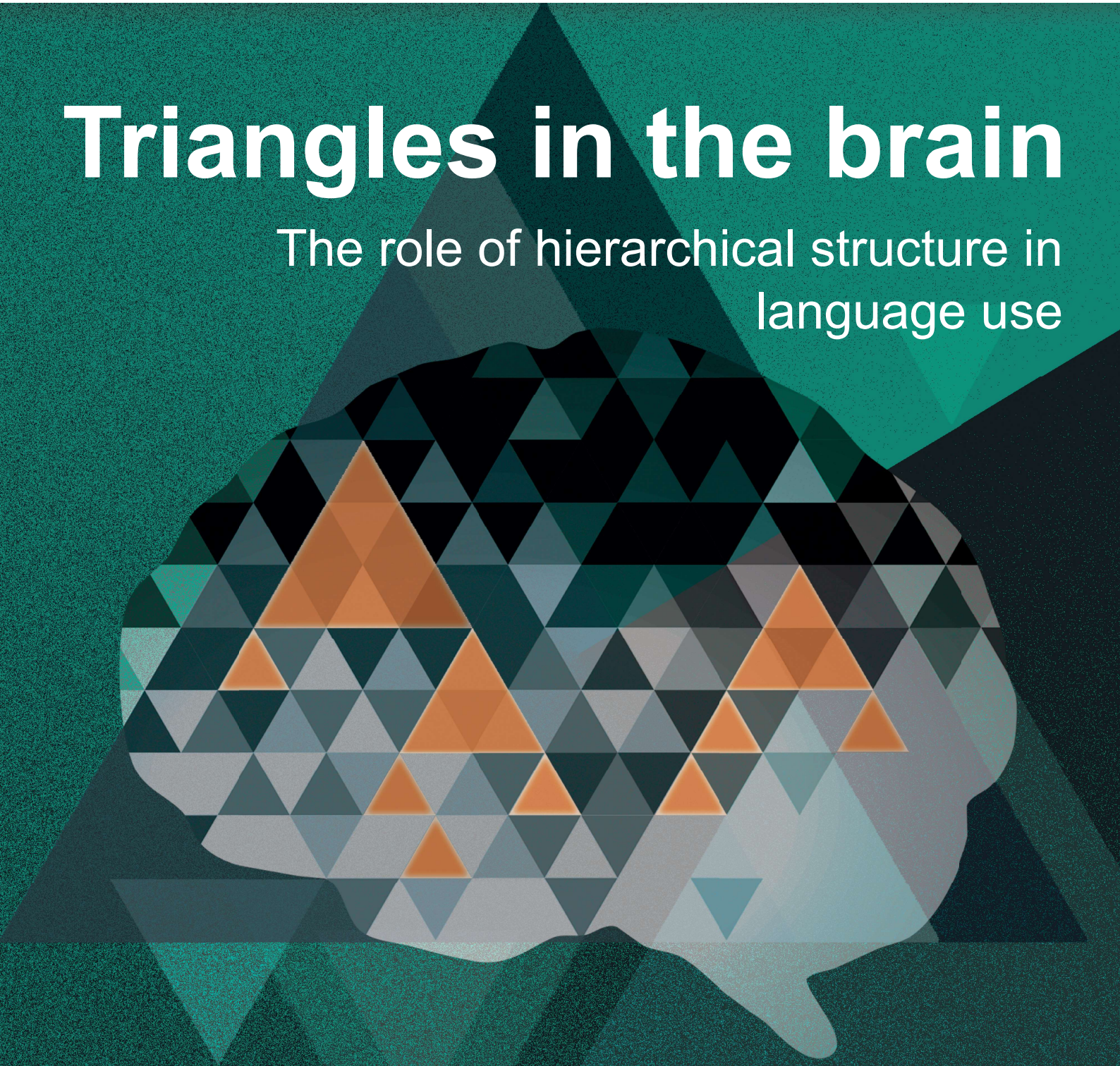
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

Triangles in the brain

The role of hierarchical structure in language use



CAS COOPMANS

Triangles in the brain:

The role of hierarchical structure in language use

Funding body
This research was funded by the Max Planck Society for the Advancement of Science (www.mpg.de/en) in the form of an IMPRS for Language Sciences PhD Fellowship (2018-2022) awarded to Cas Coopmans.

International Max Planck Research School (IMPRS) for Language Sciences
The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-46-2
Cover design by Shuang Bi
Printed and bound by Ipskamp Drukkers, Enschede

 The research reported in this thesis was conducted at the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands

Triangles in the brain:

The role of hierarchical structure in language use

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 29 maart 2023
om 12.30 uur precies

door

Casimir Willem Coopmans
geboren op 8 juni 1994
te Apeldoorn

Promotoren:

Prof. dr. Peter Hagoort

Prof. dr. Helen de Hoop

Copromotor:

Dr. Andrea E. Martin

Manuscriptcommissie:

Prof. dr. Herbert J. Schriefers

Prof. dr. David Poeppel (New York University, Verenigde Staten)

Dr. Nina Kazanina (University of Bristol, Verenigd Koninkrijk)

Contents

1 General introduction **13**

1.1 Linguistics and psychology: Structure vs. use? 14

1.2 A brief review of the empirical evidence 17

1.2.1 Hierarchy in language 17

1.2.2 Hierarchy in behavior 19

1.2.3 Hierarchy in the brain 21

1.3 Thesis outline 23

2 Hierarchy in language interpretation: Evidence from behavioral experiments and computational modeling **27**

2.1 Introduction 28

2.1.1 Behavioral evidence for hierarchical structure 29

2.1.2 Computational modeling of hierarchical structure 31

2.1.3 Background of the present study 32

2.2 Methods and results 35

2.2.1 Experiment 1 35

2.2.2 Experiment 2 38

2.2.3 Computational modeling 40

2.3 Discussion 51

2.3.1 Linear models of hierarchical structure 54

2.3.2 Structure, statistics, or both? 57

2.4 Conclusion 58

3 Constraining cognitive computational models of language **59**

3.1 Introduction 60

3.2 Possible structures, not probable strings 60

3.2.1 Evidence from co-reference and binding 61

3.2.2 Implications for computational modeling 64

3.3 The limits of variation: Impossible languages 65

3.3.1 Evidence from artificial language learning 66

3.3.2 Implications for computational modeling 68

3.4 Constraining cognitive computational models 70

3.5 Conclusion 72

4 Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech 73

4.1 Introduction 74
4.1.1 Cortical tracking of linguistic structure 75
4.1.2 Background of the present study 77
4.1.3 The present study 79
4.2 Methods 81
4.2.1 Participants 81
4.2.2 Materials 81
4.2.3 Annotations 83
4.2.4 Experimental design 85
4.2.5 Procedure 85
4.2.6 Idiom knowledge test 86
4.2.7 Speech preprocessing 86
4.2.8 EEG recording and preprocessing 86
4.2.9 Mutual information analysis 88
4.2.10 Statistical analysis of MI values 88
4.2.11 ERP preprocessing and analysis 89
4.3 Results 90
4.3.1 Speech tracking 90
4.3.2 Syntax tracking 92
4.3.3 ERPs to sentence-final verb 95
4.4 Discussion 96
4.4.1 Effects of composition in processing idioms and syntactic prose 97
4.4.2 Sentence vs. word lists: Structure and acoustics 99
4.4.3 Cortical tracking of lexicalized structure 100
4.4.4 Effects of composition on word-level speech tracking . . . 102
4.5 Conclusion 103
S4 Supplementary Information 104
S4.1 Analysis of modulation spectra 104
S4.2 EEG electrode layout 105
S4.3 Spectral power analysis 105

5 Neural source dynamics of hierarchical structure building during natural story listening 107
5.1 Introduction . . . 108
5.1.1 Neuro-computational models of sentence comprehension . 108
5.1.2 The present study . . . 113
5.2 Methods . . . 114
5.2.1 Participants . . . 114
5.2.2 Stimuli . . . 115
5.2.3 Syntactic annotations . . . 115
5.2.4 Procedure and data acquisition . . . 117
5.2.5 MEG preprocessing . . . 118
5.2.6 Source reconstruction . . . 118
5.2.7 Predictor variables . . . 118
5.2.8 Model estimation . . . 120
5.2.9 Model comparison . . . 121
5.2.10 Evaluation of the response functions . . . 122
5.3 Results . . . 123
5.3.1 Model comparison . . . 123
5.3.2 Evaluation of the response functions . . . 124
5.3.3 Region of interest analysis . . . 125
5.4 Discussion . . . 128
5.4.1 Predictive structure building in the brain . . . 128
5.4.2 Is node count the right linking hypothesis? . . . 132
5.5 Final remarks . . . 134
S5 Supplementary Information . . . 135
S5.1 Auditory stimuli . . . 135
S5.2 Comparisons against the base model . . . 135

6 Hierarchical structure in language and action: A formal comparison 143
6.1 Introduction . . . 144
6.2 Hierarchical structure in language . . . 145
6.2.1 Properties of syntactic structure . . . 147
6.2.2 Formalizing linguistic structure . . . 150
6.3 Hierarchical structure in actions . . . 160
6.3.1 Formalizing action structure (1) . . . 161
6.3.2 Formalizing action structure (2) . . . 163

6.4 Language vs. action 168
6.4.1 A formal comparison 168
6.4.2 The nature of structure 169
6.4.3 Levels of abstraction 173
6.5 Conclusions 176

7 General discussion **177**
7.1 A discussion of the main findings 177
7.2 Are there triangles in the brain? 182
7.3 Hierarchy beyond language 186
7.4 Concluding remarks 188

References **191**

Nederlandse samenvatting **233**

Research data management **239**

Acknowledgements **241**

Curriculum Vitae **245**

Publications **247**

1 | General introduction

Why aren't we puzzled when we read a sentence like "the boy who forgot his books ran home"? And why don't we get confused if someone says "the books that excite the student are heavy"? Both sentences contain sequences of words that are deviant in some way. When presented in isolation, the underlined part in "the boy who forgot his books ran home" is semantically incoherent, and the underlined part in "the books that excite the student are heavy" is clearly ungrammatical. As isolated sentences, both sequences would raise eyebrows, but when embedded in the larger sequences, their deviance goes completely unnoticed.

The reason that these locally incoherent or ungrammatical substrings are not perceived as being deviant is that phrases and sentences are more than just strings of words; their structure is organized hierarchically. While we produce language one word after another, as if the words are beads on a string, our interpretation follows from the way these words are grouped into hierarchically organized constituents. These hierarchical constituents can be denoted visually by means of geometrical structures, as shown in Figure 1.1. These representations are often called tree structures because of the way in which their increasingly diverging nodes and leaves resemble a branching tree upside down.

This hierarchical organization cannot be perceived directly – it is inaudible and invisible, hidden behind the linear surface order of the words – but it is there nonetheless. To illustrate its psychological relevance, consider again the example sentences from the first paragraph, whose tree structures are presented in Figure 1.1. What can be seen in the figures is that the sequences "his books ran home" (Figure 1.1A) and "the student are heavy" (Figure 1.1B) are never grouped together below one triangle, which is to say that they never form a constituent in these structures. Structurally speaking, "his books" belongs to "who forgot" rather than to "ran home", and "the student" is connected to "that excite" rather than to "are heavy". The fact that we do not even consider the alternative linear connections (e.g., "his books ran home") supports a conclusion that might at first thought seem paradoxical: when it comes to phrases and sentences, what reaches our eyes and ears is linear order, but what our minds

perceive is hierarchical structure (Moro, 2015). Or, to state it differently, the fact that we aren't necessarily puzzled by subsequences like "his books ran home" supports the existence of triangles in the brain.

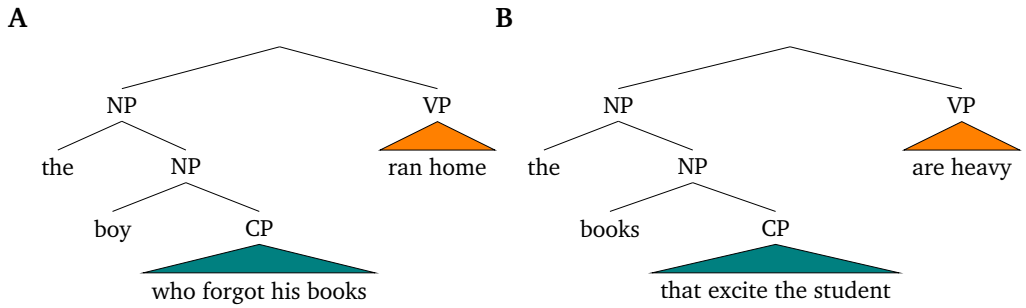


Figure 1.1: Two geometrical structures representing the hierarchical structure of sentences. The colored triangles represent syntactic constituents.

1.1 Linguistics and psychology: Structure vs. use?

Facts of this sort are discussed primarily in the field of linguistics, which is concerned with the study of language structure. However, within the psychology of language, which experimentally studies how language is used in real time, these facts remain controversial. Despite the empirical evidence from linguistics supporting the relevance of structure, psycholinguists have questioned whether hierarchical syntactic structures are always computed during actual language use. This is quite remarkable, because both linguists and psychologists of language are ultimately studying the same cognitive system. Why would the principles that explain language structure be ignored during language use?

When asked about the relationship is between linguistics and psychology, Noam Chomsky once remarked:

> "In my opinion one should not speak of a 'relationship' between linguistics and psychology, because linguistics is *part of* psychology ... In general, the following distinction is often made: linguistics is the study of language, and psychology the study of the acquisition or utilization of language. This distinction does not seem to me to make much sense. No discipline can concern itself in a productive way with the acquisition or utilization of a form of knowledge, without being concerned with the *nature* of that system of knowledge."
>
> Chomsky (1979, p. 43)

This quote expresses two views that are at the basis of the work presented in this thesis. The first is that one cannot seriously study the utilization of a form of knowledge, like language, without being concerned with the nature of that system of knowledge. Language use presupposes knowledge of language, so in order to study how our brains represent and process language, it is important to be explicit about what has to be represented and processed (Chomsky, 1965). Indeed, evidence from processing can be interpreted most directly if the representational and computational challenges faced by the parser are well understood (e.g., what are the properties of the structure that must be built?). In this sense, linguistic theory provides the basis for investigations into the psychology and neurobiology of language.

The second view is that "linguistics is part of psychology". A reformulation of this view, more focused on the specific topic of this thesis, is that grammar and processing describe a single cognitive system (Lewis & Phillips, 2015; Phillips & Lewis, 2013). This *one-system view* holds that there is only one grammatical system, used for both 'offline' acceptability judgements and 'online' language processing. Offline and online data thus represent different snapshots of the processes within this one system, and theories of grammar and processing are theories of (the outputs of) these processes, stated at different levels of description (Lewis & Phillips, 2015). A key prediction of the one-system view is that if theoretical work on the structural properties of language uncovers linguistic principles, whether invariant or language-specific, these principles should somehow be represented in the human brain, and therefore be measurable with brain recordings and in behavioral experiments. In contrast, an alternative view holds that the cognitive systems for grammar and language processing are separate and have different properties (Ferreira & Patson, 2007; Frank et al., 2012; Frazier, 2015; Townsend & Bever, 2001). Proponents of this *two-system view* argue that there are two systems, one containing a static body of knowledge (the grammar), which we rely upon when making acceptability judgments, the other a set of heuristic procedures that yield less detailed representations and that are used during everyday comprehension and production. If the two-system view is correct, this will have important implications for the possible interaction between linguists and psycholinguists. It means that grammatical theories do not need to account for psycholinguistic data and that psycholinguistic data will have no bearing on grammatical theories, because these data reflect the workings of a language processing system that does not directly recruit the grammar.

The main challenge for the two-system view is to provide an explanation for why the output of language processing is often so similar to the representations licensed by the grammar (Lewis & Phillips, 2015). If the two systems rely on different mechanisms, consistent alignment between them is unexpected, unless there is a specific account of how they interact. Conversely, the main challenge for the one-system view is explaining cases of misalignment between what is predicted by grammatical theories and what is found in measures of language processing. If there is only one grammatical system, arbitrary divergence between offline and online data is problematic. Yet, these misalignments do occur (see Section 1.2.2), and they are frequently cited in support of the representational infidelity of language processing mechanisms.

Cases of misalignment have also stirred controversy about the role of hierarchical structure in language use. In particular, it is often reported that linear properties of phrases and sentences, including sequential statistics, affect their comprehension (Arnon & Snider, 2010; Christianson et al., 2001; Ferreira & Patson, 2007; Frank & Bod, 2011; Tabor et al., 2004; Townsend & Bever, 2001). These findings are taken to show that hierarchy is somehow less fundamental, because a non-hierarchical system is simpler and therefore more parsimonious (Frank et al., 2012; Frank & Christiansen, 2018). This argument is problematic, however, for two main reasons. First, effects of linearity are not surprising because externalized language, whether spoken or signed, has serial properties due to the fundamental constraint that time is one-dimensional.[1] Even in the extreme case that the syntactic system does not care about linearity at all, linear effects are expected because the hierarchical structure has to be constructed from or flattened to a one-dimensional sequence for comprehension and production. If anything, sequential effects arising in this linearity-to/from-hierarchy transduction process are informative about how parsers select among possible representations, not about what those representations are (Phillips, 2013). Second, arguments from simplicity or parsimony should be invoked only when two competing systems are equivalent in empirical coverage. Yet, no linguistic systems exist that are both descriptively adequate and eschew hierarchy altogether. Besides, it is not at all clear that for a biological system like the human brain linearity is 'simpler' than hierarchy. Hierarchical structure allows for effective organization and control of information (Dawkins, 1976) and can be compressed

[1] Whether the visual modality in sign language affords more hierarchical cues in the externalized signal via the parallel use of the face and both hands is still an open question. Even in signs, however, the amount of hierarchical information will be limited by exactly the same temporal constraints.

efficiently in the form of generative mental programs (Dehaene et al., 2022; Restle, 1970; Simon, 1972), possibly explaining why humans encode information in a hierarchical manner whenever they can (Fitch, 2014). I will come back to this point in the general discussion.

1.2 A brief review of the empirical evidence

What is the evidence that the human mind computes hierarchical structure? I will briefly review relevant observations of three different types: observations about language, observations of people's behavior in psycholinguistic experiments, and observations of people's brain activity in neuroimaging studies. As each of the separate chapters in this thesis will discuss evidence from similar sources, this review will only highlight a few critical observations.

1.2.1 Hierarchy in language

A clear illustration of the existence of hierarchy in language is the fact that one sequence of words can be associated with multiple interpretations, even if none of the words is ambiguous. Consider the following example (from Lightfoot, 1982):

(1) John kept the car in the garage.

On one interpretation, "in the garage" refers to the place where John keeps his car – he kept it in the garage, not in the driveway. On a different reading, "in the garage" says something about "the car", so the sequence "the car in the garage" is interpreted as a unit, referring to a specific car that John kept (i.e., he did not sell it). The existence of structural ambiguity has a deep implication for the way language is mentally represented: "if some string of words can correspond to two meanings in the mind, meanings in the mind cannot be strings of words" (Pinker, 1997, p. 70). Instead, meanings in the mind hinge on hierarchical structure, as is illustrated by the two representations in Figure 1.2, which again use triangles to represent constituents.

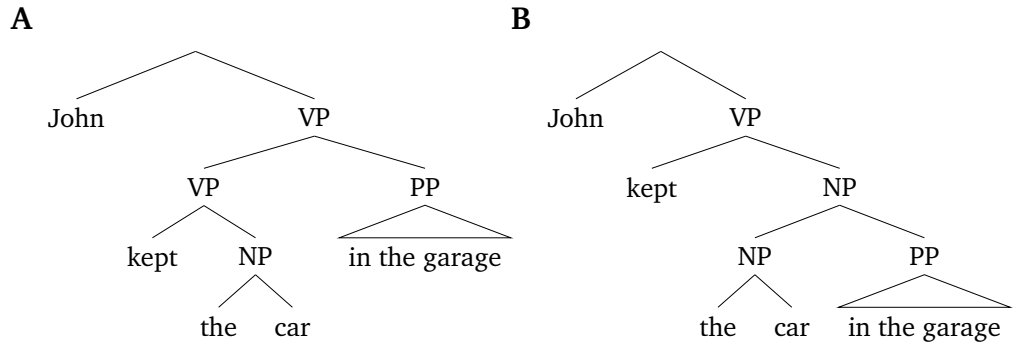


Figure 1.2: Two hierarchical representations underlying the sentence "John kept the car in the garage". The 'he-kept-it-in-the-garage' interpretation is derived from (A), the 'he-did-not-sell-it' interpretation corresponds to (B).

Here, the constituent "in the garage" modifies either the verb phrase (VP) "kept the car" or the noun phrase (NP) "the car", in line with the 'he-kept-it-in-the-garage' meaning and 'he-did-not-sell-it' meaning, respectively. This property of language, in which the meaning of a complex expression is built up from the meanings of its parts and the way in which they are structurally combined, is called compositionality (Partee, 1995).

Beyond affecting meaning, the hierarchical structure underlying phrases and sentences also affects their behavior in response to syntactic operations. To see how, consider the following passive sentence (again from Lightfoot, 1982):

(2) The car was kept in the garage.

In contrast to (1), sentence (2) is not ambiguous; it relates to just the 'he-kept-it-in-the-garage' interpretation. In short, displacement of "the car" is sensitive to the structural configuration of the sentence (Figure 1.2). The 'he-did-not-sell-it' reading could only be derived from (2) via a structural configuration that would violate a locality restriction on long-distance dependencies, so it is unavailable. Because syntactic relations and operations, including the ones underlying passivization, apply to constituents rather than to the individual words these constituents are made up of, they are said to be structure-dependent (Chomsky, 1957, 1965; Everaert et al., 2015).

Note that the observation that (1) is ambiguous but (2) is not would be puzzling if all language did was concatenate words into sequences. These examples thus make clear that in order to deal with the syntactic regularities we find in natural language, we have to be able to handle hierarchical constituent structure.

