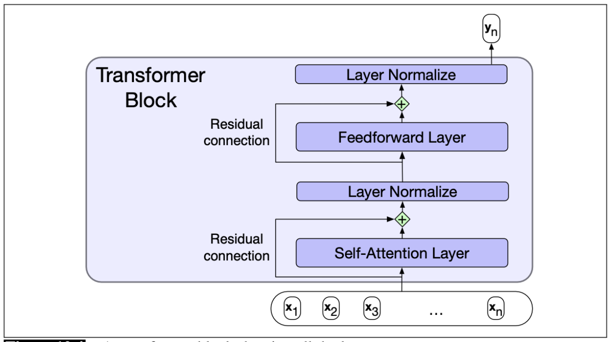


Figure 6.3: Example of a Transformer Block (Jurafsky & Martin, 2023, Ch. 10)

6.3.3 Transformer-XL

Despite the advances of Vaswani et al. (2017)'s self-attention-driven transformer in accounting for long-distance relationships between words, the context used in calculating attention is still fixed. Dai et al. (2019) introduce the Transformer-XL, which removes the fixed-length context window by adding a segment-level recurrence mechanism and a novel scheme for relative positional encodings.

For a new segment, the segment-level recurrence mechanism works by fixing and caching the hidden state sequence computed for the previous segment. This cached hidden state sequence is reused as an extended context when the model processes the next new segment (Dai et al., 2019). This recurrence mechanism is applied to every two consecutive sequences, essentially yielding an unbounded effective context.

In addition to incorporating large contexts, Dai et al. (2019) also implement relative positional encodings to account for the temporal bias of gathering and using information. Previous transformer implementations inserted this bias into the embedding layer; however, the novel scheme for relative positional encodings presented by Dai et al. (2019) inserts that information into the attention score for each layer. They do this by defining the temporal bias based on the distance between the query vector and each key vector. The benefit of this change is summarized as follows:

> By injecting the relative distance dynamically into the attention score, the query vector can easily distinguish the representations of $x_{\tau,j}$ and $x_{\tau+1,j}$ from their different distances, making

> the state reuse mechanism feasible. Meanwhile, we won't lose any temporal information, as the absolute position can be recovered recursively from relative distances. (Dai et al., 2019, p. 5)

6.4 Language Models in the Present Study

6.4.1 TXL

The syntax-less[2] language model used in this study is an implementation of the Transformer-XL (Dai et al., 2019) described above. The model used is 16-layer model with 8 attention heads and 252M parameters.[3]

The model is trained on the BLLIP-LG dataset (Charniak et al., 2000) according to the split by (Hu et al., 2020). The training set is 1.8M sentences (≈40M words). Tokenization is performed using SentencePiece (Kudo & Richardson, 2018) and a subword algorithm (Kudo, 2018) comprised of a 32K word-pieces vocabulary. Prior to training, the Penn-Treebank style linearized trees from BLLIP-LG were converted to Choe-Charniak style linearized trees like that in Figure 6.4.

This model represents a sequential view of language in which language is represented only as surface-level statistical patterns of words co-occurring together.

6.4.2 Transformer Grammar

In contrast, Transformer Grammars (Sartran et al., 2022) provide the syntax-knowledgeable language model used in this study. TGs are a novel transformer implementation that aim to combine the scalability and strong performance of Transformer-XL, the ability to model the joint-probability of a surface string x and its corresponding phrase structure tree y, $p(x, y)$, and an inductive bias that limits the model to explaining data via its built-in recursive syntactic operations. These recursive syntactic operations

[2] As a reminder, the term 'syntax-less' refers to the fact that the TXL model is unconstrained in whether or not it learns some sort of syntactic representation(s) during its training. In contrast, the TG model has a compositional bias imposed that forces it to learn explicit symbolic hierarchical representations. In other words, 'having syntax' is a matter of degree: the syntax-less model *may* learn some implicit syntactic representation, while the syntax-informed model is forced to learn explicit symbolic hierarchy.

[3] This model comes from the OpenSource implementation of TXL and TGs (https://github.com/google-deepmind/transformer_grammars).

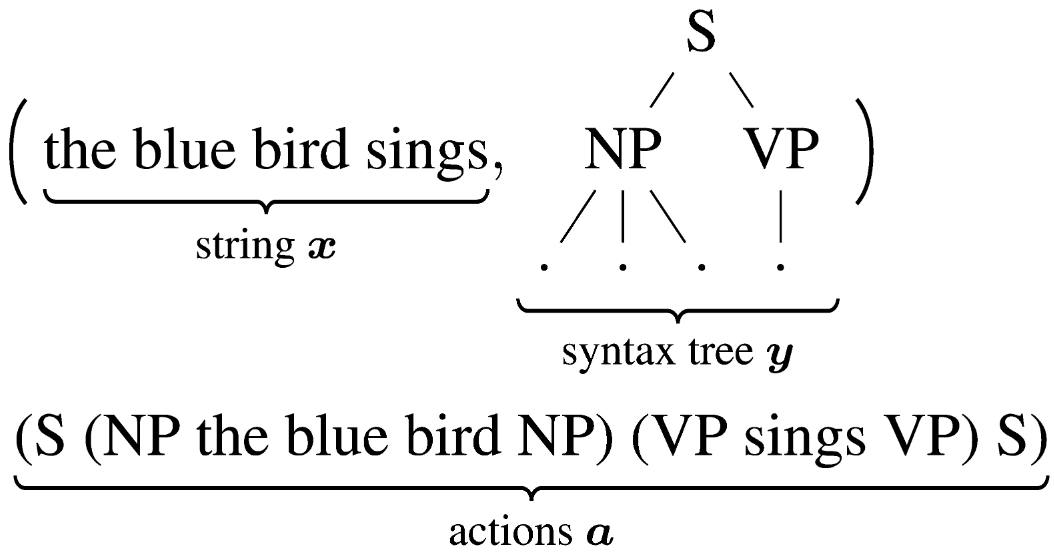


Figure 6.4: An example of a string x and tree y, which are modeled by an action sequence that generates a linearized tree representation of (x,y) (Choe & Charniak, 2016)

are added in by modifying the attention mask of the Transformer-XL, maintaining the scalability and efficiency of the model.

Obtaining the joint-probability of string/tree pairs is done by modeling a sequence of actions that generate the (x,y) pair. This modeling occurs in a top-down left-to-right fashion over an interwoven string of tokens and nonterminal nodes (Choe & Charniak, 2016; Dyer et al., 2016). These interwoven strings are the Choe-Charniak-style linearized trees presented in section 2.3. These linearized trees represent traversing hierarchical trees during the processing of natural language. An example of the string, tree, and action sequence is given in Figure 6.4.

The action sequence of length T is defined as a = (a_0, a_1, ... a_{T-1}). TGs then define a probability distribution such that p(x, y) = p(a) = $\Pi_i p(a_i|a_{<i})$. In other words, TGs model the probability of an action i (a_i) based on the probabilities of all actions taken before action i.

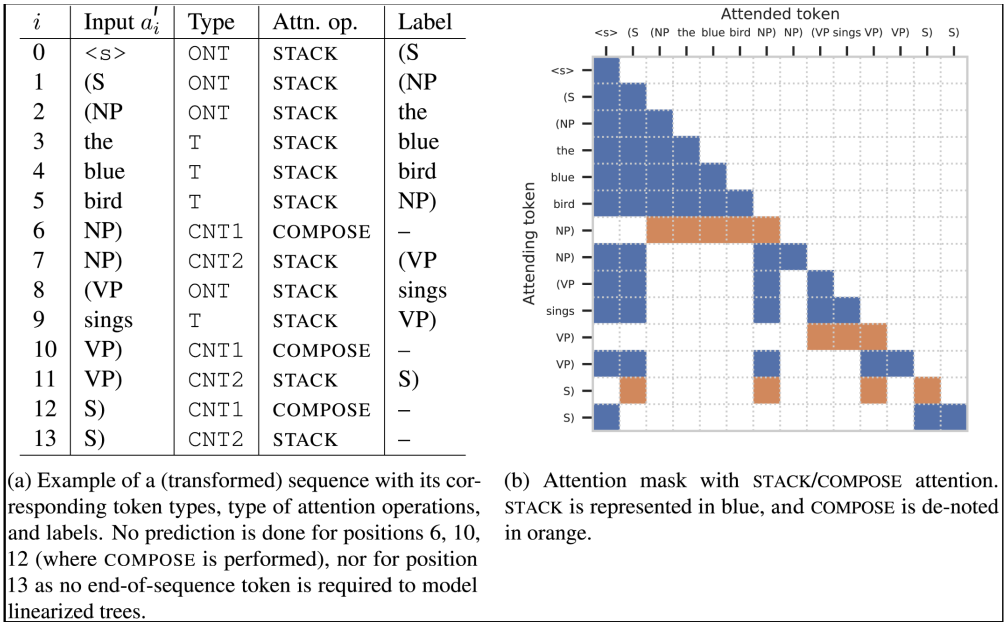
The actions that a TG can take represent a transition system, in which certain actions take place to transition from one symbol to the next. There are three transition actions that a TG can take: Open non-terminal (ONT), generate terminal symbol (T), and close most recently opened non-terminal (CNT).