1.2.2 Hierarchy in behavior

The previous section relied on judgements about meaning and acceptability, which are commonly called offline responses because they are elicited without time restrictions. As mentioned in Section 1.1, it is debated whether the principles that explain these offline data should also account for time-sensitive online responses, which reflect the real-time processes of the language system. On the one hand, supporting the one-system view, behavioral experiments have shown that hierarchical models predict reading times in naturalistic settings (Baumann, 2014; Fossum & Levy, 2012; van Schijndel & Schuler, 2015). Moreover, people are highly sensitive to structure-dependent principles when processing syntactic constructions that include a variety of complex configurations, such as reflexives (Cunnings & Sturt, 2014; Dillon et al., 2013), anaphoric pronouns (Chow et al., 2014; Kush et al., 2015), proper names (Kazanina et al., 2007), filler-gap dependencies (Phillips, 2006; Traxler & Pickering, 1996), and ellipsis (Martin & McElree, 2008; Yoshida et al., 2013). On the other hand, for many of these constructions it has been shown that non-structural factors affect their comprehension, which could be evidence for misalignment, consistent with the two-system view. Yet, the strongest evidence for misalignment would be the finding that the representations considered by the parser are wholly inconsistent with those of the grammar, such as when people derive an interpretation or analysis from a sentence that is not licensed by its grammatical structure ('illusions of grammaticality'; see Phillips et al., 2011 for a review). One such example is a phenomenon known as agreement attraction (Bock & Miller, 1991), as in (3), whose ungrammaticality often goes unnoticed in both comprehension and production.

(3) *The key to the cabinets are on the table.

Here, the verb incorrectly agrees in number with the noun to which it is linearly closest (the noun "cabinets", which is the 'attractor'), rather than the subject noun (phrase) to which it is hierarchically related ("key"). This finding receives an initially plausible explanation in terms of 'proximity concord', wherein proximity is defined in terms linear rather than hierarchical distance. If that explanation is correct, agreement attraction would show that language users construct grammatical hypotheses that are incompatible with the representations licensed by their grammar, thus motivating the two-system view.

Lewis and Phillips (2015) discuss three reasons for questioning this explanation, all based on empirical observations of attraction effects that go in a direc-

tion opposite to what is predicted by a linearity-based account. First, production studies with complex noun phrases (e.g., "the helicopter for the flight(s) over the canyon(s) were...") show *weaker* attraction effects for nouns that are linearly closer to the verb ("canyons") than for nouns that are linearly closer subject noun ("flights"; Franck et al., 2002). Second, agreement attraction is induced even by nouns to the left of the subject noun (e.g., "the drivers who the runner wave to..."), which do not linearly intervene between the subject noun and the verb (Bock & Miller, 1991; Wagers et al., 2009). And third, intervening agreement attractors only rarely have the effect that grammatical sentences are perceived as ungrammatical (Wagers et al., 2009). None of these findings is expected on an account that is based on linear proximity, suggesting that agreement attraction is not a case of true (representational) misalignment.

Instead, the pattern of results is compatible with the architectural properties of working memory (Badecker & Kuminiak, 2007; Franck & Wagers, 2020; Wagers et al., 2009). Information in working memory is accessed in a content-addressable way, which makes it prone to interference from attractors that partially match with the agreement controller in terms of content. Within this system, content is defined in terms of retrieval cues such as [+ plural], [+ nominative], and [+ subject], which are the same features as those postulated for the grammatical system, so this account of processing behavior is fully consistent with grammar-based accounts of agreement (Lewis & Phillips, 2015). Errors arise during 'online' processing because the grammatical system is implemented within a noisy memory architecture. In the case of 'offline' judgements, the parser has more time and can thus initiate multiple attempts to retrieve the right representation using the same retrieval features, thereby reducing the chance of retrieving the wrong element.

What the foregoing discussion makes clear is that, even though the one-system view predicts alignment between the outcomes of grammar and parsing, misalignment is not necessarily problematic if it can be attributed to other factors that receive independent empirical motivation (Lewis & Phillips, 2015; Phillips, 2013). Needless to say, this does not mean that it must always be possible to account for misalignments while retaining the one-system architecture. Each individual case of misalignment presents a challenge to the one-system view and is therefore potentially very important. Careful examination is necessary to determine whether it indeed supports the two-system view or can instead be explained within a one-system architecture. Importantly, this is not an ad hoc endeavor; if the one-system view is correct, it should be possible both to explain

cases of misalignment using independently motivated properties of our cognitive system, as well as to predict under which exact circumstances misalignment is likely to occur.

1.2.3 Hierarchy in the brain

The cognitive neuroscience of syntactic structure building is commonly investigated by comparing brain responses to stimuli in two conditions that differ in their hierarchical structure. To make sure that non-hierarchical accounts cannot explain the results, it is important that the two conditions do not differ in terms of their linear properties. Two experimental paradigms approach this challenge in a particularly clever way. The first paradigm was used in a seminal fMRI study by Pallier et al. (2011), who used a parametric manipulation of constituent structure. Participants read sequences that all had a fixed length of twelve words but that differed in the size of the constituents that could be constructed from these words. In six successive conditions, constituent size was parametrically increased from one (a list of twelve words) to twelve words (one twelve-word sentence). In line with the hypothesis that the neural assembly that encodes constituent structure grows with the size of the constituents, Pallier and colleagues found a set of regions in the left-hemispheric language network that showed a systematic increase in activation as constituents grew larger. This constituent-size effect has now been replicated across several modalities, including spoken (Brennan et al., 2012, 2016), written (Giglio, Ostarek, Weber, & Hagoort, 2022; Matchin et al., 2017; Zaccarella et al., 2017), and sign language (Matchin et al., 2022; Moreno et al., 2018). While these different studies report effects of constituent size in slightly different brain regions, the consistency of the effect in general shows that this paradigm effectively targets hierarchical structure building while controlling for sequential properties of the stimuli. The study reported in **Chapter 5** of this thesis therefore adopts a similar parametric design, but it manipulates constituent structure in a naturalistic way rather than relying on slightly artificial stimuli presented in isolation.

The second paradigm was pioneered in a study by Ding et al. (2016). It relies on a phenomenon known as cortical speech tracking, which refers to the brain response to slow regularities in speech. Participants listened to connected speech sequences of monosyllabic words. Within these sequences, two adjacent words could repeatedly be grouped into phrases, and two adjacent phrases could repeatedly be grouped into sentences, such as in $[_{S}\,[_{NP}$ new plans$]\,[_{VP}$ gave hope$]]$. Because the sequences were isochronously presented without prosody, only the

words were clearly defined by acoustic boundaries. That is, words were physically present in the speech signal, but phrases and sentences were not – they had to be internally constructed by the brain using knowledge of grammar. Intriguingly, electrophysiological brain activity tracked not only the presentation rate of words, but also the rates of phrases and sentences, showing a clear disconnect between what is objectively present in the input signal and what is represented in the brain signal. Moreover, when the sequences contained words that could be grouped into two-word phrases at most, the brain recordings revealed a tracking effect at the phrase rate only, not at the sentence rate. These findings have been replicated in subsequent studies (Blanco-Elorrieta et al., 2020; Burroughs et al., 2021; Ding, Melloni, et al., 2017; Getz et al., 2018; Makov et al., 2017; Sheng et al., 2019), showing that this method can be used to investigate how the brain infers hierarchical structure from a temporal speech stream (though see Kazanina & Tavano, 2022 for a critical perspective). The study reported in **Chapter 4** of this thesis therefore relies on cortical speech tracking as well, but it makes use of naturally spoken sentences rather than isochronously presented speech.

Before concluding, it is good to emphasize how surprising it really is that syntactic generalizations are structure-dependent and that the effects of structure dependence are consistently found in both behavioral and neural measures of language processing. As Andrea Moro remarked a few years ago, the existence of structure dependence is a remarkable fact "not just because it is inaccessible to our immediate introspection, but also because it is based on the only phenomenon inaccessible to our senses – namely hierarchy – whereas linear order is completely irrelevant" (Moro, 2016, p. 40). This nontrivial observation shines a very different light on Aristotle's famous dictum that "there is nothing in the mind that was not first in the senses". The dictum must be incorrect, because hierarchy is not part of the physical signal that reaches our eyes and ears. Empirical findings such as the ones discussed here therefore illustrate how language comprehension is a form of perceptual inference (Martin, 2016, 2020), combining external (perceptual) information about speech or sign with internal (linguistic) information about the properties of language in order to derive an analysis and interpretation of the incoming signal. The observation that our minds perceive hierarchical structure despite its absence in the input forms the basis of this thesis.

1.3 Thesis outline

In this thesis, I investigate the role of hierarchy in language use. Five chapters approach this topic from different angles, each asking a different question and using a different methodology. The methods are experimental, computational, and theoretical, and can be classified into one of the three strands of evidence discussed in Section 1.2.

In **Chapter 2**, I empirically address a recently expressed view in (computational) psycholinguistics which questions the importance of hierarchical structure in language. In contrast to well-known arguments from linguistics, this view holds that language use is fundamentally sequential. In response to this claim, we tested with a very simple behavioral paradigm whether people interpret ambiguous three-word phrases such as *second blue ball* in terms of their hierarchical structure or their linear surface order. Using an experimental setup in which the hierarchical and the linear interpretation of *second blue ball* do not pick out the same referent, we could determine whether people rely on hierarchical or linear structure for language interpretation. We subsequently trained and tested an artificial neural network (ANN) on a computational version of the same experimental task. We looked at a particular type of recurrent neural network architecture, the so-called long short-term memory (LSTM) neural network, about which it has been argued that it can acquire critical properties of natural language syntax. If these arguments are right, we should expect the model to behave similarly to the participants in the behavioral experiment. In several simulations, we asked whether the LSTM could reproduce the behavior of the human participants, and evaluated whether its performance varied as a function of the data on which it was trained.

In **Chapter 3**, I follow up on the computational results from **Chapter 2** with a critical discussion about the role of ANN models in the language sciences. Due the impressive performance of ANNs on a range of language tasks, ranging from machine translation to text generation, it is often argued that these models acquire many of the structural properties of human language. Yet, as any type of behavior is multiply realizable, the success of an ANN model on a given language task does not justify the behaviorist inference that the model succeeded in the way humans would. In this chapter, we argue that the results of these modeling endeavors will be most impactful in the scientific study of language if it can be shown that the model's performance comes about in a way that resembles the way in which cognition underlies our behavior. This cautionary perspective on the use of computer systems in cognitive science is by no means novel, as is

clear from the following quote, written in 1960 by three scientists who (unsurprisingly) were the foundation of the cognitive revolution:

> "It becomes obvious that there are two very different attitudes one can take toward the job [of machine translation]. In one attitude, the programmer says, 'I want to make it work any way I can, but the simpler it is, the better.' In the other attitude, he says, 'I want to make the computer do it the same way people do it, even though it may not look like the most efficient method.' As citizens we should applaud the former attitude, but as psychologists, linguists, neurologists – as students of the human being – we are bound to be more interested in the latter."
>
> Miller, Galanter, and Pribram (1960, p. 54)

As the author of this thesis, I consider myself a "student of the human being", and am therefore bound to be more interested in computer systems that acquire and use language in "the same way people do it". In **Chapter 3**, we argue that these systems currently do not exist, and describe two reasons for the misalignment between *cognitive* and *computational* approaches to language. In order to address this scientific gap, we suggest two changes to computational modeling of syntax. With these changes in mind, computational models of language can move closer towards integrating valuable insights from across the language and cognitive sciences.

In **Chapter 4**, I present data from an electroencephalography (EEG) experiment which aimed to investigate the neural processes involved in inferring the hierarchical structure of sentences from naturally produced speech. This chapter was inspired by a range of studies that show that the brain 'tracks' the presentation rate of phrases in natural speech, a phenomenon known as cortical tracking of syntax (see Section 1.2.3). This finding is remarkable because syntactic phrases are abstract units of information which are not visible in the acoustic signal in the way lower-level linguistic information is. Here, we set out to test why syntactic phrases are tracked more strongly when they are embedded in regular, meaningful sentences than when the stimulus in which they are embedded is less meaningful. One possibility is that phrases are tracked more strongly in sentences because words in sentences can be composed into meaningful constituents (reflecting the output of structure building). Alternatively, they are tracked more strongly in sentences because the lexical-syntactic information carried by content words allows words in sentences to be easily composed in the first place (reflecting the input to structure building). To determine whether it is the input to or the output of structure building that modulates cortical tracking

of syntax, we relied on a parametric modulation of linguistic information, comparing regular, meaningful sentences to carefully controlled stimuli in four conditions. These four conditions contained idioms, syntactic prose, jabberwocky sentences, and word lists, which differ from sentences in the extent to which a combination of their structure and lexical components determines their meaning. We computed cortical tracking via a measure called mutual information, which quantifies the amount of shared information between two signals, in our case the speech input and the corresponding EEG response. Because both signals were filtered in the narrow frequency range corresponding to the rate at which phrases naturally occurred in the stimuli, the shared information between these signals gives an indication of how closely the brain tracks phrases in natural speech.

In **Chapter 5**, I further investigated the processes involved in hierarchical structure building, but this time the sentences were presented as part of a coherent narrative. We specifically asked which strategy language comprehenders use to build structure. Even if we assume that parsing is incremental, starting with the first word and ending with the last, there are several ways in which hierarchical tree structures such as those in Figure 1.1 can be built. In this chapter, we compare different parsing strategies, which build structure either in a predictive or in an integratory manner. Predictive strategies build nodes in the structure before there is complete evidence for their existence, whereas integratory strategies instead take a wait-and-see approach; they build nodes only after it is certain that they are necessary. To examine which of these strategies best accounts for the comprehension of Dutch sentences, we let participants listen to audiobook stories while we recorded their brain activity using magnetoencephalography (MEG). Each word in the audiobook was assigned a so-called complexity metric, which quantifies the structural complexity of that word given each of the different parsing strategies. In effect, this means that for a predictive parser, words at the onsets of phrases are complex, because they can be used to build structure predictively. For an integratory parser, in contrast, words at the offsets of phrases are complex, because they can be integrated with the constituents to which they belong. To see how these complexity metrics relate to language-relevant brain activity, we make use of a type of forward model called temporal response function. This approach allows us to compare different parsing strategies in their ability to predict brain activity of Dutch people listening to natural stories. The predictive accuracy of these different strategies will give us insight

into the extent to which people build hierarchical structure in a predictive or in an integratory manner.

In **Chapter 6**, I compare hierarchies across domains, specifically looking at the structure of actions and action plans. Like sentences, actions are thought to be hierarchically organized. Based on this putative structural analogy, it is often suggested that language and action rely on cognitive and neural resources that are (at least partially) shared. However, while the claim that syntax is hierarchically structured is supported by empirical evidence and accompanied by formal characterizations of its properties, both of these are lacking in the study of actions. Before drawing conclusions about cross-domain convergence, it should first be shown that the structures of language and actions have the same formal properties. Unfortunately, this comparison is complicated by the fact that formal treatments of syntactic structure often rely on domain-specific constructs, defined in jargonistic terms. In this chapter, we therefore formally compare language and action using the domain-neutral vocabulary of set theory. Starting with a model that relies on the minimal assumption that the combinatorial operator for generating syntactic structure is formally equivalent to binary set formation, we aim to capture the core properties of syntax. By subsequently applying this model to the domain of actions, we can see whether the core properties of linguistic structure are also found in action structures. The conclusions of this chapter have implications for both cognitive neuroscience and comparative cognitive science.

In **Chapter 7**, I summarize the results of the different chapters, discuss the representation of hierarchical structure in the brain, and end with a broader discussion about the role of hierarchical structure in human cognition.

2 | Hierarchy in language interpretation: Evidence from behavioral experiments and computational modeling[1]

Abstract

It has long been recognized that phrases and sentences are organized hierarchically, but many computational models of language treat them as sequences of words without computing constituent structure. Against this background, we conducted two experiments which show that participants interpret ambiguous noun phrases, such as *second blue ball*, in terms of their abstract hierarchical structure rather than their linear surface order. When a neural network model was tested on this task, it could simulate such 'hierarchical' behavior. However, when we changed the training data such that they were not entirely unambiguous anymore, the model stopped generalizing in a human-like way. It did not systematically generalize to novel items, and when it was trained on ambiguous trials, it strongly favored the linear interpretation. We argue that these models should be endowed with a bias to make generalizations over hierarchical structure in order to be cognitively adequate models of human language.

[1]Adapted from Coopmans, C. W., de Hoop, H., Kaushik, K., Hagoort, P., & Martin, A. E. (2022). Hierarchy in language interpretation: evidence from behavioural experiments and computational modelling. *Language, Cognition and Neuroscience, 37*(4), 420-439.

2.1 Introduction

The ability to use language is a hallmark of the human mind. The formal structures of human language reveal the wealth of representational infrastructure that our brains deploy to guide our linguistic behavior. As such, even in a short phrase like *these two blue balls* lies a hidden signal about how the mind structures information. For this simple four-word phrase, there are 24 logically possible word orders, yet only 14 of these are attested in the world's languages (Cinque, 2005). Strikingly, the word order in English and its mirror variant (*balls blue two these*) are by far the most frequent (Cinque, 2005; Greenberg, 1963), reflecting the selection of word orders that transparently map to the hierarchical structure of the noun phrase (Culbertson & Adger, 2014; Martin et al., 2020). The word 'hierarchical' here refers to the representational format of constituent structure: words are embedded into constituents, which are in turn recursively embedded into larger constituents, creating hierarchically organized syntactic structures which are often visually denoted by means of tree structures (see Figure 2.1A). It has long been argued that the semantic interpretation of phrases and sentences is linked to this hierarchical constituent structure (Chomsky, 1957; Everaert et al., 2015; Heim & Kratzer, 1998; Jackendoff, 1972; Partee, 1975; Pinker, 1999). That is, syntactic operations are defined over hierarchical structure rather than linear order (i.e., they are structure-dependent; Chomsky, 1957), and semantic dependencies (like scope, the fact that *two* applies to *blue balls* rather than *balls* alone[2]) follow from such hierarchically organized constituent structure.

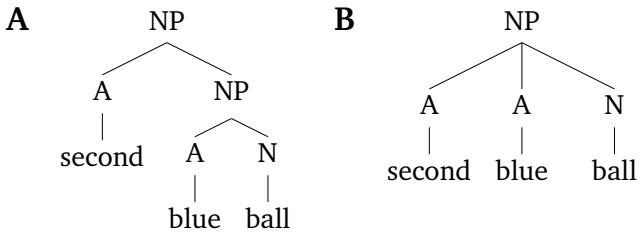


Figure 2.1: Hierarchical (A) and linear (B) representations for the phrase *second blue ball*.

Despite these arguments in theoretical linguistics, however, an alternative view holds that language use can be accounted for in terms of sequential rather than hierarchical structure (Bybee, 2002; Christiansen & Chater, 2015; Frank et al., 2012). A core aspect of this view, which has been championed by sev-

[2]Semantic scope refers to the domain in which an operator can affect the interpretation of other elements. Scope domains can sometimes be directly read off hierarchical relations between syntactic elements, that is, by virtue of the c-command relation (Reinhart, 1983).

eral authors in different proposals, is that constituency is not a basic structure but rather an epiphenomenon, emerging from frequently occurring sequential patterns in language, which are 'chunked' into sequences without much internal structure (Bybee, 2002; Christiansen & Chater, 2015; Frank et al., 2012). In other words, while this 'linearity view' does not entail that hierarchical structure does not exist, it holds that hierarchy is not fundamental in language use. This view is strengthened by the recent successes of mainstream models in natural language processing (NLP), which treat sentences as linear strings of words. These models achieve remarkably good performance, arriving at around 93% accuracy on several diagnostics (e.g., Devlin et al., 2019), and are often used to account for behavioral data in psycholinguistic experiments (e.g., Christiansen & MacDonald, 2009; Frank & Bod, 2011; Gulordava et al., 2018; Linzen et al., 2016).

Against this background, we use the interpretation of ambiguous noun phrases such as *second blue ball* as a test of the idea that constituency is not fundamental in language use. We first show in two behavioral experiments that the interpretation of these phrases is based on their hierarchical rather than their linear structure, indicating that language interpretation can in fact be biased towards hierarchical constituency. We then train and test a recurrent neural network model on our task in order to see whether it is able to reproduce such 'hierarchical' behavior. In several simulations, we evaluate whether the model generalizes in a human-like way. We show that it can simulate hierarchical behavior, but only if the training data are unambiguously hierarchical. When it is trained on ambiguous data which are equally consistent with the linear and the hierarchical interpretation of *second blue ball*, it strongly favors the linear interpretation. Moreover, the model does not systematically generalize to items that were not observed during training. Overall, this leads us to conclude that without a predisposition for hierarchical structure, the model is not a cognitively adequate model of human language (Dehaene et al., 2015; Fitch, 2014).

2.1.1 Behavioral evidence for hierarchical structure

Broadly speaking, two kinds of linguistic evidence support the claim that words, phrases and clauses have internal hierarchical structure. First, syntactic operations, such as movement, deletion and substitution target constituents rather than individual words. These operations are said to be structure-dependent, and behavioral experiments have shown that children obey structure dependence as soon as they can be tested (Crain & Nakayama, 1987). Second, structure pro-

vides the unit of semantic interpretation, as can be seen in the structural ambiguity of words (e.g., *uninstallable*), phrases (e.g., *deep blue sea*) and clauses (e.g., *she saw the man with binoculars*), as well as the structure-dependent interpretation of anaphora, disjunction, negative polarity items, and other scope phenomena (Reinhart, 1983; see Crain et al., 2017 for a recent overview of the empirical data from acquisition). These facts about language structure show that constituents behave as units, both to syntactic operations and to semantic interpretation.