These actions correspond to processes that take place during human processing, construed as traversing hierarchically structured trees.

When generating a_i on $a_{<i}$, attention, as defined by the attention mask, governs the information flow, restricting what positions can attend to which positions. Critically, the attention masks dictates that after a phrase has been composed (i.e., the CNT actions has taken place), the composed representation is the only element that can be attended to when generating future actions, and the individual words within that composed phrase can no longer be attended to. This ensures that the model will learn informative representations of the composed phrases during training.

TGs implement two types of attention: compose attention and stack attention. Compose attention is used during the process of popping the subconstituents of a composed element off the stack and pushing the composed phrase onto the stack. Stack attention is then applied such that attention only operates on the tokens in the stack. The only difference between these two types of attention is what can be attended to: subconstituents of a phrase before composition in compose attention, and only the composed phrases in stack attention. These two types of attention necessitate the duplication of closing non-terminals in order to only apply one attention mechanism per token, and no prediction is made following for compose attention to keep the number of predictions constant. An example of these mechanisms is shown in Figure 6.5.

As shown in the figure, before a phrase has been composed, it can attend to subconstituents; however, after compose attention has been applied, stack attention can only attend to open non-terminals or composed phrases. For example, when the first closing NP node is reached in line 6, compose attention is applied and only the subconstituents shown in orange are attended to. In line 7, stack attention is applied and only the open non-terminals that have been pushed to the stack can be attended to. When the opening non-terminal in line 8 is opened, only the opening non-terminals and composed NP can be attended to. Again, this implementation of the attention mask creates a syntactic bottleneck that encourages the model to learn informative representations of the composed phrases. These mechanisms are the implementation of syntax in language modeling.



i	Input a'_i	Type	Attn. op.	Label
0	<s>	ONT	STACK	(S
1	(S	ONT	STACK	(NP
2	(NP	ONT	STACK	the
3	the	T	STACK	blue
4	blue	T	STACK	bird
5	bird	T	STACK	NP)
6	NP)	CNT1	COMPOSE	–
7	NP)	CNT2	STACK	(VP
8	(VP	ONT	STACK	sings
9	sings	T	STACK	VP)
10	VP)	CNT1	COMPOSE	–
11	VP)	CNT2	STACK	S)
12	S)	CNT1	COMPOSE	–
13	S)	CNT2	STACK	–

(a) Example of a (transformed) sequence with its corresponding token types, type of attention operations, and labels. No prediction is done for positions 6, 10, 12 (where COMPOSE is performed), nor for position 13 as no end-of-sequence token is required to model linearized trees.

(b) Attention mask with STACK/COMPOSE attention. STACK is represented in blue, and COMPOSE is de-noted in orange.

Figure 6.5: An example of the TG attention mask and operations (Sartran et al., 2022, Figure 2)

The TG language model used in this study is an implementation of the TG just described. Like the TXL model described above, the TG model used is 16-layer model with 8 attention heads and 252M parameters.[4]

Just like the TXL model, the TG model is trained on the BLLIP-LG dataset (Charniak et al., 2000) according to the split and parsing done by (Hu et al., 2020). The training set is comprised of 1.8M sentences (≈40M words). Tokenization is performed using SentencePiece (Kudo & Richardson, 2018) and a subword algorithm (Kudo, 2018) comprised of a 32K word-pieces vocabulary. Prior to training, the Penn-Treebank style linearized trees from BLLIP-LG were converted to Choe-Charniak style linearized trees like that in Figure 6.4.

In sum, the only difference between the TXL and TG in this study is the additional attention mask. Their number of parameters, layers, attention heads, and training/evaluation data (excluding the annotations used for TG) are identical.

[4] This model comes from the OpenSource implementation of TXL and TGs (https://github.com/google-deepmind/transformer_grammars).

Chapter 7

Materials and Methods

7.1 Materials

In order to test the role of syntax in language comprehension, two materials are necessary: surprisal values and a dataset with empirical data. Word-by-word surprisal values are derived from the TG and TXL language models. Section 7.1.1 describes the process for calculating surprisal values from the two language models. The empirical fMRI data used here comes from an open source repository (Li et al., 2022) collected from participants while they engaged in the naturalistic task of listening to an audiobook recording of the children's story *The Little Prince*. Section 7.2 further describes this fMRI dataset.

7.1.1 Surprisal

In order to get surprisal, the transition probability from one word to the next must be calculated. The formulation of surprisal in terms of prefix probabilities from chapter 3 is repeated below for convenience. Surprisal is the $log_2(1/P(y)$ where y is the transition probability to the current word. The transition probability of a word is the ratio of the prefix-probability for the substring before the current word (denoted below as $\sum$ before) and the prefix-probability for the substring including the current word (denoted below as $\sum$ after). The prefix-probability is calculated by multiplying the individual conditional probabilities for each word in the substring, and those individual probabilities come from the language model.

Let y be the ratio of prefix probabilities

$$log_2(\frac{1}{P(y)})$$

$$= log_2(\frac{1}{\frac{\sum \text{after}}{\sum \text{before}}})$$

$$= log_2(\frac{\sum \text{before}}{\sum \text{after}})$$

$$= -log_2(\frac{\sum \text{after}}{\sum \text{before}})$$

Figure 7.1: Surprisal of Prefix-String Probabilities

One more consideration is in order before turning to the exact calculation of surprisal from our language models. The probabilities for each word produced by both TG and TXL are given as the logarithm of the conditional probability (log probability). As described so far, prefix probability is calculated by multiplying conditional probabilities. To perform an equivalent operation in logarithm space, the log probabilities of words are added together rather than being multiplied together. This equivalence is based on the following rule regarding logarithms:

$$log_2(A * B) = log_2(A) + log_2(B)$$

If we consider A the conditional probability of word 1 and B as the conditional probability of word 2, then $log_2(p(\text{word1}))$ + $log_2(p(\text{word2}))$ is equivalent to $log_2(p(\text{word1}) * p(\text{word2}))$. This gives the logarithm of the prefix probability. We will return to conversions between conditional probability and logarithm probability space below.

7.1.2 Deriving Surprisal from TXL

Word-by-word surprisal values are derived from TXL for each word in the *Little Prince* (LPP) story using a pipeline in Python. The (cleaned-up) output from TXL for the first sentence of LPP is given in Table 7.1.

Table 7.1: Log-probabilities from TXL for the first sentence of LPP

	Input	Label	Log prob
1	<s>	Once	-9.51
2	Once	when	-12.29
3	when	I	-11.50
4	I	was	-12.69
5	was	six	-11.83
6	six	years	-10.75
7	years	old	-11.55
8	old	I	-11.50
9	I	saw	-8.91
10	saw	a	-13.09
11	a	magnificent	-13.87
12	magnificent	picture	-6.36
13	picture	in	-5.99
14	in	a	-13.09
15	a	book	-11.28
16	book	about	-5.81
17	about	the	-11.50
18	the	prime	-12.64
19	prime	val	-13.92
20	val	forest	-12.17
21	forest	called	-11.75
22	called	Real	-12.33
23	Real	life	-11.68
24	life	Stor	-13.82
25	Stor	ies	-12.06
26	ies	(no prediction)	0.00

As can be seen in the table, the outputs for the first sentence are a combination of full words and subword tokens, as tokenized by Sentencepiece (Kudo & Richardson, 2018). Subword tokens are combined and matched to the full words in LPP, and log probabilities for the full words are calculated by adding together the log probabilities for all subword tokens that combined to create the full token.

The log of the prefix-probability for each token i is then calculated by adding together the log probability for token i and the log probability for all previous tokens $< i$. Each of these log prefix-probabilities is then exponentiated to get the prefix-probability. To get surprisal for each token i, we take the $-log$

of the ratio of the prefix probability of token i divided by the prefix probability of token $i - 1$. This process yields the surprisal values for the syntax-less language model that are then tested as a predictor of the BOLD signal in the fMRI data.

7.1.3 Deriving Surprisal from TG

Calculating the probability for just a string x is a straightforward computation for the language model, as there is only one probability for said string; however, calculating the probability for the joint string-tree p(x, y) pair is more complicated because the cardinality of the set of possible trees is infinite and therefore exact calculation of probability is impossible. To overcome this impossibility, we can approximate p(x) by using a reduced proposal distribution of trees. Let p(x) be defined as the marginal distribution p(x) = $\sum_{y \in Y'} p(x, y)$ where Y' is a subset of all possible trees. In line with the experiments done by Sartran et al. (2022), we limit the subset of possible trees to the 300 most probable trees for a given string x. This process is referred to as 'marginalizing' over the trees.

The 300-tree proposal distribution used here are the 300 most probable trees as generated by Noji and Oseki's (2021) supervised RNNG (also trained on BLLIP-LG). For each sentence of LPP, 300 trees were generated. These 300 trees per sentence were then pre-processed to convert them from the Penn-Treebank style linearized trees to the Choe-Charniak style linearized trees. Each of these trees were then scored by the TG, the partial output of which for the first tree of the first sentence of LPP is given in Table 7.2.

The process for calculating the log of the prefix probability for each word for each individual tree per sentence is identical to the process described above; however, the prefix probability for each word is not only the summation of the log prob for all tokens beforehand. Rather, the summation also includes all parse actions taken to get from word $i - 1$ to word i.

The process for computing the prefix probability for each word over the entire probability distribution, the 300 trees, is more complex. Remember that the prefix probability for a given token i is $\sum P(\text{derivation } 1_i), P(\text{derivation } 2_i)...P(\text{derivation } 300_i)$. For the TXL model there was only one assigned probability for each string; however, for the TG model, there are 300 derivations, each with its own probability: each tree in the proposal distribution.

Computing the prefix probability and, in turn, surprisal for each word over the entire probability distribution can be carried out in two ways, one which we will refer to as 'Manual' calculation and the

Table 7.2: Log-probabilities from TG for part of the first sentence of LPP

	Input	Label	Log prob
1	<s>	(S	-0.10
2	(S	(ADVP	-8.28
3	(ADVP	Once	-18.00
4	Once	ADVP)	-12.47
5	ADVP)	(no prediction)	0.00
6	ADVP)	(SBAR	-7.21
7	(SBAR	(WHADVP	-11.31
8	(WHADVP	when	-15.12
9	when	WHADVP)	-18.06
10	WHADVP)	(no prediction)	0.00
11	WHADVP)	(S	-0.10
12	(S	(NP	-3.41
13	(NP	I	-14.09
14	I	NP)	-12.10
15	NP)	(no prediction)	0.00
16	NP)	(VP	-8.58
17	(VP	was	-15.72
18	was	(ADJP	-9.98
19	(ADJP	(NP	-3.41
20	(NP	six	-17.60
21	six	years	-19.79
22	years	NP)	-12.10
23	NP)	(no prediction)	0.00
24	NP)	old	-7.99
26	old	ADJP)	-14.04
27	ADJP)	(no prediction)	0.00
28	ADJP)	VP)	-12.68
29	VP)	(no prediction)	0.00
30	VP)	S)	-14.17

other 'LogSumExp' calculation. Employing the manual method, we exponentiate the log of the prefix probability for each word for each of the 300 derivations to get the prefix probability for each derivation. This yields 300 partial prefix probabilities for each word, one per derivation. These conditional probabilities are summed, yielding the prefix probability over the entire probability distribution. This looks as follows:

$$\text{Prefix Probability}_i = \sum \exp(\text{LogProb}(\text{derivation}1_i))...\exp(\text{LogProb}(\text{derivation}300_i))$$

Surprisal of word i then is:

$$-log_2(\frac{\text{Prefix Probability}_i}{\text{Prefix Probability}_{i-1}})$$

LogSumExp is a helper function in Python from the SciPy package (Virtanen et al., 2020) which exponentiates a number of arguments, sums them, then takes the log of this sum. The LogSumExp of a word i looks as follows:

$$\text{LogSumExp}_i = \sum \text{Log}(\exp(\text{LogProb}(\text{derivation}1_i))...\exp(\text{LogProb}(\text{derivation}300_i)))$$

Based on the following property of logarithms:

$$log_2(A/B) = log_2(A) - log_2(B)$$

Surprisal of word i is:

$$= \text{LogSumExp}_{i-1} - \text{LogSumExp}_i$$

Both methods presented above successfully calculate surprisal over the entire probability distribution in slightly different ways. The LogSumExp of a word i gives the log of the prefix probability over the entire probability distribution, while the manual conversion gives the prefix probability itself. Calculating surprisal for word i using LogSumExp means subtracting the -LogSumExp (i.e., adding) of word $i-1$ from the -LogSumExp for word i, whereas the manual method takes the -log of the ratio of the prefix probability for word i and word $i-1$. The derivation in Figure 7.2 illustrates the equivalence of surprisal values calculated by each of these two methods.

Illustrating the equivalence of the two methods for calculating surprisal is useful for future work as it highlights that LogSumExp can be used to calculate surprisal from log probabilities equivalently, but more efficiently, than manually converting to conditional probabilities. Increasingly, language models output the log probabilities of tokens rather than the conditional probabilities, which are the canonical inputs for calculating surprisal. Therefore, using the LogSumExp function offers a more computationally efficient and succinct method for computing surprisal that is of increasing relevance — all of which results in cleaner, more usable, and more readable code for researchers.

The method(s) above yield the surprisal values for the syntax-knowledgeable language model (TG) that are then tested as a predictor of the BOLD signal in the fMRI data described in the next section.

7.2 fMRI Data

The empirical psychometric data used in this study is the blood oxygen level dependent (BOLD) signal, collected using functional magnetic resonance imaging (fMRI). The fMRI dataset used here is the English section of the Little Prince Datasets (Li et al., 2022). 49 right-handed participants (30 female, mean age = 21.3, SD = 3.6) were scanned while they engaged in the naturalistic task of passively listening to an audiobook recording of David Wilkinson's English translation of *The Little Prince*, read by Karen Savage. All participants were compensated for their participation and gave written informed consent prior to participation, in accordance with the IRB guidelines of Cornell University.

The audiobook recording was 94 minutes long and scanning for each participant was done in one session, broken up into 9 runs of about 10 minutes each. Participants listened passively to the story and ended each run by answering 4 comprehension questions about what they had heard in order to ensure

Consider the Division property of logarithms:

$$log(A/B) = log(A) - log(B)$$

Let $\text{ManualSurprisal}_i = -log(\frac{\text{Prefix Probability}_i}{\text{Prefix Probability}_{i-1}})$

Let $\text{LogSumExpSurprisal}_i = -\text{LogSumExp}_i + \text{LogSumExp}_{i-1}$

Given that:

$$\text{Prefix Probability}_i = \sum \text{exp}(\text{LogProb}(\text{derivation}1_i))...\text{exp}(\text{LogProb}(\text{derivation}300_i))$$

And

$$\text{LogSumExp}_i = \sum \text{Log}(\text{exp}(\text{LogProb}(\text{derivation}1_i))...\text{exp}(\text{LogProb}(\text{derivation}300_i)))$$

Then:

$$-\text{Log}(\text{Prefix Probability}_i) = -\text{LogSumExp}_i$$

Then:

$$\text{LogSumExpSurprisal}_i = -\text{Log}(\text{Prefix Probability}_i) + \text{Log}(\text{Prefix Probability}_{i-1})$$

And:

$\text{ManualSurprisal}_i = -\text{Log}(\text{Prefix Probability}_i) + \text{Log}(\text{Prefix Probability}_{i-1})$
by the Division property of logarithms.

Therefore:

$$\text{ManualSurprisal}_i = \text{LogSumExpSurprisal}_i$$

Figure 7.2: Illustration of equivalence in calculating surprisal via manual calculation and LogSumExp calculation

proper attention was being given to the task. Participants were given a break between each session in which they were instructed to relax but not move in the fMRI machine. The total time of the session lasted around 2.5 hours.

7.3 Methods

7.3.1 GLM Analysis of fMRI Data

When a neuron is activated in the brain, it requires oxygenated blood, thus neuron activation necessitates an increase in oxygenated blood. The 'hemodynamic response function' (HRF) describes this blood flow increase following neuronal activation. In other words, the HRF is the (extremely small) response to a stimulus. The HRF generally rises 1-2 seconds following stimulus onset, peaking between 4-6 seconds, and returning to its baseline 12-20 seconds after stimulus onset (Poldrack et al., 2011).

When neuronal activity increases and subsequently oxygenated blood flow increases, the concentration of oxygenated hemoglobin also increases. This oxygenated hemoglobin has magnetic properties that are able to be captured by fMRI, resulting in an increase in the T2*-weighted magnetic resonance signal. This signal, measuring the change in oxygenation via the magnetic properties of oxygenated hemoglobin, is called the 'blood oxygen level dependent' (BOLD) signal. The BOLD signal is directly related to the HRF, allowing for an indirect measure of neuronal properties in response to stimuli.