Furthermore, a large body of experimental evidence converges in showing how hierarchical structure explains language behavior. Of particular relevance to the current study are three behavioral paradigms which investigate noun phrase interpretation. First, Lidz and colleagues used a preferential looking paradigm to show that 18-month-old infants interpret the pronominal *one* in *Look! A yellow bottle. Do you see another one?* as anaphoric with the constituent *yellow bottle* rather than with the bare noun *bottle*, consistent with the interpretation of anaphoric *one* in adult language (Lidz et al., 2003). Second, a cross-domain structural priming study by Scheepers and Sturt (2014) showed that people find adjective-noun-noun compounds more acceptable when their structure is congruent with a mathematical equation that they have solved just before. In their study, left-branching phrases, such as *organic coffee dealer* (i.e., [[organic coffee] dealer]), received higher ratings after left-branching equations (e.g., $25 \times 4 - 3$) than after right-branching equations (e.g., $25 - 4 \times 3$). Third, Culbertson and Adger (2014) exposed English learners of an artificial language to different noun phrases with only one postnominal modifier (i.e., N-Dem, N-Num, N-Adj), based on which they had to infer the relative ordering of the modifiers in a complex noun phrase (see also Martin et al., 2020). The training data were equally consistent with two possible grammars, one of which was similar to English in terms of the linear ordering of the modifiers (i.e., balls *these two blue*), while the other was similar to English in terms of the abstract structure of the noun phrase (i.e., [[[[balls] *blue*] *two*] *these*]). The learners consistently favored the order that was structurally similar to English, despite its dissimilarity to English in terms of surface statistics. In line with this finding, a recent study on artificial rule learning showed that people from different age groups and different cultural and educational backgrounds spontaneously infer and generalize abstract hierarchical structure after exposure to sequences whose structure is fully consistent with both hierarchical rules (based on recursive center-embedding) and linear rules (based on ordinal position; Ferrigno et al., 2020). Combined,

these studies demonstrate that people represent noun phrases as hierarchical structures rather than as linear sequences. Moreover, the studies by Lidz et al. (2003) and by Culbertson and Adger (2014) indicate that this hierarchical bias does not come from the environment but rather reflects an inherent property of the linguistic system, which might also be present in other domains of cognition (Dehaene et al., 2015; Ferrigno et al., 2020; Fitch, 2014).

Evidence from the spontaneous creation of languages in language-deprived populations supports this latter point. Deaf children who are born to speaking parents and are not exposed to sign language in infancy spontaneously develop a gestural system for communication (Goldin-Meadow, 2003). This system, called homesign, has many of the properties of natural language, including hierarchically organized levels of recursive constituent structure and structure-dependent operations, such as substitution (Goldin-Meadow, 2003; Hunsicker & Goldin-Meadow, 2012). For example, in homesign, multi-gesture combinations that refer to a single nominal entity (e.g., a demonstrative gesture and a noun gesture: "that bird") function both syntactically and semantically like single-gesture nominals. They can substitute for a single noun ("bird") and can be embedded in a hierarchically structured clause, to yield a signed clause with the hierarchical structure [[that bird] pedals] rather than the flat structure [that bird pedals] (Hunsicker & Goldin-Meadow, 2012). Because the multi-gesture nominals produced by homesigners are effectively absent in the gestures of their hearing family members, they reveal that the homesigners themselves are the source of these structural properties in their linguistic system (Flaherty et al., 2021).

2.1.2 Computational modeling of hierarchical structure

While the behavioral evidence strongly supports the hierarchical view, the linearity view is strengthened by recent results from computational studies of language acquisition. Most contemporary language models are not endowed with a cognitive architecture that supports the acquisition and knowledge of linguistic information (e.g., hierarchical representations, structure dependence, or compositionality), yet they perform quite well on a range of language tasks. In particular, recent computational research with recurrent neural network (RNN) models has shown that these models often perform quite accurately on tasks which are thought to require knowledge of hierarchical structure, such as subject-verb agreement and question formation. For example, RNNs can learn to generate the correct agreement in long-distance dependencies (e.g., *The boy who likes the girls has* ...) and to move the right verb in constructing complex yes-no

questions (e.g., *Has the boy who likes the girls ...?*), seemingly without invoking hierarchical structure (Gulordava et al., 2018; Linzen et al., 2016; McCoy et al., 2018, 2020; Tran et al., 2018). Moreover, RNNs are able to generalize very well to novel grammatical constructions when these feature a mixture of examples that were observed in the training set, but they fail to systematically generalize across items in the training set to compose novel items (Baroni, 2020; Lake & Baroni, 2018; Loula et al., 2018). These findings show that RNN models show impressive generalization ability, apparently without relying on systematic compositionality.

It is often the case that the data on which these models are trained are both qualitatively and quantitatively very different from the linguistic input children receive (Linzen, 2020; Linzen & Baroni, 2021). Recent studies have sought to address this issue by exposing the model during training only to ambiguous data, from which multiple generalizations are possible (e.g., McCoy et al., 2018, 2020; Mulligan et al., 2021). During the test phase, the model is then evaluated on items for which these generalizations make different predictions. The idea behind this train-test regime is that the model's performance on test trials reveals its specific inductive biases. Comparing this performance to human behavior in the experimental paradigms discussed above (Culbertson & Adger, 2014; Ferrigno et al., 2020; Martin et al., 2020), we can evaluate whether these models generalize in a human-like way. Initial results from these studies show that some RNN architectures can make human-like syntactic generalizations, in particular when the training data contain cues to hierarchical structure (McCoy et al., 2018).

In short, while most computational language models do not explicitly incorporate structure dependence, they appear extremely proficient in a range of complex language tasks if they are trained on quantitatively and qualitatively rich data. This reveals a possible gap between the validity of these models as models of human cognition and their ability to achieve human-like behavior in certain circumstances. We approach this issue by comparing the performance of a long short-term memory (LSTM) neural network to the behavior of human participants on a task that requires hierarchically structured knowledge. The following sections first describe the task and results from the behavioral experiments.

2.1.3 Background of the present study

In two experiments, we tested whether people interpret ambiguous noun phrases such as *second blue ball* as a hierarchical structure or as a linear string. On the hierarchical interpretation, which is derived from the right-branching structure

depicted in Figure 2.1A, the structure encodes semantic scope. The ordinal *second* takes scope over the constituent *blue ball*, and the whole refers to the second among blue balls. On the linear interpretation, instead, *second* and *blue* are interpreted conjunctively, and they independently modify the noun *ball* (i.e., the ball that is blue *and* second). Here, the conjunctive (linear) interpretation is associated with the flat representation depicted in Figure 2.1B. However, we note that this is not the only possible way in which that interpretation can be represented. It could also be derived from a hierarchical structure, for instance by means of a conjunction phrase which first combines *second* and *blue*, and is then combined with *ball*. In contrast, the scopal (hierarchical) interpretation of *second blue ball* can only be derived from a nested constituent structure (i.e., Figure 2.1A). Because the hierarchical interpretation cannot be derived without hierarchical structure (as in the linear representation in Figure 2.1B), consistently hierarchical responses should be taken as evidence against the view that hierarchical structure is unnecessary to account for language interpretation.

To show how the semantics corresponding to these phrases relates to their structure (Partee, 2007; Spenader & Blutner, 2007), we provide the lambda expressions for the noun *ball* (which is of type <e,t>), the intersective adjective *blue* and the adjective *second* (which are both predicate modifiers of type <<e,t>,<e,t>>) below:

1. *ball*: λx[ball(x)]
2. *blue*: λPλx[P(x) & blue(x)]
3. *second*: λPλx[P(x) & $\exists$!y[P(y) & y < x]]

where < indicates a type of ordering relationship (i.e., y precedes x on some dimension, such as space or time).

In these expressions, P refers to a one-place predicate, i.e., a set of individuals that is the denotation of a noun such as *ball*. Hence, we get the following lambda expressions that correspond to the noun phrases *blue ball* and *second ball*, which are both of type <e,t>:

4. *blue ball*: λx[ball(x) & blue(x)]
5. *second ball*: λx[ball(x) & $\exists$!y[ball(y) & y < x]]

Combining these expressions yields the hierarchical right-branching interpretation of the complex noun phrase *second blue ball* (corresponding to Figure 2.1A), as expressed in (6):

6. Hierarchical interpretation: λx[ball(x) & blue(x) & $\exists$!y[ball(y) & blue(y) & y $<$ x]]

This means that *second blue ball* on the hierarchical interpretation refers to the set of elements x that are a member of the intersection of the set of balls and the set of blue things, such that there is exactly one other element in this intersection, which is the set of blue balls, preceding x (in one way or another). Clearly, on this interpretation, *second* applies to the set of blue balls, which means that *blue* and *ball* are combined to form a constituent that serves as the argument of *second*.

On the linear interpretation of *second blue ball* this would not be the case. Here, *second blue ball* would denote the set of elements x that are a member of the intersection of the set of balls and the set of blue things, such that there is exactly one other element in the set of balls preceding x. On this interpretation, the phrase refers to the second ball, which is blue (i.e., a green ball is in the first position). The lambda expression for the linear interpretation of *second blue ball* (corresponding to Figure 2.1B) is given in (7):

7. Linear interpretation: λx[ball(x) & blue(x) & $\exists$!y[ball(y) & y $<$ x]]

While these two interpretations could yield the same referent (Figure 2.2A), this need not be the case: based on the context in which *second blue ball* is presented, the linear and hierarchical interpretations can diverge (Figure 2.2B). This divergence forms the basis of the current study.

The idea was based on a set of acquisition experiments conducted in the 1980s, in which it was investigated how children acquire and interpret prenominal modifier sequences (Hamburger & Crain, 1984; Matthei, 1982). Matthei (1982) asked five-year old children to point to the *second blue ball* in an array of colored balls in which the linear and hierarchical interpretations yielded a different answer (Figure 2.2B). The children interpreted the phrase intersectively, pointing to the ball that was blue and in the second position, rather than to the second among blue balls. This was taken to indicate that the children had built an unembedded, linear representation. In a reply to this study, Hamburger and

Crain (1984) noted that Matthei (1982)'s results reflected the children's inability to deal with the cognitive complexity of the task, which might have concealed their hierarchical grammatical knowledge. They attempted to reduce the nature of the planning component underlying these linguistic expressions by letting children first *point to the first blue ball*, and then *point to the second one*. The children's interpretations of *one* in this scenario indicate whether they relied on a linear representation of *first blue ball*, in which case *one* can only refer to *ball*, or on a hierarchical representation, in which *one* can also refer to *blue ball*. Similar to the infants in the Lidz et al. (2003) study, four-year-old children took *one* as anaphoric with the constituent *blue ball*, indicating that they relied on a hierarchical representation of *first blue ball*. We adopted a similar experimental paradigm, but chose to use full noun phrases rather than anaphoric proforms, given the debate about whether *one* substitutes for syntactic constituents (Goldberg & Michaelis, 2017; Payne et al., 2013).

2.2 Methods and results

2.2.1 Experiment 1

The first experiment is a replication of the original study by Matthei (1982), but with only adults. 20 native speakers of Dutch (14 female, mean age = 21.9 years, age range = 19–27 years) participated in the experiment, none of whom were color-blind. Based on the results of Hamburger and Crain (1984), we expected there to be a strong hierarchical bias. In order to show such an effect empirically and to be able to test it statistically (rather than only relying on our intuition), we used a sample size of 20. All participants gave written informed consent to take part in the experiment, which was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University Nijmegen. The experiment was conducted in Dutch, but for ease of exposition, the stimuli are translated here into English, which in these sentences has the same surface word order as Dutch. Participants had to click on a target denoted by a noun phrase containing an ordinal, a color adjective and a noun referring to the shape of the target, such as *second blue ball*. Two example arrays, corresponding to the two conditions, are presented in Figure 2.2.

In the convergent condition, the hierarchical (non-intersective) and linear (intersective) interpretation converge on the same item. For example, the second blue ball in Figure 2.2A is both the second among blue balls (hierarchical) and

Figure 2.2: Example arrays for the target *second blue ball*, corresponding to a convergent (A) and a divergent (B) trial.

also the ball that is blue and in second position (linear). In the divergent condition, the linear and hierarchical interpretation yield a different answer. While the second ball in the array in Figure 2.2B is blue (linear), it is not the second among blue balls, which is in fourth position (hierarchical).

The convergent condition was not present in the original studies (Hamburger & Crain, 1984; Matthei, 1982). The responses in this condition do not dissociate between hierarchical and linear interpretations, and serve as fillers to reduce the potential influence of pragmatic factors. That is, one could argue that participants only give hierarchical answers in response to *second blue ball* on divergent trials because they take the mere presence of *blue* to indicate that they should not interpret the phrase as referring to the second ball. Had that been the intended target (e.g., in the picture of Figure 2.2B), then it could have been referred to as *second ball*, thus making the addition of *blue* redundant and therefore pragmatically odd.[3] By making sure that half of the trials contain a redundant color adjective, we intended to make participants less sensitive to the effect of redundancy on interpretation, thereby making it less likely that their behavior on divergent trials would be driven by pragmatic factors.

Each trial consisted of the written sentence "Click on the [target]" and an array of eight blue or green balls, visually presented at the same time on a computer screen. The target was always described using an ordinal, a color adjective, and the noun *ball*. The ordinals first, second, third, fourth, fifth, and sixth were used. There were 192 trials, half of which were divergent, the other half were convergent. In both conditions, all ordinals were used 16 times in the target phrase, and they were equally often combined with green as with blue. Convergent trials were created as follows: all items to the left of the target were the same as the target, and all items to the right were randomized. For the divergent trials, there were two possible targets: a linear one and a hierarchical one.[4] For every

[3]Note that even if it is the case that such pragmatic factors drive people to interpret *second blue ball* non-intersectively (i.e., not referring to the ball that is blue and in second position), they would still have to use constituent structure to interpret the phrase 'hierarchically'.

[4]This applies to all ordinals except the ordinal *first*, for which divergent trials do not exist. That is, if the first among blue balls is not the first ball, then the linear interpretation is not present and only the hierarchical option is available. Divergent trials with ordinal *first* were actually non-convergent trials with only a hierarchical option, and were therefore not analyzed.

ordinal, the position of the hierarchical target was randomly chosen among the positions to the right of the linear target. The positions to the left of the hierarchical target were then filled with the right number of items that were the same as the target. For instance, for the target *sixth green ball*, the hierarchical target could be in the positions 7 or 8. To the left of this position five green balls were placed, and one of these green balls was in sixth position (linear target). The other positions are filled with blue balls. Correct answers on convergent trials were coded as hierarchical/linear, while all other items were coded as error. On divergent trials, answers were coded as hierarchical, linear, or error.

Results

The results of experiment 1 are presented in Figure 2.3. The graph on the right contains the results for divergent trials, which show that of all correctly answered trials, participants gave a hierarchical answer 99.8% of the time. Only three answers were according to a linear interpretation. To test this effect, we applied a logistic regression model in R (R Core Team, 2020) with only an intercept to the binary output variable (hierarchical vs. linear), which showed that participants gave more hierarchical than linear answers, β = -6.27, SE = 0.58, Wald z = -10.84, $p < .001$.

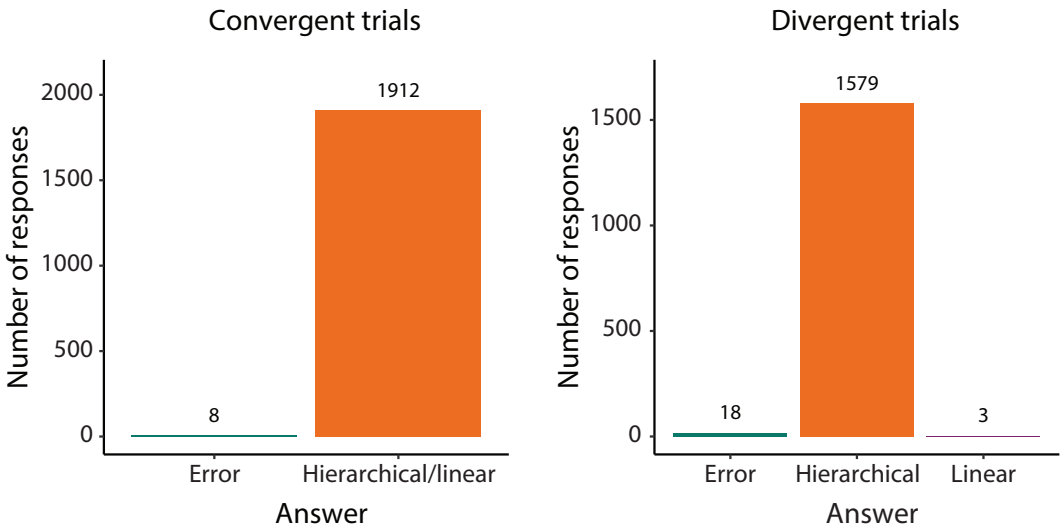


Figure 2.3: Responses in the convergent and divergent conditions of experiment 1.

While these results strongly suggest that the participants used hierarchical syntax, there is one alternative interpretation that does not need to rely on constituent structure. In this interpretation, *second* applies to the set of *blue* things

first, hence forming a complex adjective *second blue*, which is then applied to the noun *ball* (e.g., a ball that is second among blue items). This is similar to phrases in which *second* modifies an adjective, e.g., *second biggest ball* (which is the ball that is the second biggest, but not necessarily the second ball), and phrases in which *blue* is modified by an adverb, e.g., *very blue ball* (which is very blue, not very ball). Because the arrays of items contained only balls, this approach always yields the same target as the hierarchical interpretation.

Importantly, while this alternative interpretation can be represented in a constituent structure (as in the left-branching structure in Figure 2.4B), it does not, strictly speaking, need hierarchy. In the right-branching hierarchical structure in Figure 2.4A, a relationship is established between the element *second* and a constituent (i.e., a constituent is modified). Such a constituency-based relationship is not needed to represent the meaning of the left-branching structure in Figure 2.4B, which would be expressed as follows:

8. Left-branching interpretation: $\lambda x[\text{ball}(x) \,\&\, \text{blue}(x) \,\&\, \exists!y[\text{blue}(y) \,\&\, y < x]]$

On this interpretation, participants would choose the second blue thing in a sequence, which happens to be a ball (e.g., when the first position contains a blue triangle). While right-branching interpretations must rely on constituency, left-branching interpretations can, but need not do so. A second experiment was undertaken to adjudicate between the right-branching and left-branching interpretation.

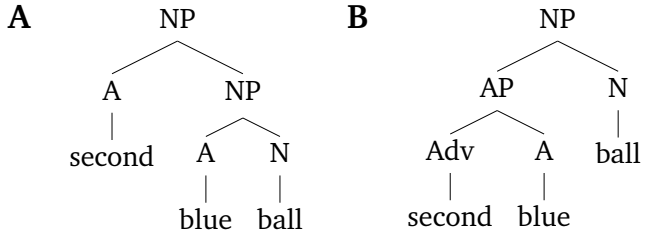


Figure 2.4: Right-branching (A) and left-branching (B) representations for the phrase *second blue ball*.

2.2.2 Experiment 2

As in experiment 1, 20 native speakers of Dutch (15 female, mean age = 23.0 years, age range = 18–28 years) took part in experiment 2 after their written informed consent was obtained. None of the participants were color-blind or had

participated in experiment 1. The experiment was almost identical to experiment 1, except that the set of items in the array also contained blue and green triangles. As there were now two shapes, the noun provided crucial information for the identification of the target. Each trial contained two potential targets. For the target *second blue ball*, the 'right-branching' interpretation, corresponding to the right-branching structure in Figure 2.4A, again refers to the second among blue balls (fifth item in Figure 2.5). The other interpretation, which could be represented in a left-branching structure (Figure 2.4B), refers to the second blue item, which is a ball (third item in Figure 2.5). The right-branching and left-branching interpretations were both always available, but never converged on the same item.

Figure 2.5: Example array for the target *second blue ball*. The left-branching target is in third position, while the right-branching target is in fifth position.

There were again 192 trials. All ordinals were used 32 times in the target phrase. For each ordinal, the target was equally often a blue ball, a blue triangle, a green ball, and a green triangle. We made sure that the left-branching and right-branching interpretations never converged on the same item by placing one item with the same color but a different shape as the target at a random position to the left of the left-branching target. In Figure 2.5, this is the blue triangle on the left, which makes the leftmost blue ball the second blue item (the left-branching target). The presence of the blue triangle does not affect the position of the right-branching target, which is the second among blue balls.[5]

Results

The results of experiment 2 are presented in Figure 2.6. Of all correctly answered trials, participants gave a right-branching answer 99.8% of the time. Only five answers were coded as left-branching answer. A logistic regression analysis of output type (right-branching vs. left-branching) confirmed that participants gave more right-branching than left-branching answers, β = -6.45, SE = 0.45, Wald z = -14.42, $p < .001$. These findings can only be captured using

[5] Again, the interpretations of targets with the ordinal *first* always converged. The responses to these targets could not distinguish between the two interpretations and were therefore not analyzed.

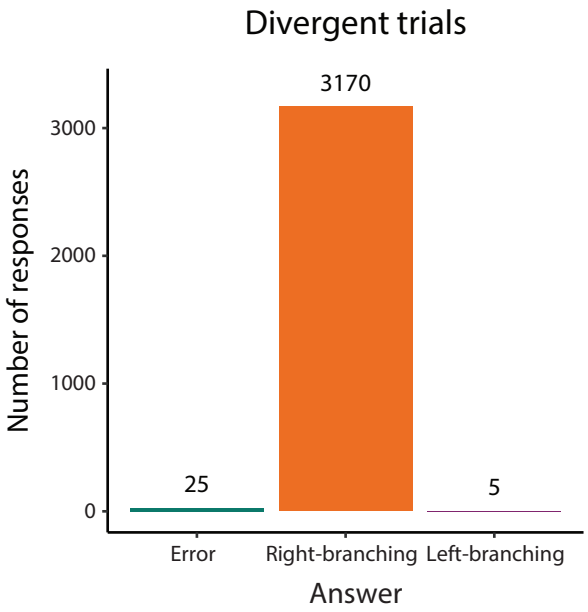


Figure 2.6: Responses in experiment 2.

constituent structure, and therefore provide strong experimental evidence for the importance of hierarchical structure for semantic interpretation.