When performing fMRI analysis, the change, in response to the stimulus, in BOLD signal at each voxel is measured.[1] We use a generalized linear model (GLM) to identify and fit this change in response to the stimulus. In the GLM, the observed time course of the BOLD signal during stimuli presentation is the dependent variable and any predictors, fit to the timecourse of the presentation of the stimuli, included in the model are the independent variables.

7.3.2 r^2 Model Comparison

To probe the brain bases of hierarchical syntax, we pursue an r^2 analysis, following Crabbé et al. (2019, §5).

[1] A voxel is a three dimensional parcellation, and can be thought of like a point on a coordinate plane, only the coordinate plane has 3 dimensions instead of 2.

Subject-Level Statistics

For each subject, we calculate how much the inclusion of the variables of interest—TG surprisal and TXL surprisal—increases cross-validated BOLD r^2 with respect to a base model with only predictors of non-interest.

Here, r^2 values indicate the voxel-wise variance explained. Thus, at the first level, two brain maps are calculated for each participant: one indicating the increase in cross-validated r^2 (how well the model explains the BOLD signal) associated with adding TG surprisal to a baseline model; and one indicating the increase in cross-validated r^2 associated with adding TXL surprisal to a baseline model. These models were cross-validated such that the model for each participant was trained on the first 8 sections of the story and tested on the 9th. This cross-validation ensures that the models are not overfitted to the data.

BOLD signal is modeled, at each voxel, for each participant, via a generalized linear model (GLM). The word-level metrics are temporally annotated at the offset of each word in the audiobook, while the speech-related metrics are annotated every 10ms. All regressors, described in Table 7.3, were convolved with the SPM canonical hemodynamic response function (Poldrack et al., 2011). As mentioned above, the BOLD signal lags behind the neuronal response by multiple seconds; convolution with the HRF accounts for this temporal difference, ensuring a good fit to the BOLD time series (Poldrack et al., 2011). All predictors were standardized (shifted to mean 0 and scaled to standard deviation 1) by scanning session and storybook section (the 9 runs mentioned above) prior to modeling.

Regressors of non-interest are included to ensure that any effects found are not due to other facets of linguistic processing (Lund et al., 2006). One notable included regressor is the lexical-semantic control. These lexical-semantic encodings are word vectors that come from a 300-dimension pre-trained English fastText model (Bojanowski et al., 2016), reduced to a dimensionality of 5. This regressor is included to verify any contribution of hierarchical structure independent from that of lexical semantics.

Throughout the GLM analysis, a liberal cortical mask (https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation_Yeo2011) was applied. This cortical mask, based on the 1,000 participants in experiments by Buckner et al. (2011) and Yeo et al. (2011), ensures that the GLM analysis is limited to cortical voxels.

Table 7.3: GLM Model Predictors

Predictor	Description	Model-Inclusion
TG Surprisal	Surprisal derived from TG at a word	Syntax-informed
TXL Surprisal	Suprisal derived from TXL at a word	Syntax-less
Word Rate	Annotation indicating the existence of a spoken word	Base, Syntax-informed, Syntax-less
Word Frequency	Log lexical frequency of a word	Base, Syntax-informed, Syntax-less
F_0	Pitch (fundamental frequency) of the voice of the narrator	Base, Syntax-informed, Syntax-less
RMS Amplitude	Root Mean Square Amplitude of the voice of the narrator (reflecting intensity)	Base, Syntax-informed, Syntax-less
Word Vector_5	5 regressors corresponding to values derived from a word's pretrained fastText vector	Base, Syntax-informed, Syntax-less

Group Level Statistics

The single-subject r^2 brain maps (one TG map; one TXL map) were entered into a paired T-test to compare the impact of the additions of TG surprisal and TXL surprisal to base models on the BOLD signal. The results of this test indicate where the addition of one variable to the base model (either TG surprisal or TXL surprisal) contributes to explaining the BOLD signal significantly better than the other.

Chapter 8

Results and Discussion

8.1 Results

The results of the procedure described in the previous section are as follows: the syntax-informed model performed above-and-beyond the syntax-less model in goodness-of-fit (r^2 values) to the measured BOLD signal timecourse in the right medial temporal gyrus (rMTG; BA21), the left pars opercularis (Broca's Area; BA44), the left temporal pole (BA38), and the right pre-frontal cortex (rPFC; BA10). These findings can be seen in Figure 8.3 and more details are given in Table 8.1. There were no regions in which the syntax-less model performed above-and-beyond the syntax-informed model. Intermediate results for TXL and TG are given in figures 8.1 and 8.2

The significant clusters found from the paired T-test were corrected using false discovery rate (FDR) $< .05$ and required a cluster size of at least 50 significant voxels. FDR controls for the the amount of type I errors, in which the null hypothesis is incorrectly rejected, allowing for increased confidence in our results.

8.2 Discussion

The findings presented here are consistent with a number of previous findings investigating hierarchical structure in the brain (e.g., Brennan et al., 2012; Hale et al., 2018; Shain et al., 2020; Stanojević et al., 2023), as well as large scale brain models of language.

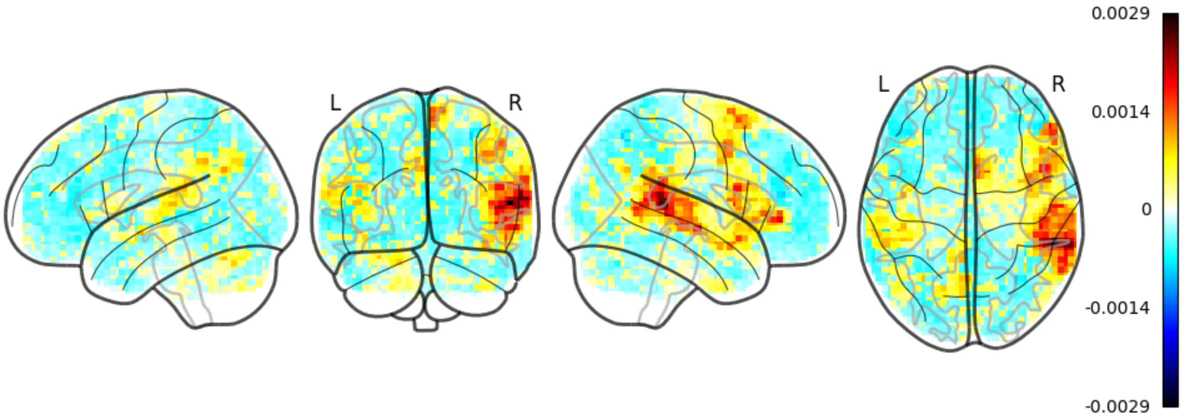


Figure 8.1: z-maps (with no thresholding) showing the r^2 increase for the syntax-less model compared to the base model.

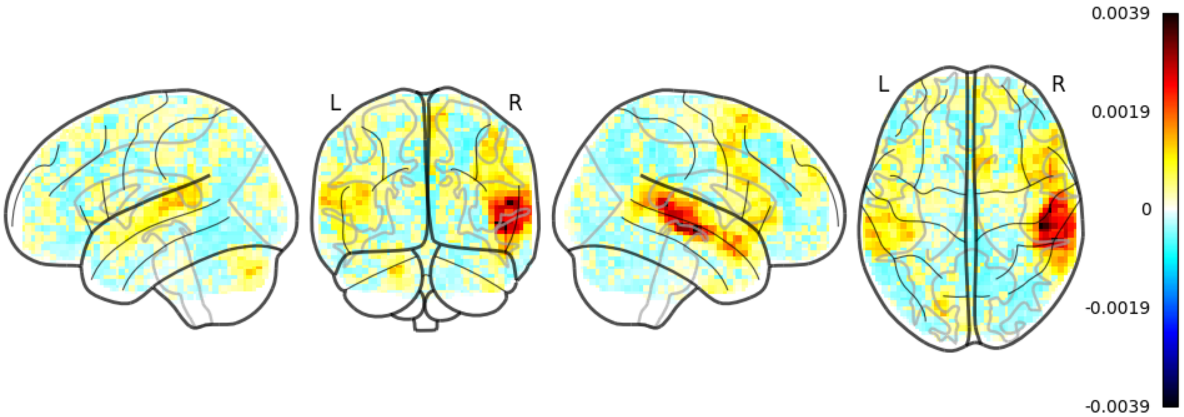


Figure 8.2: z-maps (with no thresholding) showing the r^2 increase for the syntax-informed model compared to the base model

Table 8.1: Results of paired T-test between syntax-informed and syntax-less cross-validated r^2, thresholded with an expected false discovery rate (FDR) < 0.05 and a cluster threshold of 50 voxels.