2.2.3 Computational modeling

Methods

In order to test whether a computational model would show the same bias towards the hierarchical interpretation as the participants did, we trained and tested a state-of-the-art RNN model with a long short-term memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997) on the task of experiment 1. The LSTM model, which was implemented with Keras (Chollet, 2015), had a many-to-one architecture, which is visually represented in Figure 2.7. The input to the model consisted of four one-hot vectors, sequentially presented in four timesteps. Recurrence is indicated by the fact that the model's current state is a function of its previous state (i.e., $a^{<t-1>}$) in combination with the input at the current timestep (i.e., $x^{<t>}$). The input vectors represent respectively the ordinal, color, and shape of the target, as well as the picture. Each input vector had a length of 57, where the first 9 elements were reserved for the words in the phrase (elements 1–6 represented the ordinals second through seventh, 7 and 8 represented the colors blue and green, and 9 represented the shape ball[6]) and

[6]Trials with the ordinal *first* would always be convergent and were therefore not part of the datasets. We replaced *first* by *seventh* in these datasets to make sure that the number of ordinals on which the human participants and the model were tested was the same.

the last 48 elements were reserved for the eight-element picture, wherein each element had one of two colors and the shape ball (i.e., we need three bits to represent each feature). As a result, the picture vector would have 16 ones, so we normalized it to make sure that its net content is 1, in line with the other one-hot vectors. To give an example of an input vector, the word "blue" was represented as a 57-element vector which has a one in position 7 and zeros everywhere else.

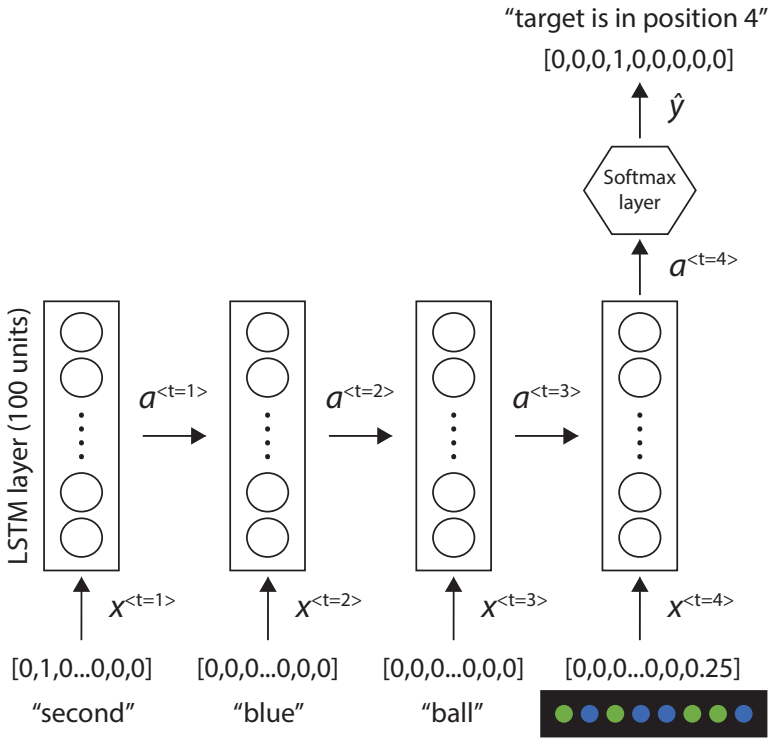


Figure 2.7: Visual representation of a trial for the LSTM, where $x^{<t>}$ represents the input at timestep t and $a^{<t-1>}$ the activation state of the model after the previous timestep.

The hidden layer consisted of 100 units, whose activation function at the last timestep was forwarded to a softmax layer, which provided the output of the network. The output was a nine-element one-hot vector which had a one at the position of the target (positions 2–8) on target-present trials or a one at position 9 to indicate that the target was absent from the picture. In short, the task of the model was to take the words and picture sequentially as input, and provide as output the position of the target.

The LSTM was trained in a supervised manner on datasets of different sizes (100-1000 trials, depending on the training set), in 50 epochs (100 steps per epoch) using the optimizer 'Adam' (optimization using stochastic gradient descent with a learning rate of 0.001) and the categorical-crossentropy loss func-

tion. For each dataset, the model was evaluated on 100 test trials, and this train-test evaluation was simulated 100 times.

Training and test datasets. We trained the LSTM on four different, artificially created datasets. In half of the trials in all datasets the target was present, in the other half the target was absent.[7] While target-absent trials were not included in the behavioral experiments, we did include them in the datasets for the network because this ensures that the network cannot succeed by only paying attention to the ordinal. In all datasets, the training and test trials were mutually exclusive, never containing identical trials. Figure 2.8 presents a visual overview of the different training/test trials.

In the 'linear' training and test set, the linear interpretation was present on target-present trials, and absent on target-absent trials. Moreover, on both target-present and target-absent trials, the hierarchical interpretation was also present, but the output showed that the training data were unambiguously about the linear interpretation, because the trials were always divergent (cf. Figure 2.2B). To give an example, if the target was *second blue ball*, then the second ball in target-present pictures was blue, but it was not the second among the blue balls (i.e., the first ball was green; see Figure 2.8). Here it becomes clear why we included target-absent trials. If the target were always present, there would be a perfect statistical relationship between the ordinal and the output (i.e., *second blue ball* would always lead to target position 2). This could serve as a context-independent statistical heuristic for the model, as it would not need to incorporate information about the color or shape of the target, or about the elements in the picture. By including target-absent trials, we made sure that the model could not succeed by relying only on the information provided by the ordinal.

The 'hierarchical' training and test set consisted of target-present trials in which the hierarchical interpretation was present and target-absent trials in which it was absent. All target-present trials were divergent (cf. Figure 2.2B), so the linear interpretation of the phrase would also be present, but the output was only in line with the hierarchical interpretation. On target-absent trials, the hierarchical interpretation was absent but the linear interpretation was still present. For example, if the target phrase was *second blue ball*, then the second

[7] This was done to make sure that the number of target-present and target-absent responses for each ordinal are roughly equal. However, it also means that output vectors with a one in position 9 ('target absent') are overrepresented in the output. We accounted for this imbalance by updating the loss function with a weighting parameter that reflected the class distribution in the training data (Chollet, 2015).

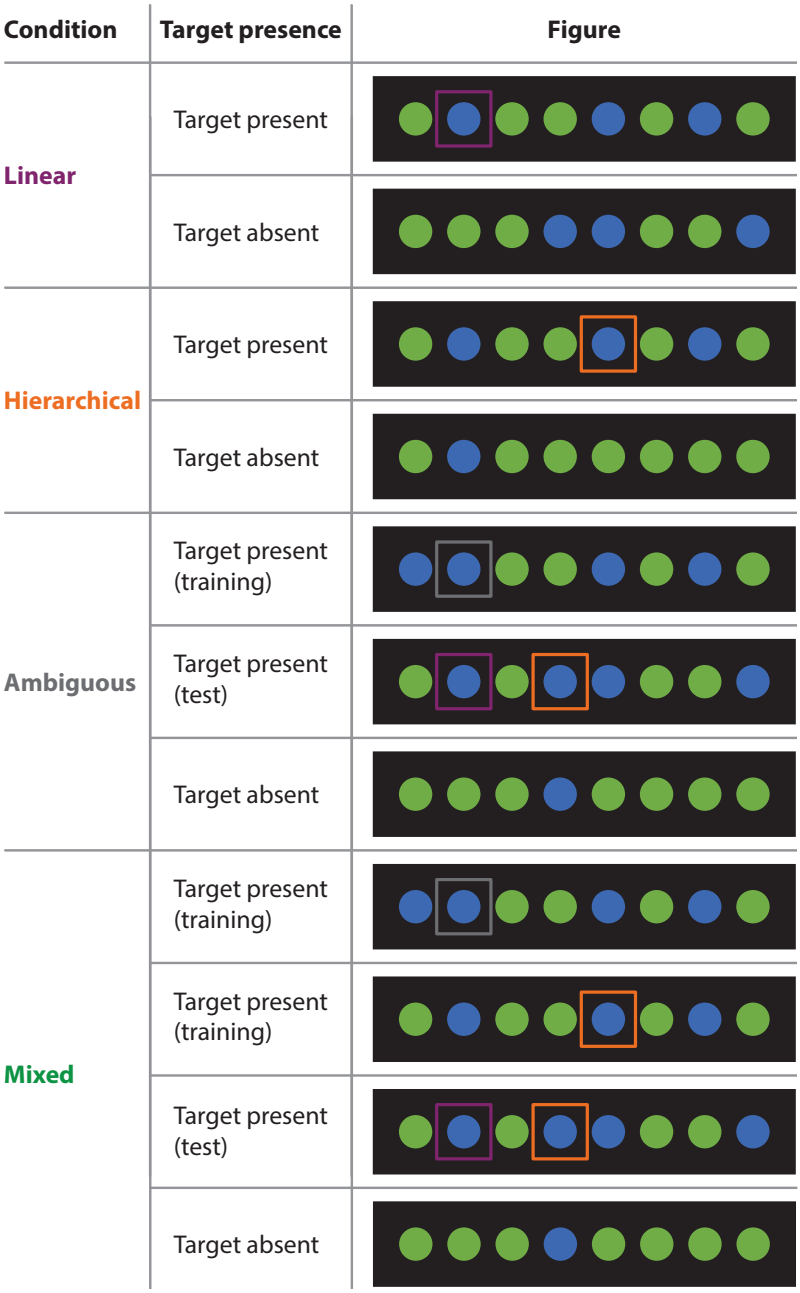


Figure 2.8: Examples of the different training/test trials in the computational simulations, ordered by condition and target presence. The target phrase for these trials is *second blue ball*. The squares in target-present trials indicate the targets for those trials.

ball was blue on both target-present and target-absent trials, but it would not be the second among blue balls (in fact, on target-absent trials the second ball would be the only blue ball; see Figure 2.8).

The 'ambiguous' training set was fully ambiguous between the hierarchical and linear interpretations of the target phrase, both on target-present and target-absent trials. While target-present training trials were always convergent (cf.

Figure 2.2A), target-present test trials were always divergent (cf. Figure 2.2B). The model's answers on these test trials are thus informative about what the model has induced from ambiguous training data. On target-absent training and test trials, neither the linear nor the hierarchical interpretation was present. The ambiguous training set had only 100 trials. This has to do with the fact that target-present training trials are always convergent and thus limited in number, and that the number of unique trials varies per ordinal (e.g., for *seventh blue ball*, there are only two different target-present pictures (one in which the eighth ball is also blue, and one in which it is green), but for *second blue ball* there are 64 different target-present pictures). To make sure that the training and test sets contain roughly the same number of all ordinals, they were both fixed at a size of 100 trials.

The 'mixed' training set contained both ambiguous and unambiguously hierarchical training trials. While the only possible generalization from these data is the hierarchical interpretation, the linear interpretation is compatible with some of the trials. By varying the percentage of ambiguous trials (and thus the ratio between ambiguous and hierarchical trials), we examined how much unambiguously hierarchical data the model needs in order to give hierarchical responses on test trials. The test trials were the same as those used after ambiguous training (i.e., divergent trials).

Generalization to novel items. To further investigate what the model has learned after the hierarchical training regime, we tested its ability to generalize to items that were not seen during training. Specifically, we looked at the model's response to phrases that included the word "red" when the training data did not contain red at all (extrapolation), or only in combination with specific ordinals (interpolation). First, we trained the model on all items (green and blue balls), and then tested it on the phrase "third red ball" and pictures which included red balls. This type of generalization is an instance of extrapolation, because the input contains features (i.e., the word "red", as well as red balls) that were not observed during training and therefore lie outside the training space (Marcus, 1998). Second, we tested the model's ability to interpolate, i.e., to generalize to an item that is composed of known features, and therefore lies within the training space (e.g., Baroni, 2020; Lake & Baroni, 2018). The model was trained on all combinations of features, including the color "red", except the item "third red ball" (e.g., "second blue/green/red ball", "third blue/green ball", and pictures which included red balls). It was then tested on "third red ball".

Here, the training set contains the distributional evidence that "red" and both "blue" and "green" pattern identically, and it contains information about how "third *x*" should be interpreted. Given that "third", "red" and "ball" have all been presented during training, the training data span a distribution that captures "third red ball", even though the combination of these items is new. Given that the new item lies within the parameter space, interpolation can be approached through linear regression. We therefore hypothesize that the model is able to interpolate from known data points to "third red ball". In order to see how well the model extrapolates and interpolates, we simulated each generalization test 100 times. Because the hierarchical model in the main experiment reached over 90% accuracy after 500 training trials (discussed in the results section, Figure 2.10B), we trained the model in each simulation on 500 trials. As in the main experiment, it was evaluated on 100 test trials.

As reported in the results, the model was not able to systematically generalize its 'hierarchical' knowledge to novel items, such as "third red ball". While the training data for the interpolation test contain the information that "red" functions the same as both "blue" and "green", it is possible that this distributional information is not sufficient to indicate the relatedness between these words. That is, there is no intrinsic relationship between the one-hot vectors [0,1...0,0] and [0,0...1,0], although they should be dependent if they are to represent the related words "red" and "blue". In an attempt to test the model's generalization ability when it receives input vectors that are closely related, we used pre-trained word embeddings from Google's word2vec (Mikolov et al., 2013), which have been shown to capture the similarity between related words. The similarity between two multidimensional word embedding vectors can be expressed in terms of the cosine of the angle between them. The closer this cosine similarity value is to 1, the smaller the angle between the vectors and thus the more similar the vectors (see the similarity matrix in Figure 2.9A).

As these word embeddings are 300-dimensional vectors, however, they might lead to overfitting given the limited size and scale of the training data. The model might overcapitalize on redundant aspects of these big vectors, disabling them from dealing with novel input. We therefore used a dimensionality reduction technique based on Principal Component Analysis to reduce the size of the word embeddings to 10 (Shlens, 2014), in line with the size of our vocabulary.[8] This reduces the size of the vectors by maximizing the variance between them, while

[8]Because we have ten words, we would need maximally ten dimensions to capture their differences. In reality, this number can be lower, because some of the words are related.

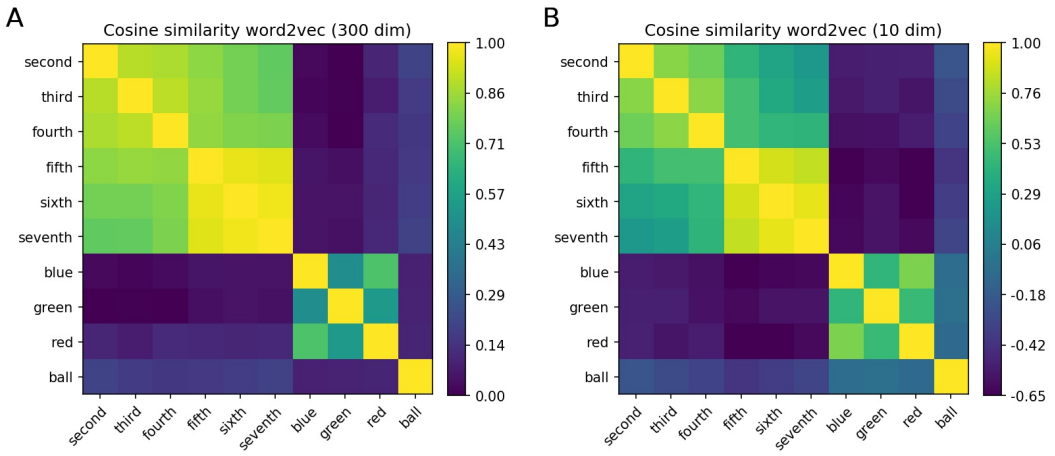


Figure 2.9: Heatmap of the cosine similarity between all 300-dimensional word embeddings (A) and between all 10-dimensional word embeddings (B). Note that in both cases the word embeddings capture the similarity between "blue", "green", and "red", as indicated by a large and positive cosine similarity.

retaining the essence of the original vectors. For our purpose it is important that the similarity between the color words, which is the property over which generalization is evaluated, is retained after dimensionality reduction (Figure 2.9B). We repeated the two generalization tests described above (including separate train-and-test evaluations) with both the full 300-dimensional word embeddings as well as the reduced 10-dimensional embedding vectors.

Results

We evaluate each model's performance by comparing its predicted output on the test input to the correct test output. Each unit in the output layer of the model contains an activation value which can be interpreted as the likelihood that that unit corresponds to the position of the target, given the input (activations sum to one). We took the index of the output unit with the maximum activation value to be the model's predicted output. This value can be seen as the specificity of the model's prediction. For instance, if a model has learned to interpret the phrase *second blue ball* hierarchically, then given the picture in Figure 2.2B it outputs a vector with a high activation value for the fourth element (i.e., the target position, which has a one in the one-hot output vector used during training) and low activation values for all the other elements. We show the specificity of the predicted output (bottom graphs in Figure 2.10 each show these predictions for the test trials of one simulation), and evaluate the accuracy of these predictions by comparing them to the correct output (i.e., top graphs in Figure 2.10 show the

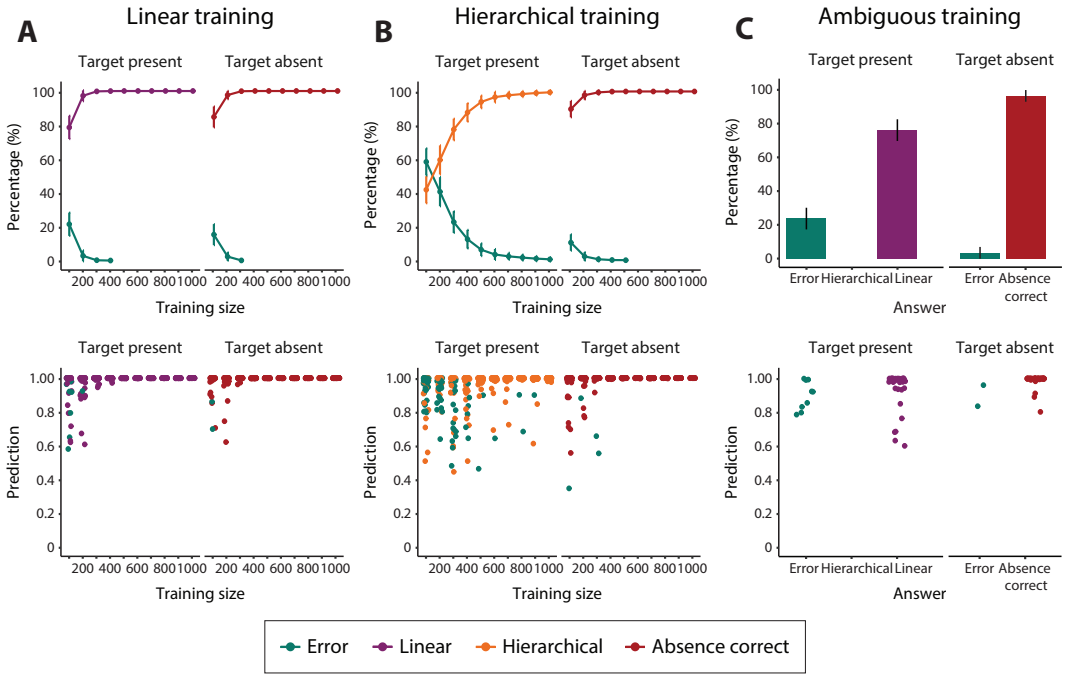


Figure 2.10: Model performance after linear training (A), hierarchical training (B), and ambiguous training (C). Results are divided into average accuracy over 100 simulations (top; error bars represent standard deviation) and specificity of predicted output (activation of output unit with largest value) on the test trials of one simulation (bottom).

average percentage correct, which is the frequency with which the predictions match their labels).

When the model was trained on linear data, it quickly reached very good performance. After 400 training trials, the model scored perfectly, reaching an average accuracy of 100% (Figure 2.10A). After training sizes of 100 and 200, the model makes on average 19 and 3 errors, respectively. These all have to do with the presence of the target: the model either gives a target-absent response on a target-present trial (i.e., 'miss'), or it gives an incorrect linear response on a target-absent trial.

After 100 hierarchical training trials, the model reaches an average accuracy of 65%. The majority of its errors are wrong (but not linear) answers on target-present trials. The model's performance steadily increases with increasing training size up to 700 trials, after which it stabilizes around 97–100% correct on target-present trials (Figure 2.10B). The hierarchical model needs more training data to reach high accuracy than the linear model, which probably has to do with the statistical variance in the hierarchical output data: whereas *second blue ball* on linear target-present trials always maps to position 2, the same target on hierarchical target-present trials can be in positions 3–8. More generally, the

effect of hierarchy on interpretation in terms of statistics (i.e., in the form of an input-output mapping in our experiment) is inconsistent because it reflects information that is not directly encoded in the linear properties of the (input or output) signal.

In order to evaluate whether the model gives more linear or more hierarchical answers after being trained on ambiguous data, we simulated this evaluation 100 times. The model was trained on 100 different datasets of 100 ambiguous (convergent) trials, and at each simulation evaluated on 100 unambiguous (divergent) test trials. The model gets absence correct on most target-absent trials ($M = 96.5$, $SD = 3.47$), see Figure 2.10C. Importantly, on target-present trials it gives mainly linear answers ($M = 76.2$, $SD = 6.43$), and never gives a hierarchical answer (see the empty column for 'hierarchical' in Figure 2.10C). On average, the model makes 14 errors, which are of the same type as those made by the 'linear' model (i.e., misses, or incorrect linear answers on target-absent trials).

To evaluate how much unambiguously hierarchical information the model needs to start generalizing hierarchically, we trained it on a mixed dataset with different ratios between ambiguous and unambiguously hierarchical trials. This ratio ranged from 10:0 (fully ambiguous) to 0:10 (fully hierarchical). Note that these mixed training data are always fully compatible with the hierarchical interpretation. What varies is the number of trials that is also compatible with the linear interpretation. Each mixed training set contained 100 trials, and we simulated each train-test evaluation 100 times. Figure 2.11 presents the responses for each of the different ratios. What is clear from the figure is that the more unambiguous evidence for the hierarchical interpretation in the training set, the more the model converges on the hierarchical interpretation in the test set. What is notable is that this increase is gradual: there is never a point at which the model 'realizes' that the hierarchical interpretation is the only correct generalization (i.e., the model does not induce a rule). Instead, it always gives a substantial proportion of linear answers, even when 90% of the training data are unambiguously hierarchical and only 10% are ambiguous. Moreover, the number of errors on target-present trials increases as there is more unambiguous evidence for the hierarchical interpretation. This matches the patterns seen after linear and hierarchical training. The model initially only considers the linear interpretation, on which it does not make many errors (cf. Figure 2.10A), but the increasing evidence for the hierarchical interpretation is also taken as increasing evidence against the linear interpretation, so the model will give less linear responses.

However, it still does not always get the hierarchical answer right, which is why its error rate increases (cf. Figure 2.10B). In all, these results again show that the model can learn to answer 'hierarchically', but that it needs (a considerable percentage of) unambiguous trials to overcome its non-hierarchical bias.

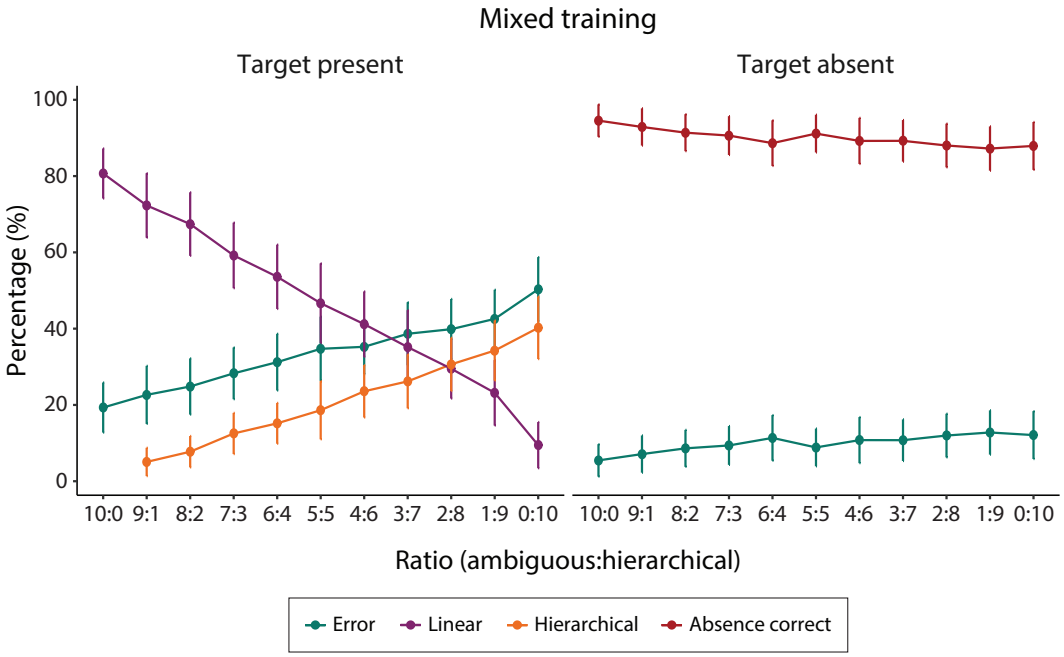


Figure 2.11: Model performance on hierarchical test trials after mixed training data. The training sets are composed of different ratios between ambiguous and unambiguously hierarchical trials, ranging from fully ambiguous data (10:0) to fully hierarchical data (0:10).

Extrapolation and interpolation. We then probed the hierarchical model's ability to extrapolate and interpolate to novel items that were not seen during training. Figure 2.12 presents the model's accuracy, defined as the percentage of correct hierarchical answers, on both generalization tests as a function of the input vectors that represented the words in the phrases. On the extrapolation test, the model did not generalize very well, regardless of whether it was trained on one-hot vector representations (Mean accuracy = 12.6, SD = 9.20), reduced word embeddings (M = 12.7, SD = 5.44) or full word embeddings (M = 10.7, SD = 5.89). In order to see whether these accuracies differ from chance level, we ran 100 simulations in which the training data consisted of pseudorandom mappings between input (phrase, picture) and output (target position). These contained the same information as the other simulations, and included one-hot vectors as the input layer. The model was tested on "third red ball". Given that there are 6 attested outputs in the hierarchical training regime for the ordinal

"third" (i.e., positions 4 through 9), and that there is no consistent statistical relationship between the target and the output (i.e., there is nothing to learn, beyond the fact that "third" cannot be in the positions 1-3), this model scores around chance level of 16.7% accuracy. Comparison of the four groups (one-hot, reduced embeddings, full embeddings, random) through a one-way ANOVA in R (R Core Team, 2020) reveals that the accuracies between groups were different, $F(3,396) = 16.7$, $p < .001$, but a post-hoc Tukey test showed that none of the conditions scored above chance. In fact, they all scored slightly below chance: one-hot vs. random: $\Delta = -4.63$, 95% CI [-7.12, -2.15], $p < .001$; reduced word embedding vs. random: $\Delta = -4.53$, 95% CI [-7.01, -2.05], $p < .001$; full word embedding vs. random: $\Delta = -6.53$, 95% CI [-9.01, -4.05], $p < .001$.

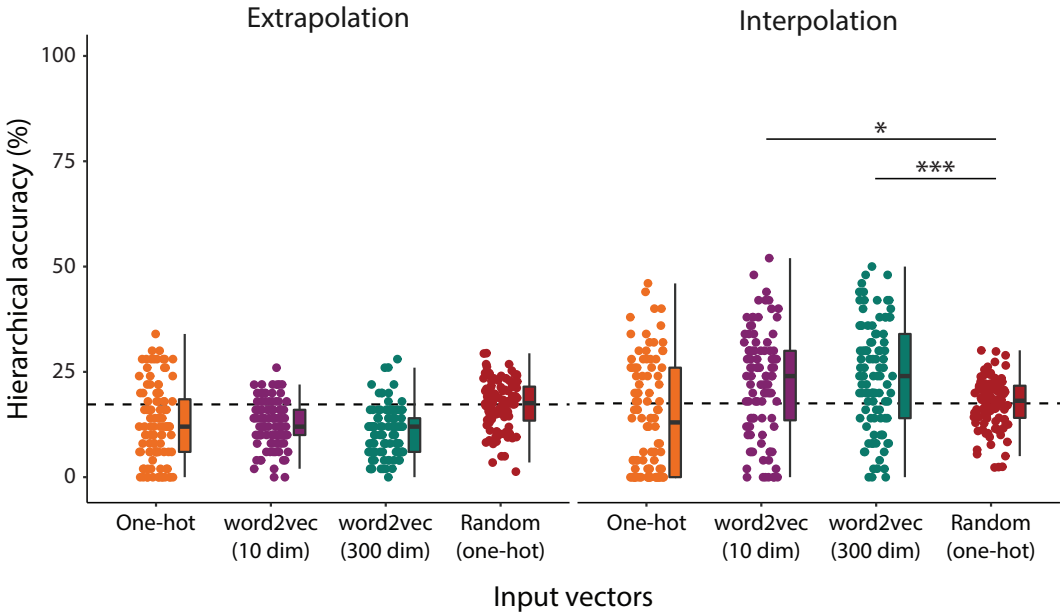


Figure 2.12: Percentage of correct hierarchical responses on both generalization tests after training and testing on different input vectors. Each drop reflects the average hierarchical accuracy on one simulation run (100 simulations per evaluation). The dashed horizontal line reflects the mean accuracy of the model after pseudorandom training, thus representing chance level. *$p < .05$, ***$p < .001$.

On the interpolation test, the model reached higher accuracy for each type of input vector: one-hot vectors ($M = 13.9$, $SD = 13.1$), reduced word embeddings ($M = 22.2$, $SD = 12.4$) and full word embeddings ($M = 23.8$, $SD = 22.2$). We again consider chance level to be around 16.7%, because the input "third red ball" during training could only be followed by a one-hot output vector with a one in either of the six positions 4–9. To evaluate each model against this chance level, we computed the model's performance after it was trained on

pseudorandomly generated data, as described above. Comparison of the four groups (one-hot, reduced embeddings, full embeddings, random) again reveals that the accuracies between groups were different, $F(3,396) = 15.5$, $p < .001$. A post-hoc Tukey test showed that the accuracy for the full and reduced word embeddings was higher than expected by chance (full word embedding vs. random: $\Delta = 6.32$, 95% CI [2.15, 10.5], $p < .001$; reduced word embedding vs. random: $\Delta = 4.66$, 95% CI [0.49, 8.84], $p = .02$). Interpolation accuracy for one-hot vectors was not different from chance. Despite this slight increase in accuracy for the model when trained on word embeddings, overall these findings show that the model was not able use the information it had induced from hierarchical training to systematically generalize to unseen items.

2.3 Discussion

In two behavioral experiments, we show a strong preference for hierarchy in human language interpretation: people's interpretation of ambiguous noun phrases categorically follows from their hierarchically organized syntactic structure. In line with a long tradition of research, our findings support the idea that humans represent noun phrase structures in terms of hierarchical relations rather than linear order (Alexiadou et al., 2007; Cinque, 2005; Culbertson & Adger, 2014; Hamburger & Crain, 1984; Jackendoff, 1972; Martin et al., 2020; Pinker, 1999). In addition, we trained and tested an LSTM model on a computational version of the experimental task, and showed that the model can learn to give hierarchical answers if it is trained on unambiguously hierarchical datasets. However, when the training data contain both unambiguously hierarchical as well as ambiguous trials, the model strongly favors the linear interpretation, even though the hierarchical interpretation is a better fit to the data. Moreover, the 'hierarchical' model does not systematically generalize to novel items that are not seen during training. These findings show that the model behaves unlike humans when the training data are ambiguous, and suggest that it needs different inductive biases in order to achieve human-like generalization.

A comparison between the performance of the model and the behavior of the human participants reveals a number of critical differences. First of all, while the model learned to give hierarchical answers, it only did so when it was explicitly fed unambiguously hierarchical information during supervised training. When the training data were ambiguous with respect to the correct representation underlying the noun phrases, the model had a strongly linear bias, never giving a

hierarchical answer during the test phase. When the training data were mixed to contain both ambiguous and unambiguously hierarchical trials, such that the hierarchical interpretation was the only generalization fully compatible with the data (i.e., the linear interpretation was only compatible with ambiguous trials), the model still had a strongly linear bias. This suggests that the model can learn to answer hierarchically, but that it needs a substantial percentage of unambiguous trials to overcome its non-hierarchical bias (cf. McCoy et al., 2018).

The point about the apparent need for unambiguously hierarchical information during supervised training is relevant because children are not taught to interpret language hierarchically, but come to do so naturally, despite strongly deficient and ambiguous input data (e.g., Berwick, Pietroski, et al., 2011; Crain, 1991; Gleitman & Newport, 1995; Kam & Fodor, 2012; Legate & Yang, 2002; Lidz et al., 2003). While we do not believe that adult language users have not been exposed to unambiguous data, it does seem to be the case that humans have a bias to interpret language in accordance with its underlying hierarchical structure (Crain & Nakayama, 1987; Crain et al., 2017; Ferrigno et al., 2020; Flaherty et al., 2021; Hunsicker & Goldin-Meadow, 2012; Kam & Fodor, 2012; Martin et al., 2020; Yang et al., 2017). The effect of such biases is particularly clear when people consistently generalize over hierarchical structure rather than linear order, despite the fact that these generalizations are underdetermined by the training data (Culbertson & Adger, 2014; Ferrigno et al., 2020; Martin et al., 2020; Morgan & Ferreira, 2021). This learnability scenario also applies to the interpretation of phrases such as *second blue ball*, even when the input does contain unambiguous data. There might indeed be positive evidence in the linguistic input to suggest that such a phrase should be interpreted as a hierarchical structure, but this does not yet rule out the interpretation derived from a linear structure. As is the case in most linguistic examples of ambiguity, evidence for interpretation A is not necessarily evidence against interpretation B. In our behavioral experiments, however, participants categorically interpreted *second blue ball* hierarchically, completely ignoring the linear interpretation, even though that linear option was always present. The strong preference to interpret these phrases hierarchically is suggestive of an inductive bias for hierarchy. Computational models without such a hierarchical inductive bias will often interpret ambiguous linguistic input in line with the linear generalization, because that is the simpler statistical mapping between input and output sequence (Frank et al., 2013; McCoy et al., 2018, 2020). Indeed, it has been shown that RNNs have an

architectural bias for dependencies over shorter (linear) distance (Christiansen & Chater, 1999).

In addition, the hierarchical model was not capable of systematic generalization to novel items. We showed this by evaluating its ability to extrapolate (i.e., generalize to "third red ball" when the training data does not contain "red") and interpolate (generalize to "third red ball" when the training data contains "third", "red" and "ball", but not in combination) as a function of different types of input vectors (i.e., one-hot vectors and word embeddings). On the extrapolation test, the model did not perform above chance level, even if it was trained and tested on word embeddings from word2vec (Mikolov et al., 2013). This is in line with previous studies which show that RNNs are not able to generalize to items that are not observed during training (Hupkes et al., 2020; Lake & Baroni, 2018; Loula et al., 2018), a consequence of the training algorithm also called input independence (Marcus, 1998, 2001). Note that the model's responses to items with the word "red", while labelled as errors, are technically not incorrect. Because the training data never contained "red" as possible input, every induction for a new item containing "red" is statistically legitimate (Marcus, 1998, 2001). Importantly, however, they differ sharply from what humans do. That is, modification in natural language is systematic, in that it applies in the same way to all variables of the right type. If someone knows how to interpret "second *blue* ball" and "second *green* ball", they interpret "second *red* ball" in a similar way, even if they have never seen "red" as a possible attribute. A well-known example of the productive and systematic nature of linguistic knowledge is children's behavior on the Wug Test: young children know that the plural form of a pseudoword such as *wug* would be *wugs*, even though they have never heard this word before (Berko, 1958).

On the interpolation test, we found that if the model was trained on one-hot vectors, it performed at chance level. When it was trained on word embeddings, however, it scored somewhat higher than chance level, suggesting that it was able to take advantage of the inherent similarity between the word embeddings that represent related words, such as "blue" and "red". In addition, it is possible that the model picks up the statistical information that "red" and both "green" and "blue" occur in the same distributional environments, which would allow it to interpret "third red ball" correctly. However, we again believe that the reason behind this performance differs in a fundamental way from the reason why human cognition can support interpolation (and extrapolation). Human knowledge of linguistic modification relies on a symbolic representation of the way in

which ordinals modify their arguments (i.e., ordinal(x); see the lambda expressions in (1)-(8)), which is why this relation obeys consistency and systematicity. The answers of the participants in our behavioral experiments were categorical: they consistently interpreted the phrases in the same way. The model's performance, instead, is stochastic: it gets the answer to "third red ball" right on about one-fourth of the trials, while making an error on all the other trials. The fact that the model was not able to consistently draw the right generalizations (i.e., the highest average accuracy was 23.8% for the full word embeddings, but even this model sometimes reached 0% accuracy, see Figure 2.12) shows that the model was not capable of systematic generalization. Rather, in line with previous work, it appears that the model is to some extent capable of generalizing in an item-based manner, correctly interpreting novel items when they are composed of known features (Baroni, 2020; Lake & Baroni, 2018; Loula et al., 2018).

The model's inability to systematically generalize to unseen items shows that it achieved its performance on hierarchical test trials without resorting to hierarchical constituent structure (Fodor & Pylyshyn, 1988; Marcus, 2001; Pinker, 1999; Pinker & Prince, 1988). To be clear, this is not to say that hierarchical structure per se is necessary for a system to be able to generalize. A computational system that relies only on linearly structured representations might be able to generalize, certainly if these representations contain symbolic variables to which specific instances can be bound. Our point is that the inability to systematically generalize to novel items suggests that the model does not rely on the type of symbolic constituent structure we believe underlies the responses of the human participants (Martin, 2020; Martin & Doumas, 2017, 2019; Puebla et al., 2021).

To sum up, we showed that an LSTM learns to provide output that is in line with hierarchical representations. However, the way in which the model generalizes is quite different from linguistic generalization by humans: when given ambiguous training data, it never provided hierarchical answers, and when tested on novel items, it did not systematically generalize. These two limitations show that the model's inductive biases and its ostensibly hierarchical knowledge are fundamentally different from human knowledge of language.

2.3.1 Linear models of hierarchical structure

While many contemporary language models achieve impressive performance on a range of language tasks (e.g., machine translation, question answering), they

often break down when evaluated on targeted syntactic tests. The reason is that they are fundamentally sequence-based models: they map one *sequence* onto another *sequence* (hence the term seq2seq models; Sutskever et al., 2014), and thus learn sequentially organized statistical patterns that cannot capture the full complexity of hierarchical syntax. While statistical signatures of hierarchical constituent structure can be found in the sequential structure of a sentence (Thompson & Newport, 2007), and while sequential statistics affect language processing (Townsend & Bever, 2001), that is not to say that sequence statistics is a sufficient basis for language (Chomsky, 1957). Because these models are inherently linear, they do not have a natural way to capture structural ambiguities (e.g., that *she saw the man with binoculars* has two meanings) and structural generalizations between different constructions (e.g., how *what did she see the man with?* relates to only one of these two meanings), which follow from the structured nature of linguistic representations and the structure dependence of linguistic operations.

In addition, the strongly linear bias of these computational models does not readily explain why structure dependence is so pervasive (Berwick, Pietroski, et al., 2011; Crain & Pietroski, 2001; Fodor & Crowther, 2002; Heinz & Idsardi, 2011). If statistical information about sequential properties, such as linear order, were available as the basis for grammatical acquisition, one would expect speakers to adopt linear procedures, and therefore languages with linear dependencies to emerge, because that type of information is abundantly available. For instance, in the large majority of subject-verb agreement dependencies, the subject noun and the verb are adjacent. A language model which is trained and tested on these data can thus predict the correct verb inflection in most cases without accessing syntactic structure (Linzen et al., 2016). When the model is tested on structurally more complex examples, which are less likely to be found in the training data and which require hierarchical structure, its accuracy drops dramatically (Marvin & Linzen, 2018). Yet, for humans this never happens: children universally adopt structure-dependent rules in the face of overwhelming evidence that is in line with linear alternatives (e.g., Crain & Nakayama, 1987; Crain et al., 2017; Gleitman & Newport, 1995; Lidz et al., 2003; Yang et al., 2017).

We noted in the previous sections that under the experimental circumstances in which the model was tested, it appears that statistical analysis of sequentially presented data is not sufficient to model human language behavior. This divergence between model performance and human behavior could be attributed

to roughly two independent factors: differences in cognitive architecture and differences in input data. Regarding input data, we acknowledge that the training data for computational models usually consists of raw texts, which lack rich sources of information that contribute to disambiguating the intended meanings of utterances (see Bender & Koller, 2020 for discussion). While this limits the generalizability of our findings in the same way as it limits most NLP work, we do recognize that other NLP models are trained on much more and more diverse data than what we used in our simulations. It is certainly possible that the LSTM would have performed differently had it been trained on more naturalistic data. Assuming that naturalistic language data contain more evidence in favor of hierarchical structure, we predict that the model's performance on divergent test trials will reveal a stronger preference for the hierarchical interpretation, in line with what we show in our mixed train-test regime (see Figure 2.11).

That being said, even within the limited scope of our training simulation we showed that the model learned to behave 'hierarchically'. It was only after further investigation (in particular, extrapolation and interpolation) that we concluded that this behavior did not arise in the same way as the linguistic behavior of our participants. The difference in quality and quantity of the training data, therefore, does not undermine our argument that hierarchical performance is not directly indicative of human-like hierarchical representations. We believe that progress towards human-like linguistic generalization will benefit from a significant adjustment to the cognitive architecture of these models, such that they are biased to encode constituent structure (for related proposals, see Guest & Martin, 2021a; Linzen, 2020; Linzen & Baroni, 2021). This might eventually turn out to be unnecessary in the sense that a preference for constituency could be learned from the environment, so it need not be innate (e.g., Perfors et al., 2011). Our current results do not speak to the question of innateness. However, what is crucial is not whether these biases are innate or learned, but whether they precede the acquisition of specific grammatical properties. Given the evidence for structure-dependent generalizations in both child and adult linguistic behavior (Crain & Nakayama, 1987; Crain et al., 2017; Flaherty et al., 2021; Gleitman & Newport, 1995; Hunsicker & Goldin-Meadow, 2012; Kam & Fodor, 2012; Lidz et al., 2003; Martin et al., 2020; Yang et al., 2017), we believe that the incorporation of a notion of hierarchy into computational language models is the logical next step in order to build plausible models of human cognition.