		MNI Coordinates			
Region	**Cluster size (mm^3)**	**X**	**Y**	**Z**	**Peak Stat**
R Medial Temporal Gyrus (BA21)	5960	52.0	-28.0	-6.0	4.96
L Pars Opercularis (Broca's Area; BA44)	712	-42.0	14.0	28.0	4.74
		-42.0	4.0	34.0	3.82
L Temporal Pole (BA38)	424	-48.0	12.0	-18.0	4.17
R Pre-frontal Cortex (BA10)	464	14.0	62.0	22.0	3.92

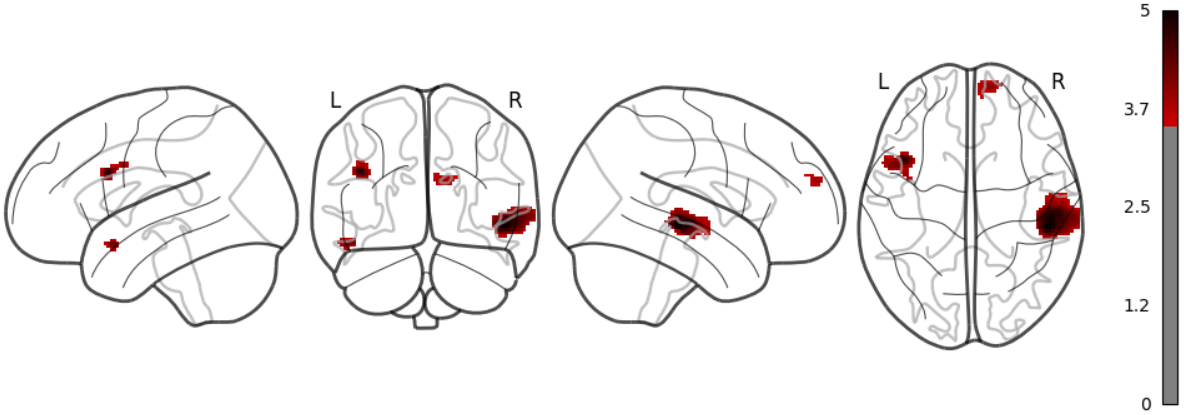


Figure 8.3: Results (z-valued) of paired T-test between syntax-informed and syntax-less cross-validated r^2, thresholded with an expected false discovery rate (FDR) < 0.05 and a cluster threshold of 50 voxels.

8.2.1 Large Scale Brain Models of Language

When isolating brain regions that subserve specific aspects of linguistic processing, it is important to remember that these computational processes are part of a large dynamic network of language-related brain regions that all interact with each other. The goal is to determine the contributions of specific regions as part of this larger, dynamic network.

Complex Structure in *The Little Prince*

Many large-scale brain models of language posit different brain regions that subserve basic and complex syntactic processing. One example of complex structure is quotation. Quotation, a form of indirect speech, is a complex structure due to its multiple-embedding of S levels (exemplified in Figure 8.4). One interpretation of the rMTG result is that this region subserves quotation processing. This interpretation aligns with the results of Crabbé et al. (2019) and Wehbe et al. (2014), who investigate how the parser size of a beam and syntax, respectively, correlate with fMRI data during language processing. The stimuli used by Crabbé et al. (2019) are the same as those used here, while Wehbe et al. (2014) had participants read Chapter 9 of *Harry Potter and the Sorcerer's Stone* (Rowling, 2012). Both sets of stimuli are rife with quotation. *The Little Prince* contains 6,712 words that are in direct quotations, 44% of the total words in the story (15,389). Chapter 9 of Harry Potter contains 2,005 words in direct quotation, 39% of the total words in the chapter (5,137).

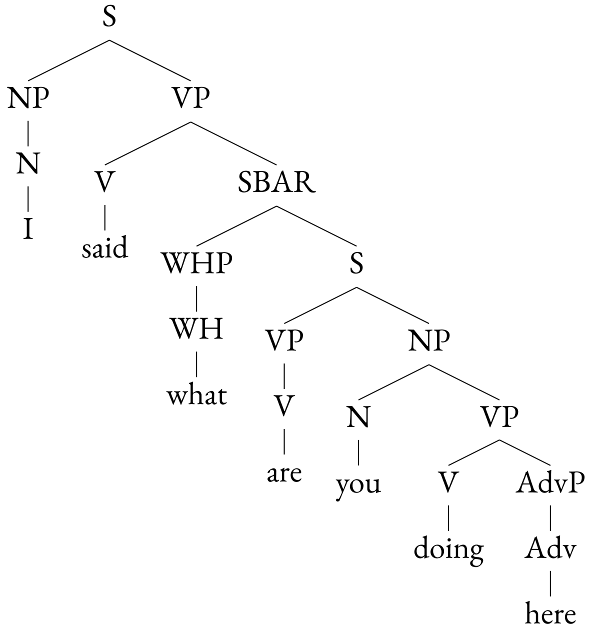


Figure 8.4: An example of indirect speech that is attested in the stimulus text (*The Little Prince*) used in the fMRI study.

We posit that the complex structure building associated with this quotation in-part drives the findings in the rMTG across these three studies. Indeed, this explanation is further supported by the findings of Bašnáková et al. (2014), who compared brain activation of participants using fMRI while they listened to indirect speech and direct control utterances. They found that indirect speech engaged a number of structures, notably the PFC, the rMTG, and bilateral IFG, compared to the direct control utterances.

The Neuroanatomical Pathway Model of Language

The 'Neuroanatomical Pathway Model of Language' (Friederici, 2017; Friederici & Gierhan, 2013) posits 4 pathways, 2 ventral and 2 dorsal, in the left hemisphere that constitute the language network, which subserves linguistic comprehension and production. These pathways refer to brain regions that are connected by different white matter fiber tract bundles. Two of these pathways are directly involved with syntactic structure building: the ventral pathway from the frontal operculum to the anterior superior temporal gyrus (aSTG) and the dorsal pathway from the pars opercularis (BA44) to the posterior temporal gyrus/sulcus (pSTG/pSTS). This model argues that the frontal operculum and aSTG, which are connected by the uncinate fasciculus (UF), are involved in local structural processing. A possible specific explanation is that the aSTG gets information about local phrases from the auditory cortex, representations of which are made available in aSTG, including BA38, upon reaching a phrasal head (e.g., N, V, A). This information is then transferred to the frontal operculum via the UF. This information is further transferred to BA44 if additional processing is necessary. In sum, this ventral network is "responsible for the most basic syntactic processes, that is, local syntactic computations" (Hickok & Small, 2016, p. 352). The current findings in BA38 are consistent with the role Friederici assigns to this pathway. TGs capture local syntactic computation for each phrase, which is the exact purpose of the aSTG and the frontal operculum posited by this brain model.

The second, dorsal, syntactic network deals with global computation, including sentence-level hierarchy and syntactically complex sentences, such as those with embedding. This network includes BA44, which supports building hierarchical structure with non-adjacent dependencies, and the pSTG/STS, which integrates syntactic and semantic information into this structure. The current findings in BA44 could be explained by this dorsal pathway, as TGs capture the exact non-adjacent dependencies and complex structure building supported by this network. The lack of a significant result for TG-derived surprisal in the pSTG/STS could be because the contribution of these regions in semantic integration has been regressed out by the lexical-semantic predictors, or that TGs and TXLs both capture this semantic integration.

As described by this large-scale brain model, language-related functions are largely lateralized to the left hemisphere; however, language-related functions often implicate homologues in the right hemisphere (see e.g., Stowe et al., 2005 for a review or Vigneau et al., 2011 for a meta-analysis). A number of studies have specifically implicated the rMTG in complex structure building (e.g., Crabbé et al., 2019; Wehbe et al., 2014). It is possible that current finding in the rMTG is due to the complex structure building associated with quotation due to its multiple embedding. The left and right temporal lobes are connected via the corpus callosum fiber tract (Goldstein et al., 2024), which could facilitate the transfer of complex structural information between the two regions. Indeed, the left MTG, the rMTG, and the pSTS have been shown to be functionally connected (e.g., Dronkers et al., 2004; Turken & Dronkers, 2011). This additional recruitment of the right hemisphere also aligns with the theory of Just and Varma (2007), which describes how different brain centers are dynamically recruited when handling resource-heavy demands, such as complex structure building.

Memory, Unification and Control (MUC)

The Memory, Unification, and Control (MUC; Hagoort, 2005, 2013) model posits three functional components of language: Memory, Unification, and Control. Memory is the only domain-specific function. Related to language, memory refers to the storage of linguistic knowledge and the building blocks of structure (e.g., phrasal heads). This component is subserved by the temporal cortex. The current results in BA38 could be explained by the fact that TG captures information about linguistic building blocks. Phrasal head information is captured by TGs in their specific phrasal composition.