In support of the value of this idea, recent results show that endowing neural networks with (syntactic) inductive biases for hierarchy improves their perfor-

mance on complex syntactic tasks (e.g., Chen et al., 2017; Hale et al., 2018; Kuncoro et al., 2018; McCoy et al., 2020; Shen et al., 2019; Wilcox et al., 2019). These biases can be implemented in several ways, by means of both implicit and explicit representations of hierarchy. As an example of the former, the Ordered Neurons LSTM has an architecture in which its memory cells are structurally ordered in such a way that when a higher ordered neuron is updated, lower ordered neurons are forced to be updated as well. Different neurons therefore vary in update frequency, due to which they also vary in the timescale of the information they encode, with higher ordered neurons encoding longer timescales (Shen et al., 2019). As higher nodes in a tree structure represent information spanning over longer timescales, higher ordered neurons learn to encode higher nodes. This network thus comes to represent the hierarchical structure of sentences by discovering an implicit connection between timescale and node height. In contrast to this fully data-driven approach, the Tree-LSTM model is built to represent the hierarchical structure of sentences explicitly (Chen et al., 2017). This model is given the correct syntactic tree structure for every input sentence (which it has to translate), such that its internal representations are biased to encode constituent structure. In contrast to the implicit link between node height and timescale in the Ordered Neurons LSTM (Shen et al., 2019), the Tree-LSTM incorporates syntactic trees explicitly (Chen et al., 2017). An important similarity between the two approaches, however, is that they both rely on the modeler's assumptions about the type of structure that must be represented.

2.3.2 Structure, statistics, or both?

A commonly articulated reason to favor linearity is that hierarchical structure is complex. Therefore, if language use can be equally well captured by a purely linear system, the linear system should be favored on grounds of parsimony (e.g., Frank et al., 2012; Frank & Christiansen, 2018). However, while the hierarchical structure of natural language syntax is indeed more complex than can be modeled by linear grammars (Chomsky, 1956), equivalent metrics of parsing complexity have not been defined for 'linear' vs. 'hierarchical' language use. Thus, without an implementation of syntactic structure building, or at least the identification of the core computations at stake, the simplicity statement is ill-posed. Furthermore, the psycholinguistic evidence that hierarchical structure building is costly comes from the comparison of putatively 'more complex' structure with 'less complex', but still hierarchical, structure (e.g., King & Just, 1991; Waters & Caplan, 2004). And last, appeals to simplicity can only be made when competing

theories have equivalent empirical coverage, which is not the case here because the linearity view cannot account for our behavioral results. To deal with these situations, hierarchical structure might only be used in very specific situations, such as when sentence meaning depends on precise hierarchical structure (*second blue ball* vs. *big blue ball*; e.g., Frank et al., 2012). However, this requires the postulation of both a linear and a hierarchical grammar processor, resulting in a two-system cognitive architecture that is more complex than a one-system architecture that only uses hierarchical syntax (Lewis & Phillips, 2015).

While the debate about hierarchical and linear systems is often couched in terms of hierarchy versus statistics, these two are not mutually exclusive (see Martin, 2016, 2020; Yang, 2004). We believe that probabilistic processes do play an important role in language, as has been shown extensively (e.g., Marcus, 2001; Pinker, 1999; Townsend & Bever, 2001), but that they operate within the boundaries imposed by hierarchical structure, during both language processing (Martin, 2016, 2020) and acquisition (Lidz & Gagliardi, 2015; Yang, 2002, 2004). Finding out where the boundaries lie, i.e., what is the representational level over which probabilities are computed, is an important avenue for future research.

2.4 Conclusion

In conclusion, we have shown that hierarchical structure is a key component of human language interpretation, and that an LSTM only reproduces such hierarchical behavior under highly specific training circumstances. We conclude that without a predisposition to generalize hierarchically, the model is not a cognitively adequate model of human language (Fitch, 2014; Martin, 2020; Martin & Doumas, 2017, 2019, 2020). Beyond language, hierarchical structure might form the basis of other domains of cognition and information processing (e.g., Dehaene et al., 2015; Doumas et al., 2008; Ferrigno et al., 2020; Fitch, 2014; Hummel & Holyoak, 1997; Martin & Doumas, 2019, 2020; Tenenbaum et al., 2011). Figuring out how the brain builds hierarchically structured representations from linear input therefore remains a central question in the science of the human mind.

3 | Constraining cognitive computational models of language

Abstract

Over the past several years, it has been shown that artificial neural network (ANN) models can learn to generate many complex syntactic constructions, seemingly without relying on the type of symbolic structured knowledge conventionally used by language scientists. Despite these positive results, we show that these models do not meet the key scientific demand of cognitive fidelity: they are both too weak and too strong with respect to critical properties of human language. ANNs are too weak because they fail to learn central properties of syntax, such as structure-*dependent* restrictions on interpretation. At the same time, ANNs are too strong because they successfully learn regularities that are not possible in human language. For example, they can make structure-*independent* generalizations that humans never appear to make, generalizations that empirically fall beyond the boundaries of possible human languages. To address these scientific problems, we propose two broad changes to cognitive computational modeling of syntax. The first is a change to the learning objectives used, which should be focused on structure interpretation rather than sequence generation. The second is a change to their cognitive architecture, admitting constraints on possible human languages. We believe that with these changes in mind, ANNs for language can move closer towards integrating valuable insights from across the language and cognitive sciences.

3.1 Introduction

When *cognitive* and *computational* approaches to language are compared, it is striking that they are often not aligned in terms of their goals, scientific methods, and evaluation criteria.[1] For instance, while many cognitive approaches emphasize the importance of *structure* in explaining linguistic phenomena, language learning, and language behavior, computational models are mostly evaluated on their ability to recognize or generate *sequences*. Here, we provide suggestions for narrowing this gap. Specifically, we argue for a conceptual shift in ANN syntactic modeling, which we believe should be focused more on the interpretation of syntactic *structures*, and less on the probability of *sequences*. The chapter is structured as follows: In Section 3.2, we demonstrate the importance of structure dependence in language by presenting evidence from co-reference and binding. These linguistic phenomena illustrate that the relationship between form and meaning in natural language is mediated by hierarchical structure. Current ANNs struggle to capture this mapping, and one core reason for this is that they are trained and tested on a learning objective that is orthogonal to the goal of language comprehension. In Section 3.3, we discuss impossible languages, which are formal languages that are logically possible but that are never realized in natural language. Current language models remain cognitively inadequate because they can easily learn these impossible languages. To address these limitations, we suggest several constraints on and changes to scientific practice in cognitive computational modeling of language. We believe that computational research incorporating these constraints and changes will benefit from integrating what is already known about the structure of natural language systems, and will provide a stronger contribution to the scientific study of language.

3.2 Possible structures, not probable strings

A large body of both theoretical and empirical evidence shows that syntactic generalizations are structure-dependent: they are stated in terms of relations defined over the hierarchical structure of phrases and sentences, rather than the sequential order of their words (Chomsky, 1957, 1965; Coopmans, de Hoop,

[1] We use the term computational models in a particular way, referring to commonly used neural network architectures (e.g., recurrent neural networks, long short-term memory (LSTM) networks, Transformers) that are evaluated in terms of their syntactic abilities. These sequence models are initially domain-general and become language models once they are trained on huge text corpora. Our usage thus excludes tree-bank parsers, Bayesian models, and formal models of language learning, among others.

Kaushik, et al., 2022; Everaert et al., 2015). The semantic interpretation of phrases and sentences is also structure-dependent: their meaning is derived from hierarchically organized constituent structure (Heim & Kratzer, 1998). As a result, the pairing of form and meaning in natural language (mediated by syntax) need not be one-to-one: some forms do have one meaning, but others have multiple meanings (i.e., structurally ambiguous sentences), and still others have zero meanings that are grammatically licensed (i.e., ungrammatical sentences; Pietroski & Hornstein, 2020). An overarching goal of syntax research is explaining this relation, that is, how people are able to identify both what a sentence *can* mean (i.e., its possible interpretations, represented as <form, meaning> mappings) as well as what it *cannot* mean (i.e., <form, *meaning> mappings). In the context of this objective, explaining why certain forms are ungrammatical is only a subgoal, because ungrammaticality is a special case of a sentence having zero meanings (Berwick, Pietroski, et al., 2011; Pietroski, 2015; Pietroski & Hornstein, 2020).

Despite the subordinate theoretical importance of ungrammaticality in cognitive approaches to language, this is exactly what computational language models are commonly evaluated on (for a recent review of these models, see Linzen & Baroni, 2021). These studies investigate whether a statistical model that is trained on a large corpus of (mostly grammatical) sentences can learn to assign higher probabilities to grammatical sentences than to ungrammatical ones. Because of the narrow focus of this objective, any generalization the model comes up with might be tailored to recognizing ungrammaticality and might therefore not generalize to other <form, meaning> mappings. Indeed, it is very well possible that this sequence-based evaluation metric inhibits the model from capturing structure-dependent generalizations, as much of the training data will be in line with structure-independent alternatives. This is problematic because a sequential analysis of sentence structure will often lead the model astray, making inaccurate predictions about the interpretation of novel sentences. The following section provides an example.

3.2.1 Evidence from co-reference and binding

The sentences below reveal an interesting asymmetry in the possibility of co-reference between “he” and “the boy”. Note first that sentences (1) through (4) all have an interpretation in which “he” and “the boy” *do not* pick out the same individual. Notably, (1), (2) and (3) are ambiguous; they also have a reading in which “he” and “the boy” *do* pick out the same referent. The co-referential

interpretation is blocked in (4), in which "he" and "the boy" must have disjoint reference.

(1) When **the boy** finished his essay, **he** was happy.

(2) When **he** finished his essay, **the boy** was happy.

(3) **The boy** was happy when **he** finished his essay.

(4) **He** was happy when **the boy** finished his essay.

Clearly, no generalization based on linear order alone is able to capture these facts. Instead, this asymmetry is captured by a generalization conventionally called Binding Principle C, a structure-dependent constraint that makes use of c-command (Chomsky, 1981; Lasnik, 1976; Reinhart, 1983). Principle C states that the interpretation of referring expressions such as "the boy" cannot be made referentially dependent on another element that c-commands them. C-command is a structural relation, formally defined as follows: a node α c-commands another node β if β is (contained in) the sister node of α. As can be seen in Figure 3.1B, which shows the structure corresponding to sentence (4), "the boy" is c-commanded by "he", so a dependency relation between these two elements is ruled out by Principle C. This is not the case in example (2), because the pronoun "he" is embedded in an adverbial clause and "the boy" is not c-commanded by it (see the structure in Figure 3.1A).

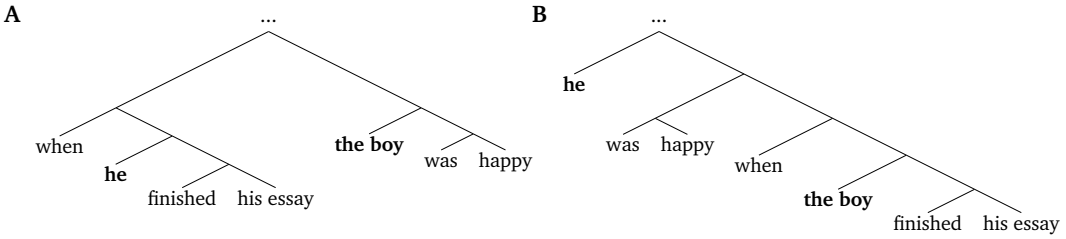


Figure 3.1: Two hierarchical structures in which "he" does (B) or does not (A) c-command "the boy".

Principle C also explains the interpretation of the so-called cross-over construction in (6) (Chomsky, 1986; Freidin & Lasnik, 1981). First consider (5), which is ambiguous; a relation between "who" and "he" can, but need not, be established. In (6), however, no such dependency relation between "who" and "he" can be established.

(5) **Who** said **he** finished his essay?

(6) **Who** did **he** say finished his essay?

Again, linear order alone cannot explain this asymmetry. The difference between these examples is that "who" in (5) is the subject of "said", which is higher in the tree than "he" (see ~~who~~ in Figure 3.2A), while "who" in (6) is the subject of "finished", which occupies a position lower than "he" (see ~~who~~ in Figure 3.2B). The reason that a dependency relation between "who" and "he" in (6) is ruled out is that "he" c-commands the original position of "who" (see Figure 3.2B), a violation of Principle C. This is not the case in (5), as "who" is never in a position where it is c-commanded by "he" (see Figure 3.2A).

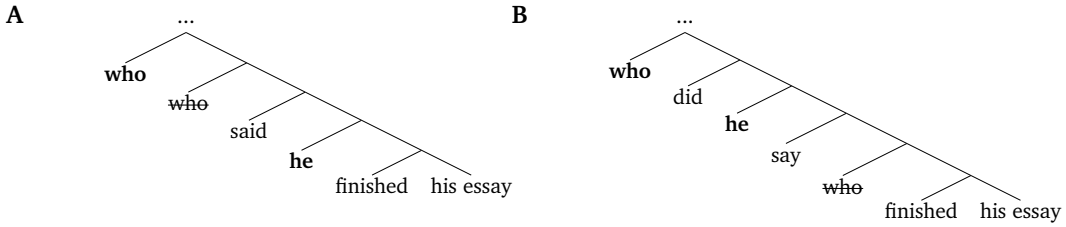


Figure 3.2: Two hierarchical structures in which "he" does (B) or does not (A) c-command the original position of "who" (represented as ~~who~~).

These examples establish that syntactic generalizations are made over *structures* (i.e., the possibility of a referential dependency is based on a particular structural relationship between the two elements), rather than linear order. If language relied on structure-independent principles referring to linear order, the asymmetries would be unexpected, because the examples in (1) and (2) are identical to (3) and (4), respectively, in terms of the linear order of "he" and "the boy". The same goes for (5) and (6), which are sequentially similar (i.e., "who" precedes "he"), but structurally very different. Generalizations based on the sequential structure of sentences alone would therefore make incorrect predictions regarding the interpretation of these sentences. Because (1) licenses the same interpretations as (2), it is not unreasonable to suppose that the interpretations of (3) and (4) are similarly linked, which they are not. Likewise, if (6) is treated as analogous to (5), then (6) should also be ambiguous, contrary to fact. Without considering their underlying hierarchical structure, the unambiguity of both (4) and (6) is mysterious. Before we discuss the implications of these facts for computational modeling of language, we should note that this is but one example of recovering meaning, not specific to co-referential dependen-

cies. Scope-taking is a fundamental part of human language that delimits the meanings that can be expressed in a structure-dependent way.

3.2.2 Implications for computational modeling

The reason that these examples are discussed in so much detail is that they present both an analytic problem and an acquisition problem (Lidz, 2018). The analytic problem was discussed in the previous section, namely that there is an asymmetry in the possibility of co-reference: the fact that sentences (1)-(3) are ambiguous while (4) is not requires an explanation. The acquisition problem is that it is unclear how this generalization can be acquired, in particular if the learner only has access to surface forms, which is the case for ANN models of language.[2]

Computational language models are trained on huge text corpora, often with the task of predicting the next word. After training, their knowledge of syntax can be tested via their response to minimal pairs of grammatical and ungrammatical sentences (e.g., Goldberg, 2019; Gulordava et al., 2018; Linzen et al., 2016; Marvin & Linzen, 2018; Warstadt et al., 2020; Wilcox et al., 2019; Xiang et al., 2021). If the model consistently assigns a higher probability to the grammatical sentence in the minimal pair, it is assumed that it has learned the (principles underlying the) relevant grammatical construction. This evaluation metric is clearly unsuitable for the Principle C constructions discussed above, as they are all grammatical. The relevant difference between (3) and (4) is the availability of a co-referential interpretation in (3), which is unavailable in sentence (4). But the absence of the co-referential interpretation in (4) does not result in ungrammaticality, so this sentence will still be part of the training corpus. It thus seems that a computational model that pays attention only to the frequency of forms, rather than to the relationship between forms and their meanings, will be unable to capture facts of this sort.

Apart from the much-discussed difficulty of linking the non-occurrence of a construction with its ungrammaticality (Chomsky, 1957; Yang, 2015), the additional problem here is that the relevant non-occurrence has to do with meaning. The critical data point for the learner is the non-occurrence of an interpretation, which is not visible in the thing that language models count, namely surface forms (Chierchia, 2004; Crain et al., 2017). On any associative or distributional

[2]Note that the acquisition problem arises irrespective of the correctness of Principle C as generalization. Even if it can be derived from a more primitive syntactic principle (Hornstein, 2009) or instead reflects a pragmatic principle (Ambridge et al., 2014), the question remains how a distributional analysis of surface forms yields the right generalization.

account, what has to be counted are <form, meaning> mappings, not surface forms; it is insufficient to only keep track of which sentences occur (Crain, 1991; Crain & McKee, 1985; Crain & Thornton, 1991). And more specifically, what has to be counted are the interpretations *that do not occur*. From the non-existence of certain interpretations, the learner then has to induce the constraints on how sentence structures map onto sentence meanings.

In other words, the subtle differences between the examples above are differences in meaning rather than form. As meaning is not a part of the sequence, these differences cannot be captured in n-gram statistics and will therefore be difficult to learn for surface-based computational models. Indeed, the one study that tested an ANN's ability to learn co-referential restrictions on referring expressions showed that it was insensitive to Principle C (Mitchell et al., 2019). In stark contrast, children of around three years of age have been found to adhere to Principle C in the interpretation of referring expressions, accepting co-reference in (1)-(3) but not in (4) (Crain & McKee, 1985; Eisele & Lust, 1996; Kazanina & Phillips, 2001; Lidz et al., 2021; Lukyanenko et al., 2014). At the same age, they also know how to interpret cross-over constructions, correctly rejecting the bound interpretation in (6) but not in (5) (Crain, 1991; McDaniel & McKee, 1992), suggesting that the Principle C analysis that unifies the two phenomena is on the right track as an organizational descriptor of the language system.

It has been argued recently that one cannot arrive at meaning from access to forms (Bender & Koller, 2020; Lake & Murphy, 2021). The discussion above aims to show that one also cannot get syntax, or knowledge of a structure-dependent system, from access to forms alone. It follows that focusing exclusively on the acceptability of strings, rather than on the meanings associated with structures, is unlikely to produce a system matching the properties of human language. Changing the learning objective from generating sentences to interpreting structures therefore seems like a principle scaffold in building ANNs' capacity to learn the structure-dependent nature of linguistic principles.

3.3 The limits of variation: Impossible languages

It was long thought that there are no bounds to the variety of human languages. The linguist Martin Joos famously argued that languages could "differ from each other without limit and in unpredictable ways" (Joos, 1957, p. 96). Current views instead hold that typological variation is constrained: not all imaginable combinations of grammatical properties are permissible (Baker, 2001; Cinque,

2013; Newmeyer, 2005; Rizzi, 2009). The fact that certain logically possible languages are never realized, and that languages consistently display the same properties and adhere to the same formal principles, suggests that there are impossible languages (Chomsky, 1991; Moro, 2015, 2016; Newmeyer, 2005).

One such property that does not seem to be adopted in the syntactic rules of natural languages is structure independence: syntactic rules or operations do not refer to the linear order of words in a sentence (Everaert et al., 2015; Rizzi, 2013). A particular view holds that the pervasiveness of structure dependence relates to the fact that structure-independent languages fall outside the boundaries of possible human languages (Baker, 2001; Chomsky, 1991; Moro, 2015, 2016). On this account, structure dependence restricts the class of possible languages, rendering structure-independent languages naturally impossible. We should stress the word *naturally* here, because 'impossible languages' are neither *formally* nor *logically* impossible (i.e., one can think of languages that possess these properties, such as formal languages). Nor are they literally unattainable (i.e., they could be learned in principle), but they appear to be naturally or *biologically* impossible in the sense that they are not acquired by the mechanisms that support natural language acquisition (Hauser, 2009; Lightfoot, 1982; Newmeyer, 2005; Wexler, 1990). This idea receives support from the results of artificial language learning experiments, which suggest that people naturally treat possible and impossible languages as different in kind.

3.3.1 Evidence from artificial language learning

The results of different types of artificial language learning experiments demonstrate that people acquire 'possible' artificial languages differently than 'impossible' artificial languages. In an early study by Read and Schreiber (1982), 7-year-old children were trained to repeat word sequences that were part of a sentence produced by one of two experimenters. They had to learn implicitly which sequences to repeat by looking at what the other experimenter repeated back. In one of the conditions, the to-be-repeated sequence was a syntactic constituent (i.e., subject noun phrase), which in the training data varied in length and semantic content. Most children successfully mastered the task, consistently repeating back the right constituent. When the to-be-repeated sequences were non-constituents, defined in terms of their sequential properties (e.g., the first four words of the sequence), none of the children was successful. These results indicate that children readily infer structural notions like constituent, while be-

ing unable to generalize structure-independent rules that refer to properties of the linear sequence.

Converging evidence comes from a series of neuropsychological studies with a cognitively-impaired subject who has a remarkable talent for learning natural languages (Smith & Tsimpli, 1995). In one of the studies, the subject had to learn an artificial language that had both 'possible' grammatical rules, which were structure-dependent, as well as 'impossible' grammatical rules, which referred to the linear order of words in a sentence (e.g., create the emphatic form of a sentence by adding a suffix to the third word in that sentence; Smith et al., 1993). While being able to learn the structure-dependent rules, the subject failed to learn the structure-independent ones. Interestingly, neurotypical control participants also failed to learn these impossible rules in the linguistic context of the experiment. Yet, when the task was presented to them as a non-linguistic puzzle, they could work out the problem with ease. The latter shows that the rules were not too difficult; when the neurotypical controls could rely on "central strategies of general intelligence", they were able to solve the problems (Smith & Tsimpli, 1995, p. 154). The difference in behavior between the subject and the neurotypical controls suggests that people can learn impossible rules in principle, but only when they do not rely on the mechanisms that support natural language learning.