The unification component refers to the combination of elements from the memory component in a meaningful way. This component is subserved by left inferior frontal cortex (LIFC), including BA44. In the MUC model, unification occurs over multiple levels of representations, including syntax. The current findings in BA44 are directly in line with this purported purpose of LIFC.

The current findings in the rMTG are not directly explainable by the MUC model; however, this model does not specify different brain regions subserving basic and complex structure building. The findings in the rMTG could still align with the MUC model if additional brain regions are recruited during particularly complex unificational processes.

An alternate explanation for the findings in the rMTG comes from research on empathy (de Greck et al., 2012; Jie et al., 2021; Rankin et al., 2006, e.g.,). These studies find a distinct correlation between the rMTG and measures of empathy, indicating that higher levels of empathy result in greater activation of the rMTG. Given that *The Little Prince* is a story about a young boy learning about the world by traveling the universe, teaching the boy and the reader about love, loss, friendship, and loneliness. These themes are likely to induce increased empathy in participants as they listened to the storybook, which could explain the findings in the rMTG. This explanation leaves unclear how the addition of constituency in the syntax-knowledgeable model better captures higher levels of empathy than the purely sequential syntax-less model.

8.2.2 Methodological Improvement

The methodology employed throughout this study marks an improvement over previous studies comparing language models with syntax to those without, discussed in chapter 3. For example, Brennan and Hale (2019) compare a trigram model and simple recurrent neural network against a phrase-structure context-free grammar. In contrast, the only difference between TG and TXL is the additional attention mask in TG. This additional attention mask biases the Transformer toward a more hierarchical mode of operation. Put simply, the architectures of TG and TXL only differ in whether or not 'syntax' is turned on. This comparison is methodologically cleaner than the ad-hoc comparison between models with vastly different underlying architectures. Despite this ad hoc comparison, their results ultimately align with those presented here.

The results presented here align with those previously found for different implementations of syntax: dependency grammar (Lopopolo et al., 2021), bag of syntactic features (Wehbe et al., 2014), CCG (Stanojević et al., 2023), and long-left contexts (Toneva et al., 2022). This convergence strengthens the classic view of the brain's language network as including specialized component(s) that support hierarchical syntactic processing (Friederici, 2017; Friederici & Gierhan, 2013; Hagoort, 2005, 2013).

8.2.3 On The Scope of This Project

There is great debate about LLMs acting as feasible computational frameworks for understanding human linguistic processing and the neurobiology of language (see Millière, 2024, §IV.ii, for a review). The results of this study do not assess the potential adequacy of LLMs in eventually achieving this end; however, the results do indicate that LLMs have not yet achieved this goal.

The surprisal values derived from the TG and TXL language models for *The Little Prince* here are heavily correlated with each other (r = 0.70), indicating that TXL, in its unconstrained learning, has learned *something* about syntax; however, the fact that TG-derived surprisal still better correlates with the BOLD signal in a number of brain regions indicates that whatever, and however, the TXL model is learning does not match human processing as well as TG. The only difference between the TG and TXL is the compositional bias in TG. The success of TG in modeling human neural data then could indicate that biasing the TG model toward these explicit symbolic hierarchical representations better constitutes the type of linguistic processing that humans do. In other words, LLMs have the *potential* to be adequate models of human linguistic processing, possibly with additional symbolic constraints or biases; however, they do not currently achieve this end.

8.2.4 Summary: In Support of Hierarchical Structure

TG-derived surprisal significantly correlated with the BOLD signal in BA21, BA44, BA38, and BA10, illustrating that biasing the Transformer toward explicit symbolic syntactic representations (constituency) improves its correlation in these regions. The identification of specific brain regions related to processing hierarchical structure pushes back against recent arguments for a distributed cortical network in which lexico-semantic processing and structure building operations are not dissociable (Caucheteux et al., 2021; Fedorenko et al., 2012, 2020). The generalized linear models used here to model brain activity control for lexical semantics via word-embeddings (Bojanowski et al., 2016). Despite this control, syntactic processing in comprehension can still be localized to BA44 and BA38. This result would be surprising if the lexical and syntactic aspects of language processing were truly not dissociable. Null results in earlier neuroimaging studies may be explainable in terms of methodological differences.

Overarchingly, these findings push back against recent arguments (e.g., Contreras Kallens et al., 2023; Goldstein et al., 2022a) for a sequential view of language in which representations are based on surface level word co-occurrence patterns rather than any sort of hidden hierarchical structure and processing is based only on prediction using sequential information.

Chapter 9

Conclusion

This study investigated the role of hierarchical syntactic structure in processing natural language. In particular, it used the Transformer Grammar (Sartran et al., 2022) as the syntax-informed language model from which surprisal values were derived. This is in contrast to the TXL model (Dai et al., 2019), which only models language based on surface-level appearance of words. In deriving these surprisal values, the usefulness of the helper function LogSumExp (Virtanen et al., 2020) was illustrated, in order to accommodate Log-probabilities surprisal (Hale, 2001).

The surprisal values from these models were used as regressors in separate Generalized Linear Models (GLMs) to explore how much they contributed to modeling fMRI data (Li et al., 2022) collected while participants listened to the audiobook narration of *The Little Prince*. Comparing the increase in r^2 values of each model compared to a base model via a paired T-test, the syntax-informed GLM model accounts for variance in the BOLD signal timecourse above-and-beyond the syntax-less model in the pars opercularis, the left ATG, the rMTG, and the rPFC.

The results support two conclusions: 1) Broca's Area (pars opercularis) and left anterior temporal gyrus play a major role in structure building, confirming numerous previous results (in line with a number of large scale brain models of language) and 2) right hemisphere homologues, notably the rMTG, might be recruited to assist in difficult complex structure building such as quotation.

Taken together, these results affirm that syntax is part of the best account of processing natural language, as well as the brain regions that subserve this aspect of processing. These findings offer no support for a purely sequential view of language, suggesting instead that language is organized with multiple levels of abstraction, including hierarchical structure, which guide natural language comprehension and directly correspond to neurobiological properties.