These findings are corroborated by the results of fMRI experiments by Musso and colleagues, who taught their participants languages with possible and impossible grammatical rules (Musso et al., 2003). Similar to the study by Smith et al. (1993), possible languages contained rules that were structure-dependent. Impossible languages were selectively manipulated versions of real languages, whose rules referred to the linear properties of words in a sentence (e.g., create the interrogative form of a declarative by inverting the linear sequence of the words). After several training sessions, participants could learn both rule types, but they appeared to use different brain systems for them. As performance accuracy increased, activity in the left inferior frontal gyrus (LIFG) increased as well, but only for the learning of possible languages (see Tettamanti et al., 2002 for similar results). When participants were learning impossible languages, LIFG activity decreased as a function of increasing performance accuracy. Subsequent studies with artificial grammars reported similar results, showing that LIFG (in particular, BA44) responds more strongly to hierarchical long-distance dependencies following from a 'possible' phrase-structure grammar than to linear local dependencies determined by an 'impossible' finite-state grammar (Bahlmann et

al., 2008; Chen et al., 2021; Friederici et al., 2006). The observation that the processing of possible and impossible languages involve different brain systems indicates that the brain makes a functional distinction between the two types of languages, and that that distinction is based on whether the languages follow structure dependence or not.

3.3.2 Implications for computational modeling

It thus appears that people naturally generalize structure-dependent rules but have difficulty with languages that exhibit structure-independent regularities. Most ANNs, instead, have been shown to have a preference to generalize linearly rather than hierarchically (Christiansen & Chater, 1999; Coopmans, de Hoop, Kaushik, et al., 2022; Frank et al., 2013; McCoy et al., 2020; Petty & Frank, 2021), which shows that they are able to acquire dependencies that make reference to linear order. Indeed, the few studies that have looked at whether neural network models can learn structure-independent regularities show that the models have no difficulty doing so (Fong & Berwick, 2008; Fong et al., 2013; Mitchell & Bowers, 2020). Fong et al. (2013), for instance, tested whether statistically-trained parsers assign the right parse tree to linearly inverted questions. The training corpus contained declarative sentences with their corresponding parse trees, and contained either natural questions or artificially modified versions of these questions in which the sequence of words was inverted. If the parser was trained on fully natural data, it was not able to assign the right parse tree to reversed questions. Instead, if these unnatural constructions were also part of the training data, the parser performed quite well on reversed questions. This latter finding reveals a non-human capacity to learn, because the training corpus contained (impossible) grammatical idiosyncrasies in which the parse tree for declarative sentences followed completely different rules from that for questions.

More recently, Mitchell and Bowers (2020) trained and tested an LSTM on its ability to handle number agreement in several impossible structures, including constructions in which the linear sequence of words following a special marker was reversed (e.g., “the man whose dog barks ate the apples” becomes “the man whose dog <marker> apples the ate barks”). Here, the correct generalization is structure-independent in terms of both the domain of application (i.e., sequence reversal) and the conditions of application (i.e., it depends on whether the to-be-reversed elements linearly precede or follow the special marker). Mitchell and Bowers (2020) showed that the LSTM’s performance was not affected by

the structure-independent modifications; it handled number agreement in these impossible structures as effectively as in natural language constructions.[3] These findings show that ANNs that are not equipped with constraints on possible languages have no trouble learning impossible languages, which makes them too strong as cognitive models of human language (Adger, 2018; Fong & Berwick, 2008; Pinker & Prince, 1988; Smith & Tsimpli, 1997). It thus seems that computational models of language which aim for cognitive adequacy should be endowed with constraints on possible languages.

Building models that are unable to acquire impossible languages requires being explicit about what is computed by the neural network (Dunbar, 2019; Guest & Martin, 2021a; Rawski & Heinz, 2019). One promising approach towards this goal involves probing the network's units or internal states to see how they relate to its behavior in response to a particular syntactic construction (e.g., Giulianelli et al., 2018; Lakretz et al., 2021). These states might be used as latent proxies for the network's representation of linguistic knowledge and can thus provide information about the way the model infers structure from the input.

Yet, beyond this engineering problem, there is a more fundamental problem, at least for those computational models that are treated as linguistic theories (e.g., Baroni, 2022; Cichy & Kaiser, 2019; Ma & Peters, 2020). Linguistic theories aim to explain why certain logically possible languages are never spontaneously acquired even though they are compatible with much of the data in the linguistic environment. This fundamental question about why human language is the way it is is not being addressed by current language modeling approaches. In fact, if the sequentially-oriented statistical learning mechanisms of ANNs alone were the basis for grammatical acquisition, one would expect human languages to look quite differently (Adger, 2019; Coopmans, de Hoop, Kaushik, et al., 2022; Crain & Pietroski, 2001; Heinz & Idsardi, 2011; Jackendoff, 1988). For instance, ANN-generated languages might contain linear agreement dependencies (Adger, 2018) or probabilistic variation (Hudson Kam & Newport, 2005), neither of which are characteristically found as grammatical properties of human languages. From a scientific perspective, the current computational approach thus misses an important point, which is not that models can learn structure-dependent rules by computing statistics over sequences (which is an engineer-

[3] Another interesting finding was that the LSTM weights that handled agreement in possible and impossible structures overlapped substantially, suggesting that the model does not make a fundamental distinction between the two types of structures. This contrasts with humans, who seem to employ different neurocognitive resources for possible and impossible languages (Musso et al., 2003; Tettamanti et al., 2002).

ing problem), but that languages with sequential dependencies are ultimately never spontaneously acquired by language learners (Berwick, Pietroski, et al., 2011). Explaining why structure-dependent rules are universally attested remains a fundamental challenge for cognitively-oriented computational work on language.

3.4 Constraining cognitive computational models

Broadly speaking, a cognitively faithful computational language model must meet two requirements. On the one hand, it should learn (or at least approximate) what humans learn by making the inductions humans (could) make. Thus, it must succeed in learning possible languages – ideally, following the learning trajectory children follow. This is a prominent topic of current research, which focuses on the question whether neural network models can produce the linguistic behaviors humans produce (Coopmans, de Hoop, Kaushik, et al., 2022; Goldberg, 2019; Gulordava et al., 2018; Linzen et al., 2016; Lakretz et al., 2021; Martin & Doumas, 2017, 2019; Marvin & Linzen, 2018; Warstadt et al., 2020; Wilcox et al., 2019; Xiang et al., 2021). However, a positive finding on this benchmark is not necessarily evidence for a human-like cognitive architecture, because the model's learning mechanisms might work equally well for properties that are never found in human languages (for a general perspective, see Guest & Martin, 2021b). On the other hand, the model must not make inductions humans do not make, and therefore not learn what humans do not learn. In other words, an adequate model must not be able to learn impossible languages. While this aspect has been a central topic in linguistics, it is often overlooked in the computational modeling of language. The few studies that have been done on this topic suggest that current models are not constrained enough to be prevented from learning impossible languages (Coopmans, de Hoop, Kaushik, et al., 2022; Fong et al., 2013; Mitchell & Bowers, 2020).

To address these challenges, we propose two changes to the practice in computational language research: one to the modeling objective and one to the model's architecture (see also Berent & Marcus, 2019; Guest & Martin, 2021a; Marcus, 2001; Martin, 2016, 2020; Rawski & Heinz, 2019). The first change involves shifting the focus from the (un)grammaticality of forms to the meaning associated with these forms. The second involves endowing the models with constraints on possible <form, meaning> mappings, such that these mappings are structure-dependent. One approach towards integrating a constraint like

structure dependence in ANNs would be to split the model's task up into two parts. The first part is independent of the experimental objective and consists of assigning constituent structure to sequential input. In the second part, the model will perform its specific language task (e.g., next-word prediction, natural language inference) using that constituent structure rather than the (sequence of the) individual words. As a consequence, the model will treat constituents as the relevant unit for the language task and might therefore be able to use the structural relations that can be derived from them (e.g., c-command). The specific implementation of this proposal will have to be spelled out in more detail, but we do think that an approach along these lines is necessary for ANNs to be useful as cognitive computational models of language.

In particular, we suspect that incorporating our suggestions will yield progress on the two scientific requirements identified above. As constraints delimit the space of possible hypotheses within which the statistical learning mechanisms of ANNs can discover patterns, they reduce the number of possible generalizations and therefore ought to facilitate learning. One clear benefit of this approach is that it will allow for more targeted testing of the learnability of domain-specific syntactic principles, such as Principle C. If models are prevented from (or at least biased against) forming linear generalizations, they might be able to learn such principles using training data that are quantitatively and qualitatively similar to the primary linguistic data children are exposed to (Futrell et al., 2019; Wilcox et al., 2019).

The incorporation of constraints will also prevent shortcut learning. Neural network models often rely on heuristics to master the task at hand (Geirhos et al., 2020; Malhotra et al., 2020). This can be seen most clearly in their unpredictable behavior in response to adversarial examples, which are those cases where a functionally marginal change in the input, often invisible to the human eye (e.g., in perceptual classification tasks), drastically worsens model performance (Dujmović et al., 2020; Szegedy et al., 2014). One of the reasons that many neural network models are vulnerable to adversarial attacks is that they treat all information as equally relevant and therefore fail to make a distinction between those features of the input that are merely contingent associations (statistics) and those that are inherent properties of the to-be-represented domain (Heaven, 2019; Marcus, 2018). While ANNs are sensitive to any type of statistical regularity in linguistic input, there are statistical regularities that humans do not appear to be sensitive to. In fact, many statistically prominent features in the input are simply ignored, and this strongly facilitates language learn-

ing (Hudson Kam & Newport, 2005; Gagliardi & Lidz, 2014; Pinker & Prince, 1988; Yang, 2013). By using adversarial examples as a test objective, it is possible to determine whether a model uses those features that are inherent properties (e.g., in subject-verb agreement, the structural relation between subject and verb) or instead relies on superficial information that only happens to be correlated with these inherent properties (e.g., the linear adjacency between noun and verb). Extrapolating this idea from the perceptual to the cognitive domain, one could use the learnability of impossible languages as an adversarial example to expose shortcut learning in computational modeling of language (Fong & Berwick, 2008). If, after being trained on artificial data, the model succeeds in acquiring properties that are never found in human language, it is reasonable to suppose that it relied on information humans do not use. The apparent success of such a model is therefore less informative about human cognition, and it must be augmented by structure-dependent constraints on form-meaning mappings in order to be regarded as a cognitively faithful model of human language.

3.5 Conclusion

By focusing on probable sequences rather than on possible mappings between structure and meaning, ANNs can approach, though not fully capture the hierarchical nature of language. We argue that current language models are both too strong and too weak as cognitive models of language, and propose two changes to these models to make them meet the demand of cognitive fidelity. The first involves changing the learning objective from the generation of sequences to the interpretation of structures. The second involves incorporating constraints on possible structures. We believe that if both suggestions are taken seriously, computational language modeling research will have greater impact in the scientific study of language.

4 | Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech[1]

Abstract

Recent research has established that cortical activity 'tracks' the presentation rate of syntactic phrases in continuous speech, even though phrases are abstract units that do not have direct correlates in the acoustic signal. We investigated whether cortical tracking of phrase structures is modulated by the extent to which these structures compositionally determine meaning. To this end, we recorded electroencephalography (EEG) of 38 native speakers who listened to naturally spoken Dutch stimuli in different conditions, which parametrically modulated the degree to which syntactic structure and lexical semantics determine sentence meaning. Tracking was quantified through mutual information between the EEG data and either the speech envelopes or abstract annotations of syntax, all of which were filtered in the frequency band corresponding to the presentation rate of phrases (1.1–2.1 Hz). Overall, these mutual information analyses showed stronger tracking of phrases in regular sentences than in stimuli whose lexical-syntactic content is reduced, but no consistent differences in tracking between sentences and stimuli that contain a combination of syntactic structure and lexical content. While there were no effects of compositional meaning on the degree of phrase-structure tracking, analyses of event-related potentials elicited by sentence-final words did reveal meaning-induced differences between conditions. Our findings suggest that cortical tracking of structure in sentences indexes the internal generation of this structure, a process that is modulated by the properties of its input, but not by the compositional interpretation of its output.

[1] Adapted from Coopmans, C. W., de Hoop, H., Hagoort, P., & Martin, A. E. (2022). Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiology of Language, 3*(3), 386-412.

4.1 Introduction

How the brain parses a continuous speech stream into discrete, hierarchically organized units of linguistic representation remains an important question in the neurobiology of language (Giraud & Poeppel, 2012; Martin, 2016, 2020; Meyer et al., 2020). A possible mechanism that the brain might use to extract linguistic information relies on phase alignment between neural activity and quasi-regular properties of the speech signal. This process, called cortical tracking, results from the tendency of neural systems to adjust to the timing of (quasi-)regular aspects of external stimuli, and has been argued to facilitate segmentation and parsing of continuous speech (for reviews, see Ding & Simon, 2014; Giraud & Poeppel, 2012; Kösem & van Wassenhove, 2017; Obleser & Kayser, 2019; Peelle & Davis, 2012; Rimmele et al., 2018; Schroeder & Lakatos, 2009; Zoefel & VanRullen, 2015).

Cortical tracking is well established for low-level aspects of the linguistic signal, which have clear correlates in the physical instantiation of speech (e.g., the speech envelope). Strikingly, recent work has shown that words and phrases, which are not clearly discernable in the speech signal and have to be internally constructed, are also cortically tracked (Ding et al., 2016; Ding, Melloni, et al., 2017). Moreover, these high-level linguistic properties influence lower-level speech processing, as shown by the fact that cortical tracking of the speech envelope is modulated by the listener's knowledge of the language (Broderick et al., 2019; Di Liberto et al., 2018; Kaufeld et al., 2020).

These studies indicate that the inferred content of a signal affects the extent to which the brain tracks that signal (see also Keitel et al., 2018; Martin, 2020; ten Oever & Martin, 2021). What it is still elusive, however, is which aspects of *content* determine cortical speech tracking. In a recent paper, Kaufeld et al. (2020) showed that the neural signal aligns more strongly with periodically occurring linguistic units, such as syntactic phrases, when these contain meaningful information and are therefore relevant for linguistic processing. Specifically, cortical tracking of phrase structure was stronger for regular sentences than for control stimuli that were matched in terms of either lexical semantics (word lists) or both prosody and syntactic structure (jabberwocky sentences), suggesting that this neural response is driven by the compositional meaning of sentence structures. However, the difference between sentences and these control conditions can be described not only in terms of the output of compositional processing (i.e., the fact that sentence structures have a meaningful compositional interpretation), but also in terms of the factors that go into structural composition. To

investigate which of these aspects of *content* affect cortical tracking of linguistic structure, the current electroencephalography (EEG) study investigates cortical tracking of linguistic units (phrases, words, syllables) when these are embedded in stimuli that are parametrically varied in terms of the amount of linguistic information. These stimuli ranged from regular, compositional sentences to structure-meaning divergent forms (idioms, syntactic prose), structures with reduced lexical-syntactic content (jabberwocky) and unstructured word lists. We thus test how the relationship between structure and meaning in spoken language affects cortical tracking of linguistic information.

4.1.1 Cortical tracking of linguistic structure

Low-frequency cortical activity closely tracks the amplitude envelope of the speech signal (Ahissar et al., 2001; Doelling et al., 2014; Gross et al., 2013; Kayser et al., 2015; Keitel et al., 2017, 2018; Luo & Poeppel, 2007). Because the low-frequency periodicity of the speech envelope correlates with the syllable rate (i.e., in the theta band), it has been argued that cortical activity in this frequency range tracks syllable-sized linguistic units (Giraud & Poeppel, 2012; Luo & Poeppel, 2007; Peelle & Davis, 2012; Poeppel & Assaneo, 2020). However, speech contains temporal regularities at multiple timescales; high-level linguistic units, such as syntactic phrases, also exhibit quasi-regular temporal structure, yet only a small number of studies have investigated cortical tracking of phrase structure.

A main method to study tracking of abstract structure has relied on careful control of the presentation rate of linguistic information, whereby this information is 'frequency tagged'. The idea behind this approach is that when information is presented repeatedly at a specific frequency, the neural response to that type of information synchronizes with its presentation rate. In a series of M/EEG studies, Ding and colleagues have shown that neural activity becomes phase-locked to the presentation rate of phrases and sentences, even though these abstract units are not physically discernable in the auditory signal itself (Blanco-Elorrieta et al., 2020; Ding et al., 2016; Ding, Melloni, et al., 2017; Getz et al., 2018; Makov et al., 2017; Sheng et al., 2019). Such phase-locked responses are found only if the input can be grouped into phrases, showing that they are based on linguistic knowledge, not acoustic information (Ding et al., 2016; Martin & Doumas, 2017). And while it has been disputed that what is tracked is really abstract structure rather than lexical semantics (Frank & Yang,

2018), recent studies have shown that lexical accounts cannot fully explain the data (Burroughs et al., 2021; Jin et al., 2020).

These frequency-tagging studies are very artificial because they rely on synthesized speech that is isochronously presented, but similar effects are reported in studies with more naturalistic materials. In one such study by Keitel et al. (2018), participants listened to naturally spoken sentences that were embedded in noise, after which they had to perform a comprehension task. All sentences were annotated for the occurrence of phrases, words, and syllables, yielding linguistically relevant frequency bands that were specific for their stimulus materials. Within each frequency band, speech tracking was quantified through mutual information between the speech envelope and neural activity. At the timescale of words and phrases, tracking was stronger for correctly comprehended than for incorrectly comprehended sentences, showing that speech tracking in these frequency bands is related to successful language comprehension.

Using a similar approach, Kaufeld et al. (2020) presented participants with naturally spoken stimuli in three conditions: regular sentences, jabberwocky sentences (i.e., same prosody and structure, but different lexical content), and word lists (i.e., same lexical content, but different structure and prosody). Backward versions of all stimuli were used to control for acoustic differences. At the phrasal timescale, speech tracking was stronger for regular sentences than for both jabberwocky sentences and word lists, while these differences were absent in the acoustic control conditions. These findings thus show that the brain is more attuned to phrases when they contain meaningful information and are therefore relevant for language comprehension (Kaufeld et al., 2020). In particular, the fact that phrase-level speech tracking is stronger for sentences than for jabberwocky suggests that this response is modulated by the semantic content of phrases (see also Brennan & Martin, 2020; Martin, 2020; Martin & Doumas, 2017).

It is still an open question, however, whether *semantic content* should be interpreted as lexical-semantic content – the fact that sentences are structured sequences composed of real words – or rather, compositional-semantic content – the fact that these real words in sentences compose into meaningful constituents. The most prominent difference between regular and jabberwocky sentences is that the former contain real content words, which are replaced by pseudowords in jabberwocky sentences. Real words and pseudowords differ in both semantic and lexical-syntactic content, with the latter strongly affecting linguistic structure building (Hagoort, 2005, 2017; Matchin & Hickok, 2020). It is thus possible

to interpret the difference in phrase-level speech tracking between sentences and jabberwocky in two ways: either it reflects the fact that words in sentences can be composed into meaningful constituents (i.e., reflecting the *outcome* of structure building; Kaufeld et al., 2020), or it reflects the fact that the lexical-syntactic information carried by content words allows words in sentences to be easily composed in the first place (i.e., reflecting the *input* to structure building). In the latter case, these findings reflect the brain's attempt to build a structural representation of the linguistic input, regardless of its interpretation. The present study aims to tease apart these two possibilities.

4.1.2 Background of the present study

We contrast regular sentences, whose meaning is compositionally derived from their structure and lexical components, with stimuli in which the mapping between structure and meaning is less transparent. If it is indeed the case that phrase-level speech tracking is driven by the structure-meaning correspondence of sentences, the tracking response should be stronger for sentences than for controls that are divergent in their structure-meaning relationship. As examples of the latter, we used one naturally occurring stimulus (idioms) and one artificial stimulus (syntactic prose), both of which contain the same structural and lexical-semantic information as regular, compositional sentences, but are putatively less compositional in the sense that their meaning does not derive fully from a combination of their structure and lexical components. Parametrically reducing the amount of linguistic information, we also included jabberwocky sentences and unstructured word lists.

We note that compositional processing is not an all-or-none phenomenon (Baggio, 2021; Titone & Connine, 1999), and idioms and syntactic prose are not processed entirely noncompositionally. However, a compositional analysis of the sentences in these conditions either does not yield a sensible interpretation (syntactic prose) or does not yield the intended interpretation (idioms). We therefore assume that compositional processes will be overall less engaged in the comprehension of idioms and syntactic prose than in the comprehension of regular sentences.

Idioms are conventionalized co-occurrence restrictions whose figurative meaning must be learned (Cacciari, 2014; Cacciari & Glucksberg, 1991; Jackendoff, 1995, 2017). They adhere to basic grammatical rules but are semantically idiosyncratic: the figurative meaning of idioms is not fully derived from a semantic composition of their component parts (Cacciari & Glucksberg, 1991;

Jackendoff, 1995, 2017; Sprenger et al., 2006). As an example, consider the Dutch idiom *een vinger aan de pols houden* (literally: "to keep a finger on the pulse"), whose figurative meaning is "to check whether everything goes right". Clearly, this figurative meaning is noncompositional and conventionalized, but in terms of structure the idiom is not an unanalyzed whole. The idiom is a verb phrase whose verb inflects in the past tense in the same way it does in regular sentences (i.e., as in English, *houden* "to keep" is irregular, inflecting to *hield* "kept" in the past tense), and it has the regular argument structure of the verb *houden* "to keep", which is used ditransitively and can be modified by adverbs in the usual way.

The idea that idioms contain regular syntactic structure is supported by evidence from language processing, which shows that the structure of idioms is accessed in both comprehension and production (Cutting & Bock, 1997; Konopka & Bock, 2009; Peterson et al., 2001; Sprenger et al., 2006). This structure is linked to the idiom's meaning in a highly idiosyncratic way, but language users who process the idiom in real time cannot know this beforehand and will therefore initially attempt to derive its interpretation compositionally. Behavioral experiments show that while effects of compositionality can be found in the early stages of idiom comprehension, literal processing can to some extent be terminated after the phrase or sentence is recognized as being an idiom, at which point its idiomatic meaning is retrieved from semantic memory (Cacciari, 2014; Cacciari & Corradini, 2015; Cacciari & Tabossi, 1988; Holsinger & Kaiser, 2013; Libben & Titone, 2008; Peterson et al., 2001; though see Smolka et al., 2007). Evidence from electrophysiological brain recordings also suggests that compositional processes can be interrupted in the comprehension of idioms (Canal et al., 2017; Rommers et al., 2013; Vespignani et al., 2010). We therefore consider idioms suited to serve as experimental sentences whose meaning is not fully derived from their component parts. These effects of compositionality might not be apparent immediately (i.e., before the idiom recognition point), but we suspect that compositional processes will be overall less engaged for idioms than for regular sentences.