Appendix A

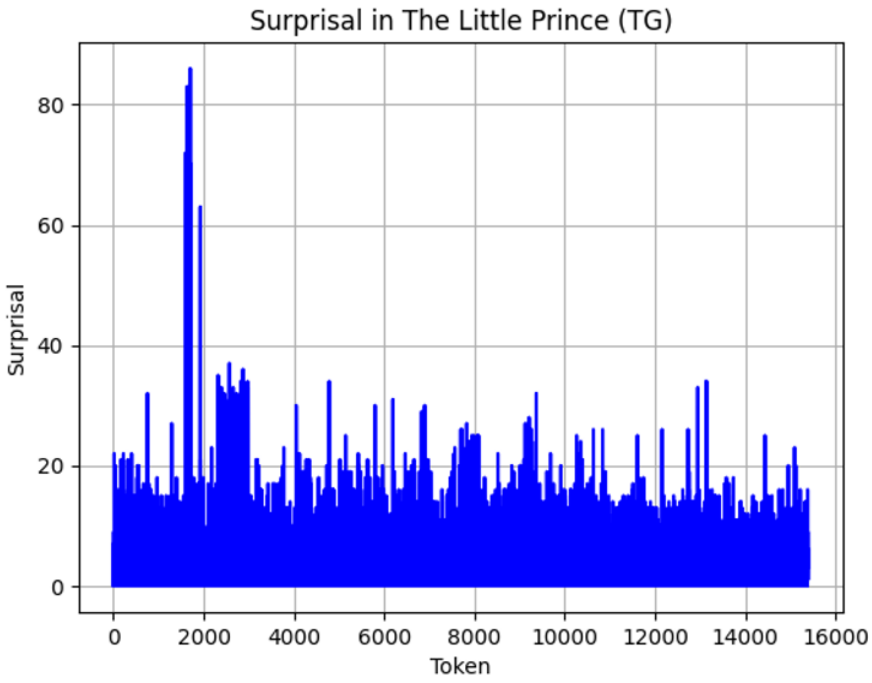


Figure A.1: TG Surprisal for each token in *The Little Prince*

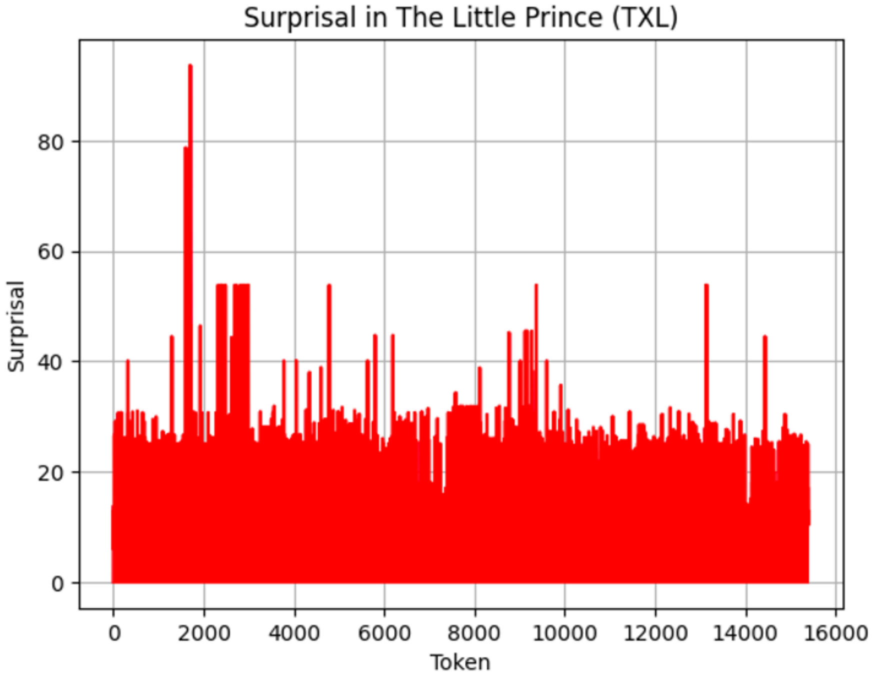


Figure A.2: TXL Surprisal for each token in *The Little Prince*

Bibliography

Adger, D. (2003). *Core syntax. a minimalist approach*. Oxford University Press.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. [Pages: xiv, 524]. V. H. Winston & Sons.

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex (New York, N.Y.: 1991)*, *24*(10), 2572–2578. https://doi.org/10.1093/cercor/bht112

Bates, E., Elman, J., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness : A connectionist perspective on development*. A Bradford Book.

Bever, T. (1970, January). The cognitive basis for linguistic structures. https://doi.org/10.1093/acprof:oso/9780199677139.003.0001

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv:1607.04606*.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus [Number: 1]. *Journal of Eye Movement Research*, *2*(1). https://doi.org/10.16910/jemr.2.1.1

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349. https://doi.org/10.1080/01690965.2010.492228

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173.

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

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), e0207741. https://doi.org/10.1371/journal.pone.0207741

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

Buckner, C., & Garson, J. (2019). The stanford encyclopedia of philosophy. In E. N. Zalta (Ed.). https://plato.stanford.edu/archives/fall2019/entries/connectionism/

Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C., & Yeo, B. T. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(5), 2322–2345. https://doi.org/10.1152/jn.00339.2011

Caucheteux, C., Gramfort, A., & King, J.-R. (2021, July). Disentangling syntax and semantics in the brain with deep networks. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 1336–1348, Vol. 139). PMLR. https://proceedings.mlr.press/v139/caucheteux21a.html

Charniak, E. (1996). *Statistical language learning*. MIT press.

Charniak, E., Blaheta, Don, Ge, Niyu, Hall, Keith, Hale, John, & Johnson, Mark. (2000). BLLIP 1987-89 WSJ Corpus Release 1. https://doi.org/10.35111/FWEW-DA58

Choe, D. K., & Charniak, E. (2016). Parsing as Language Modeling. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2331–2336). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1257

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124. https://doi.org/10.1109/TIT.1956.1056813

Chomsky, N. (1957, May). Syntactic Structures. In *Syntactic Structures*. De Gruyter Mouton. https://doi.org/10.1515/9783112316009

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (50th ed.). The MIT Press. https://www.jstor.org/stable/j.ctt17kk81z

Chomsky, N. (1980). A review of bf skinner's verbal behavior. *The Language and Thought Series*, 48–64.

Chomsky, N. (1986). *Barriers*. https://mitpress.mit.edu/9780262530675/barriers/

Chomsky, N. (1995). *The Minimalist Program*. MIT Press. https://mitpress.mit.edu/9780262531283/the-minimalist-program/

Chomsky, N. (2001, April). Derivation by Phase. In *Ken Hale: A Life in Language*. The MIT Press. https://doi.org/10.7551/mitpress/4056.003.0004

Chomsky, N. (1955/1975). *The Logical Structure of Linguistic Theory*. Springer. https://link.springer.com/book/9780306307607

Chomsky, N. (1981b). Principles and parameters in syntactic theory. In N. Hornstein & D. Lightfood (Eds.), *Explanation in linguistics* (pp. 32–75). Longman.

Chomsky, N. (1970a). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in english transformational grammar* (pp. 184–221). Ginn.

Clark, A. (2001). *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press USA.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Crabbé, B., Fabre, M., & Pallier, C. (2019, November). Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1150–1160). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1106

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019, June). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. https://doi.org/10.48550/arXiv.1901.02860

de Greck, M., Wang, G., Yang, X., Wang, X., Northoff, G., & Han, S. (2012). Neural substrates underlying intentional empathy. *Social Cognitive and Affective Neuroscience*, *7*(2), 135–144. https://doi.org/10.1093/scan/nsq093

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1-2), 145–177. https://doi.org/10.1016/j.cognition.2003.11.002

Dunagan, D., Stanojević, M., Coavoux, M., Zhang, S., Bhattasali, S., Li, J., Brennan, J., & Hale, J. (2023). Neural Correlates of Object-Extracted Relative Clause Processing Across English and Chinese. *Neurobiology of Language*, *4*(3), 455–473. https://doi.org/10.1162/nol_a_00110

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348. https://doi.org/10.1016/j.cognition.2020.104348

Fedorenko, E., Nieto-Castañon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513. https://doi.org/10.1016/j.neuropsychologia.2011.09.014

Fiengo, R. (1977). On Trace Theory. *Linguistic Inquiry*, *8*(1), 35–61. Retrieved December 4, 2023, from https://www.jstor.org/stable/4177972

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Friederici, A. D. (2017, November). *Language in Our Brain: The Origins of a Uniquely Human Capacity*. The MIT Press. https://doi.org/10.7551/mitpress/11173.001.0001

Friederici, A. D., & Gierhan, S. M. E. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain : Why cognitive science will transform neuroscience.* Wiley-Blackwell.

Goldstein, A., Covington, B. P., Mahabadi, N., & Mesfin, F. B. (2024). Neuroanatomy, Corpus Callosum. In *StatPearls*. StatPearls Publishing. Retrieved February 26, 2024, from http://www.ncbi.nlm.nih.gov/books/NBK448209/

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

Goodkind, A., & Bicknell, K. (2018, January). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)* (pp. 10–18). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0102

Graf, T., Monette, J., & Zhang, C. (2017). Relative clauses as a benchmark for Minimalist parsing. *Journal of Language Modelling*, *5*, 57–106. https://doi.org/10.15398/jlm.v5i1.157

Grune, D., & Jacobs, C. J. H. (2008). Introduction to parsing. In *Parsing techniques: A practical guide* (pp. 61–102). Springer New York. https://doi.org/10.1007/978-0-387-68954-8_3

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423. https://doi.org/10.1016/j.tics.2005.07.004

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00416

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*, 1–8. https://doi.org/10.3115/1073336.1073357

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

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018, July). Finding syntax in human encephalography with beam search. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2727–2736). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1254

Hale, J., Lutz, D., Luh, W.-M., & Brennan, J. (2015, June). Modeling fMRI time courses with linguistic structure at various grain sizes. In T. O'Donnell & M. van Schijndel (Eds.), *Proceedings of the 6th workshop on cognitive modeling and computational linguistics* (pp. 89–97). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-1110

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, *8*(1), 427–446. https://doi.org/10.1146/annurev-linguistics-051421-020803

Harris, Z. (1951). *Methods in structural linguistics*. University of Chicago Press.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Henderson, J. M., Choi, W., Lowder, M. W., & Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, *132*, 293–300. aheadofprint. https://doi.org/10.1016/j.neuroimage.2016.02.050

Hickok, G., & Small, S. L. (2016). *Neurobiology of language* (1st ed.). Academic Press.

Hornstein, N., & Nunes, J. (2008). Adjunction, labeling, and bare phrase structure. *Biolinguistics*, *2*, 57–86. https://doi.org/doi:10.5964/bioling.8621

Hornstein, N., Nunes, J., & Grohmann, K. K. (2005). *Understanding minimalism*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840678

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020, July). A Systematic Assessment of Syntactic Generalization in Neural Language Models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT Press.

Jelinek, F., Lafferty, J., & Mercer, R. (1992). Basic methods of probabilistic context free grammars. *Speech Recognition and Understanding, NATO ASI Series*, *75*. https://doi.org/10.1007/978-3-642-76626-8_35

Jie, J., Fan, M., Yang, Y., Luo, P., Wang, Y., Li, J., Chen, W., Zhuang, M., & Zheng, X. (2021). Establishing a counter-empathy processing model: Evidence from functional magnetic resonance imaging. *Social Cognitive and Affective Neuroscience*, *17*(3), 273–289. https://doi.org/10.1093/scan/nsab097

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (1st). Prentice Hall PTR.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87 4*, 329–54. https://api.semanticscholar.org/CorpusID:3793521

Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(3), 153–191. https://doi.org/10.3758/CABN.7.3.153

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*, 170–210. https://api.semanticscholar.org/CorpusID:9860676

Kudo, T. (2018, July). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 66–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1007

Kudo, T., & Richardson, J. (2018, August). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. https://doi.org/10.48550/arXiv.1808.06226

Lakoff, G. (1971). On generative semantics. In D. Steinberg & L. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 232–296). Cambridge University Press.

Langendoen, D. T. (1975). Finite-state parsing of phrase-structure languages and the status of readjustment rules in grammar. *Linguistic Inquiry*, *6*(4), 533–554. Retrieved February 24, 2024, from http://www.jstor.org/stable/4177899

Lappin, R. L., Shalom, & Johnson, D. E. (2000a). He structure of unscientific revolutions. *Natural Language and Linguistic Theory*, 665–771.

Lasnik, H., & Lohndal, T. (2013). Brief overview of the history of generative syntax. In M. den Dikken (Ed.), *The cambridge handbook of generative syntax* (pp. 26–60). Cambridge University Press. https://doi.org/10.1017/CBO9780511804571.004

Lennenberg, E. H. (1967). *Biological foundations of language*. John Wiley; Sons, https://onlinelibrary.wiley.com/doi/10.1002/bs.3830130610

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. Retrieved November 10, 2023, from https://www.sciencedirect.com/science/article/pii/S0010027707001436

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., Brennan, J. R., Yang, Y., Pallier, C., & Hale, J. (2022). Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data*, *9*(1), 530. https://doi.org/10.1038/s41597-022-01625-7

Lopopolo, A., van den Bosch, A., Petersson, K.-M., & Willems, R. M. (2021). Distinguishing Syntactic Operations in the Brain: Dependency and Phrase-Structure Parsing. *Neurobiology of Language*, *2*(1), 152–175. https://doi.org/10.1162/nol_a_00029

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., & Nichols, T. E. (2006). Non-white noise in fMRI: Does modelling have an impact? *NeuroImage*, *29*(1), 54–66. https://doi.org/10.1016/j.neuroimage.2005.07.005

Mandler, G. (2002). Origins of the cognitive (r)evolution. *Journal of the History of the Behavioral Sciences*, *38*(4), 339–353. https://doi.org/10.1002/jhbs.10066

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, *19*(2), 313–330.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

Millière, R. (2024). Language models as models of language. *Oxford Handbook of the Philosophy of Linguistics*.

Moore, J. (2017). John b. watson's classical s–r behaviorism. *The Journal of Mind and Behavior*, *38*(1), 1–34. Retrieved January 29, 2024, from http://www.jstor.org/stable/44631526

Newell, A. (1990). Unified theories of cognition. *Unified theories of cognition.*, xvii, 549–xvii, 549.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, *19*(3), 113–126. https://doi.org/10.1145/360018.360022

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*(3), 447–461. https://doi.org/10.1017/langcog.2016.20

Noji, H., & Oseki, Y. (2021, August). Effective Batching for Recurrent Neural Network Grammars. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics:*

ACL-IJCNLP 2021 (pp. 4340–4352). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.380

OpenAI. (2023). GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774

Pereira, F. C. N., & Wright, R. N. (1991). Finite-state approximation of phrase structure grammars. *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, 246–255. https://doi.org/10.3115/981344.981376

Phillips, C. (2013, August). Language Down the Garden Path: The Cognitive and Biological Basis for Linguistic Structures. In M. Sanz, I. Laka, & M. K. Tanenhaus (Eds.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199677139.001.0001

Pilehvar, M. T., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan Claypool.

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R., & Cotterell, R. (2023, July). On the Effect of Anticipation on Reading Times. https://doi.org/10.48550/arXiv.2211.14301

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). Statistical modeling: Single subject analysis. In R. A. Poldrack, J. A. Mumford, & T. E. Nichols (Eds.), *Handbook of Functional MRI Data Analysis* (pp. 70–99). Cambridge University Press. Retrieved December 13, 2023, from DOI: %2010.1017/CBO9780511895029.006

Putnam, H. (1967). The 'innateness hypothesis' and explanatory models in linguistics. *Synthese*, *17*(1), 12–22. https://doi.org/10.1007/BF00485014

Pylyshyn, Z. W. (1984). *Computation and Cognition : Toward a Foundation for Cognitive Science.* A Bradford Book.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Rankin, K. P., Gorno-Tempini, M. L., Allison, S. C., Stanley, C. M., Glenn, S., Weiner, M. W., & Miller, B. L. (2006). Structural anatomy of empathy in neurodegenerative disease. *Brain*, *129*(11), 2945–2956. https://doi.org/10.1093/brain/awl254

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009, August). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In P. Koehn & R. Mihalcea (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Nat-*

ural Language Processing (pp. 324–333). Association for Computational Linguistics. Retrieved November 27, 2023, from https://aclanthology.org/D09-1034

Rowling, J. (2012). *Harry potter and the sorcerer's stone*. Pottermore Limited.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations by Error Propagation. https://doi.org/10.7551/mitpress/5236.003.0012

Rumelhart, D. E., Hinton, G. E., McClelland, J. L., et al. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*(45-76), 26.

Sartran, L., Barrett, S., Kuncoro, A., Stanojević, M., Blunsom, P., & Dyer, C. (2022). Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational Linguistics*, *10*, 1423–1439. https://doi.org/10.1162/tacl_a_00526

Saussure, F. ([1916] 1959). *Course in general linguistics* (W. Baskin, Trans.). Philosophical Library.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Shannon, C. E. (1948). A mathematical theory of communication [Conference Name: The Bell System Technical Journal]. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Skinner, B. F. (1963). Operant behavior. *American Psychologist*, *18*(8), 503–515. https://doi.org/10.1037/h0045185

Smolensky, P., Legendre, G., & Miyata, Y. (1993). Integrating connectionist and symbolic computation for the theory of language. *Current Science*, *64*(6), 381–391. Retrieved February 6, 2024, from http://www.jstor.org/stable/24098873

Stabler, E. P. (2013). Two Models of Minimalist, Incremental Syntactic Analysis. *Topics in Cognitive Science*, *5*(3), 611–633. https://doi.org/10.1111/tops.12031

Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M., & Hale, J. T. (2023). Modeling Structure-Building in the Brain With CCG Parsing and Large Language Models. *Cognitive Science*, *47*(7), e13312. https://doi.org/10.1111/cogs.13312

Stowe, L. A., Haverkort, M., & Zwarts, F. (2005). Rethinking the neurological basis of language. *Lingua*, *115*(7), 997–1042.

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, *2*(11), 745–757. https://doi.org/10.1038/s43588-022-00354-6

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules.* [Pages: x, 445]. The MIT Press.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models [Publisher: Nature Publishing Group]. *Nature Human Behaviour*, 1–18. https://doi.org/10.1038/s41562-023-01783-7

Turken, A. U., & Dronkers, N. F. (2011). The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses. *Frontiers in Systems Neuroscience*, *5*, 1. https://doi.org/10.3389/fnsys.2011.00001

Uriagereka, J. (2011, March). 239 Derivational Cycles. In *The Oxford Handbook of Linguistic Minimalism*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199549368.013.0011

van der Burght, C. L., Friederici, A. D., Maran, M., Papitto, G., Pyatigorskaya, E., Schroën, J. A. M., Trettenbrein, P. C., & Zaccarella, E. (2023). Cleaning up the Brickyard: How Theory and Methodology Shape Experiments in Cognitive Neuroscience of Language. *Journal of Cognitive Neuroscience*, *35*(12), 2067–2088. https://doi.org/10.1162/jocn_a_02058

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, August). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. [Place: United States]. *NeuroImage*, *54*(1), 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020).

SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waugh, L. R., Monville-Burston, M., & Joseph, J. E. (Eds.). (2023). *The cambridge history of linguistics*. Cambridge University Press.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, *9*(11), e112575. https://doi.org/10.1371/journal.pone.0112575

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Wilson, K. V. (1980). *From Associations to Structure* [Google-Books-ID: xkFcnDqLi_4C]. Elsevier.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011