In syntactic prose, real words are used to construct syntactically correct but nonsensical sentences (e.g., Bastiaansen & Hagoort, 2015; Kaan & Swaab, 2002; Marslen-Wilson & Tyler, 1980; Mazoyer et al., 1993). As an example, consider the Dutch sentence *een prestatie zal het concept naar de mouwen leiden*, which translates as "an achievement will lead the concept to the sleeves". This sentence adheres to the rules of Dutch syntax, including constraints on word or-

der and argument structure, but a compositional analysis of the sentence does not yield an interpretation that makes sense. Not many studies have investigated the brain processes involved in comprehending syntactic prose, but one relevant study found increased EEG gamma-band power for regular sentences compared to syntactic prose (Bastiaansen & Hagoort, 2015). Notably, two other EEG studies reported similar effects in the gamma band when comparing regular sentences and idioms (Canal et al., 2017; Rommers et al., 2013), tentatively suggesting that the contrasts between sentences and both idioms and syntactic prose affect similar neurocognitive processes.

In addition to these two conditions, we also used jabberwocky sentences and word lists (see also Kaufeld et al., 2020). With these five conditions in total (see examples in Table 4.1), our design parametrically varies the amount of linguistic information present in the stimuli. All conditions except jabberwocky sentences contained real content words, and all conditions except word lists had the same syntactic structure. Moreover, for all syntactically structured conditions, a compositional interpretation can be derived. However, a compositional combination of the words in idioms does not yield their figurative meaning, a compositional combination of the words in syntactic prose does not yield a coherent semantic interpretation, and a compositional combination of the (pseudo)words in jabberwocky sentences is underspecified. In other words, regular sentences differ from the other syntactically structured conditions not in whether they allow compositional processing in principle, but in whether a compositional combination of the structure and lexical components yields a straightforward meaningful interpretation.

4.1.3 The present study

Participants listened to spoken stimuli in these conditions while their EEG was recorded. We quantified cortical tracking between the speech envelopes and the EEG data by means of mutual information (MI), which is an information-theoretic measure that quantifies the statistical dependence between two random variables (Cogan & Poeppel, 2011; Gross et al., 2013; Ince et al., 2017; Kayser et al., 2015; Keitel et al., 2017). MI was computed in three frequency bands, corresponding to the occurrence of phrases (1.1–2.1 Hz), words (2.3–4.7 Hz), and syllables (3.4–4.9 Hz) in our stimuli (Kaufeld et al., 2020; Keitel et al., 2018). Following previous research, we controlled for spectral differences between sentences and word lists by including backward versions of both stimuli. These backward versions preserve many of the spectral properties of their

Table 4.1: Dutch example stimuli of all five conditions.

Condition	Stimulus	Lexical semantics	Syntactic structure	Meaningful compositional interpretation
Sentence	De jongen gaat zijn zusje met haar huiswerk helpen. *the boy will his sister with her homework help* "The boy will help his sister with her homework."	X	X	X
Idiom	De directie zal een vinger aan de pols houden. *the directorate will a finger on the pulse keep* Literal: "The directorate will keep a finger on the pulse." Figurative: "The directorate will check whether everything goes right."	X	X	
Syntactic prose	Een prestatie zal het concept naar de mouwen leiden. *an achievement will the concept to the sleeves lead* "An achievement will lead the concept to the sleeves."	X	X	
Jabber-wocky	De jormen gaat zijn lumse met haar luisberk malpen. *the jormen will his lumse with her luisberk malp* "The jormen will malp his lumse with her luisberk."		X	
Word list	De gaat jongen zusje huiswerk zijn haar helpen met *the will boy sister homework his her help with*	X		

Note. English translations are provided below. Only the underlined words in the idiom stimulus are part of the conventionalized idiom.

forward version (especially rhythmic components) but are unintelligible (Gross et al., 2013; Kaufeld et al., 2020; Keitel et al., 2017; Park et al., 2015).

We were particularly interested in the coherence between speech and EEG in the phrase frequency band. For this measure of phrase-level speech tracking we consider two possibilities. If it is affected by the extent to which a compositional analysis of the input yields a meaningful structural representation, we expect higher MI for regular sentences than for all other conditions. Instead, if phrase-level speech tracking reflects the construction of a structural representation regardless of its compositional interpretation, we do not expect MI for regular sentences to differ from MI for idioms and syntactic prose. Yet, we do predict MI to be higher for regular sentences than for jabberwocky and word lists, because the latter two contain less information based on which a structural representation can be constructed (i.e., cues from argument structure, word order).

4.2 Methods

4.2.1 Participants

We recruited 40 participants (30 female, mean age = 24.6 years, age range = 19–31 years) from the participant pool of the Max Planck Institute for Psycholinguistics. All participants were right-handed native speakers of Dutch, who reported normal hearing and did not have a history of language impairment. After receiving information about the experimental procedures, participants gave written informed consent to take part in the experiment, which was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University Nijmegen. They were reimbursed for their participation. After preprocessing, we excluded two participants due to low numbers of artifact-free trials. The analyses reported are based on a sample of 38 participants. As there are no previous studies that are comparable to our experiment in terms of method, analyses and design, this sample size was based on EEG studies from our lab that either used similar analyses (but had a lower number of conditions; Kaufeld et al., 2020; n = 29) or had a similar design (but used different analyses; Nieuwland et al., 2019; n = 40).

4.2.2 Materials

Experimental items

An example of one stimulus item for each condition is given in Table 4.1. The Sentence condition contained sentences with a compositional meaning, which is derived from a combination of the word meanings and their structural combination. To give an example of a translated stimulus item, the meaning of "The boy will help his sister with her homework" is a function of the meaning of the individual words and the syntactic structure of the sentence. The structure was the same for all syntactically structured conditions (i.e., Sentence, Idiom, Syntactic prose, and Jabberwocky), which start with a noun phrase (NP) and an auxiliary verb (e.g., "The boy will . . . "), followed by a verb phrase consisting of an NP, a prepositional phrase (PP), and a non-finite lexical verb (V), which is phrase-final in Dutch (e.g., ". . . help his sister with her homework").

For the Idiom condition we selected a set of commonly used and well-known Dutch idioms that had the same NP–PP–V structure. The majority of these idioms were selected from stimulus lists shared by Hubers et al. (2018) and Rommers et al. (2013). The idioms were embedded in carrier sentences by the addition

of a sentence-initial NP and an auxiliary verb, which are not part of the conventionalized structure. We only analyzed those idioms that were known to the participants. Idiom knowledge was established for each participant by means of a post-experiment questionnaire (see Section 4.2.6).

Syntactic prose sentences are grammatically well-formed and contain real words, but these are difficult to compose into a coherent semantic representation. The stimulus sets in both the sentence condition and the syntactic prose condition were matched with the idioms on the total number of syllables and on the lexical frequency of the content words (frequencies extracted from the SUBTLEX-NL database of Dutch word frequencies; Keuleers et al., 2010).

Jabberwocky sentences were generated with the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010), which generates pseudowords that obey the phonotactic constraints of Dutch. We created jabberwocky versions of all items in the sentence condition by substituting each content word with a pseudoword that was matched in number of syllables, subsyllabic structure, and syllable transition frequency. The function words (auxiliaries, determiners, prepositions, pronouns) were kept the same, allowing for the construction of the same syntactic structure with a compositional interpretation.

Items in the Word list condition contained the same words as those in the corresponding sentence item, but were scrambled in such a way that no syntactic combinations could be formed.

We created 85 stimuli for all conditions, of which the first five served as practice trials, which were not analyzed. Only the idiom condition had 90 items, which allowed us to preserve roughly the same number of trials as in the other conditions after excluding unknown idioms.

Audio recordings

The stimuli were recorded in a sound-attenuated booth by a female native speaker of Dutch (sampling rate = 44.1 kHz (mono), bit depth = 16). After recording, the intensity of all stimuli was scaled to 70 dB in Praat (Version 6.1.02; Boersma & Weenink, 2019). Backward stimuli for the sentence and word list conditions were created by reversing each stimulus recording in Praat.

Figure 4.1 shows the modulation spectra of all forward conditions as well as the backward version of sentences and word lists. These figures indicate that the forward conditions are prosodically very similar (Figure 4.1A), except for the word list condition (Figure 4.1B), which deviates from the sentence condition at several frequencies (see Supplementary Information S4.1). While not

ideal, we believe that this prosodic difference between stimuli with regular syntactic structure and those without structure is inherent in the contrast between these conditions. Because our main interest is the comparison between the syntactically structured conditions (Sentence vs. Idiom, Syntactic prose, and Jabberwocky), it is important that these conditions do not systematically differ in acoustic properties.

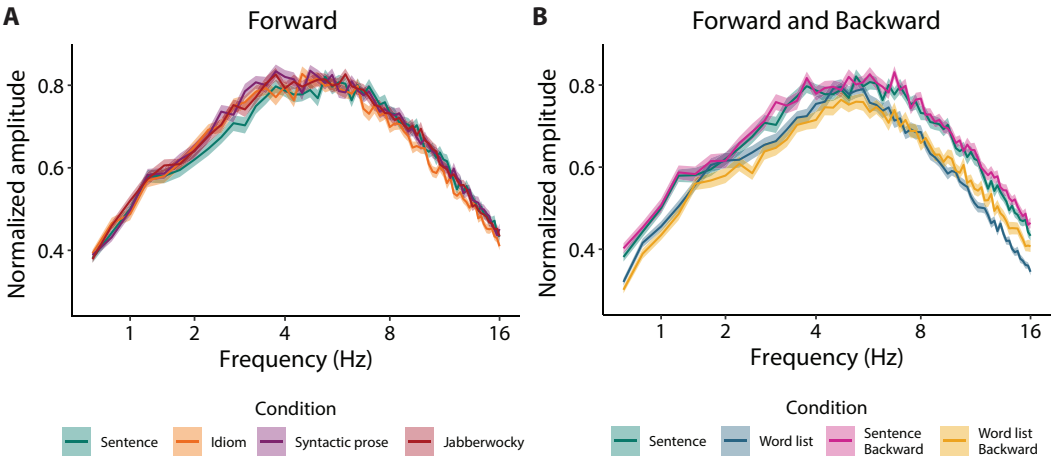


Figure 4.1: Modulation spectra of the forward versions of all conditions, computed following the procedure described in Ding, Patel, et al. (2017). Backward versions of sentences and word lists were included because of the differences between the forward versions of these two conditions.

4.2.3 Annotations

We manually annotated the forward recordings in Praat (Boersma & Weenink, 2019) with respect to the presence of phrases, words, and syllables. Specifically, for each stimulus we annotated the position in the recording where a linguistic unit ends (Figure 4.2A). For both words and syllables, this corresponds to the boundary between successive units. For phrases this corresponds to the position of closing phrase boundaries. For example, in [*de jongen*] [*gaat* [*zijn zusje*] [*met haar huiswerk*] *helpen*], the closing bracket denotes the offset of a phrase whose onset is denoted by the corresponding opening bracket. As word lists by definition do not contain phrases, we marked 'phrases' in these stimuli by annotating the offsets of the words that are at positions of closing phrase boundaries in the corresponding sentence item. In the example sentence above, phrases are closed after the second, fifth, eighth, and ninth word, leading to the following phrase annotation for the corresponding word list: [*de gaat*] [*jongen* [*zusje huiswerk*]

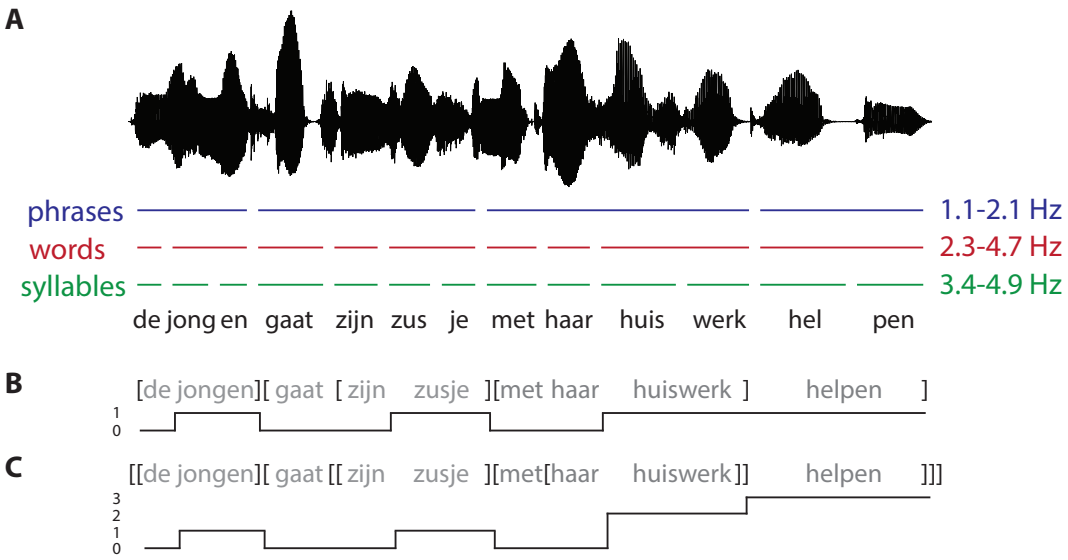


Figure 4.2: Three different annotations of linguistic structure for the Dutch translation of the sentence "the boy will help his sister with her homework". (A) Schematic illustration of the three different timescales of the linguistic units of information (phrases, words, and syllables) contained in the sentence. From the annotation of these timescales, we derived frequency bands for each linguistic unit. (B) Phrase-level annotation, where words that integrate a phrase are coded as 1 for their entire duration, while all other words are coded as 0 (bracket presence). (C) Phrase-level annotation, where the value assigned to each word corresponds to the number of phrases that the word integrates (bracket count).

[*zijn haar helpen*] *met*]. Converting the onsets and offsets of these annotations to frequencies resulted in the following frequency bands: 1.1–2.1 Hz for phrases, 2.3–4.7 Hz for words, and 3.4–4.9 Hz for syllables.

To provide additional evidence that our results index cortical tracking of abstract (syntactic) information, rather than mere acoustic differences between the conditions, we performed an additional MI analysis in which the speech stimuli were replaced by abstract versions of these stimuli in which we only encoded phrase-structure information (Brodbeck et al., 2018; Kaufeld et al., 2020). For each forward stimulus, we marked all time points corresponding to phrase-final words with a 1 and marked all other time points with a 0 ('bracket presence'; Figure 4.2B). Phrase-final words are those words at which syntactic and/or semantic composition can take place. For example, the time points corresponding to the underlined words in the sentence [*de jongen*] [*gaat* [*zijn zusje*] [*met haar huiswerk*] *helpen*] (i.e., *boy, sister, homework, help*) were marked by a 1, because they close syntactic phrases, while all other time points were marked by a 0.

Again, annotating phrase-final words in word lists is impossible, so we marked phrases in the same way as described above, marking time points corresponding to words with a 1 if these words are in a position that indexes a phrase-final word in the corresponding sentence.

These abstract annotations of bracket presence are actually insufficient to represent phrase structure, because sentences are hierarchically embedded structures rather than linearly concatenated phrases. To represent this property, we incorporated 'bracket count' as yet another type of abstract annotation (Brennan & Martin, 2020; Brennan et al., 2012, 2016; Nelson et al., 2017), which is correlated with bracket presence but contains more detailed syntactic information. This variable counts the number of phrases that are completed at a particular word (derived from bottom-up tree traversal), corresponding to the number of closing brackets in [[*de jongen*] [*gaat* [[*zijn zusje*] [*met* [*haar huiswerk*]] *helpen*]]]. The value assigned to each word for its entire duration corresponds to the number of phrases that the word integrates (Figure 4.2C).

4.2.4 Experimental design

Participants listened to all stimuli in all seven conditions, which were presented in a block design. The order in which the seven blocks were presented was pseudorandomized, with the following constraints: the two backward conditions were never presented in adjacent blocks, and the block with word lists and the block with idioms always preceded the block with sentences. Regarding the word lists, this presentation order was used to reduce the possibility that participants would project (their memory of) the phrase structure of the sentences onto the word lists. Regarding the idioms, this order was used to reduce the possibility that participants would try to derive their meaning compositionally. Within each block, the order of the items was randomized.

4.2.5 Procedure

Participants were individually tested in a soundproof booth. They were instructed to attentively listen to the audio, which was presented over loudspeakers, while looking at a fixation cross displayed at the center of the screen. After each trial, participants had to advance to the next trial by pressing a button. They were allowed to take short breaks between blocks. The EEG experiment lasted approximately 60–70 minutes and was followed by an idiom knowledge test.

4.2.6 Idiom knowledge test

The EEG experiment was followed by a digital questionnaire in which participants were asked to indicate whether they knew the figurative meaning of the idioms that were presented in the experiment. For each idiom, they had to indicate this on a keyboard. If they answered "yes", they had to type the meaning using the keyboard. If they answered "no", they were asked to indicate what they thought the meaning could be. Idioms were coded as 'known' when the participant answered "yes" and gave a correct description of the meaning of the idiom. For each participant, we included only idioms coded as known into subsequent analyses. On average, participants knew 78 of the 90 idioms (86.7%, range = 64–90).

4.2.7 Speech preprocessing

The speech envelope is the acoustic power of the speech signal at a given time in a given frequency range. Here, we estimated the broadband speech envelope by averaging across all ranges, following the procedure described in Chandrasekaran et al. (2009) and adopted by subsequent studies (Gross et al., 2013; Kaufeld et al., 2020; Kayser et al., 2015; Keitel et al., 2017). Using the Chimera toolbox (Smith et al., 2002), we band-pass filtered the auditory signal into 8 frequency bands between 100–8000 Hz (third-order Butterworth filter, forward and reverse), such that the bands spanned equal widths on the cochlear frequency map (1.i in Figure 4.3). The cutoff frequencies of the bands (in Hz) were: 100, 228, 429, 743, 1233, 2000, 3198, 5071, and 8000. We computed the Hilbert transform of the signal in each of these frequency bands and took the absolute value as an estimate of the narrowband envelope (1.ii in Figure 4.3). We downsampled each narrowband speech envelope to 150 Hz, and averaged across all 8 bands to derive the broadband speech envelope (1.iii in Figure 4.3).

4.2.8 EEG recording and preprocessing

The EEG was recorded using an MPI custom actiCAP 64-electrode montage (Brain Products, Munich, Germany), of which 59 electrodes were mounted in the electrode cap (see Supplementary Information S4.2 for electrode layout). Eye blinks were registered by one electrode below the left eye, and eye movements were registered by two electrodes, placed on the outer canthi of both eyes. One electrode was placed on the right mastoid, the reference electrode was placed on the left mastoid and the ground was placed on the forehead.

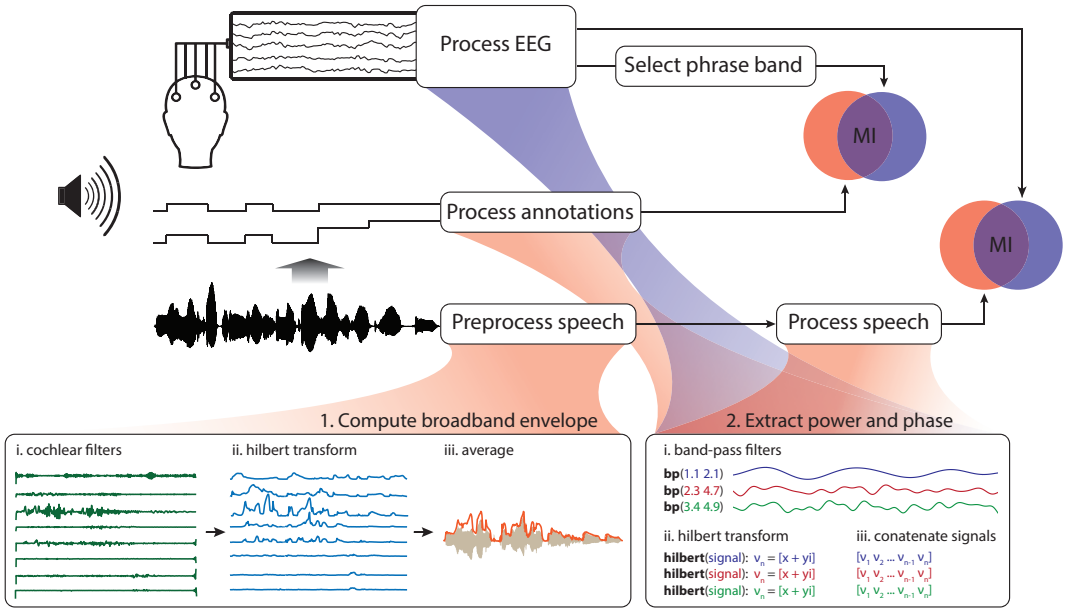


Figure 4.3: Visual representation of the analysis pipeline.

The EEG signal was amplified through BrainAmp DC amplifiers and referenced online to the left mastoid. The data were acquired at a sampling rate of 500 Hz, using a band-pass filter of 0.016–249 Hz.

Preprocessing was performed using the Fieldtrip toolbox (Oostenveld et al., 2011) in MATLAB (Version 2016a). If channels were broken or showed heavy drifts, they were replaced by a weighted average of their neighbors. The data were then low-pass filtered at 50 Hz (36 db/oct), re-referenced to the average of all electrodes and segmented into epochs ranging from the onset to the offset of the audio recording. We manually rejected trials that contained (movement) artifacts and trials in which an unknown idiom was presented (on a by-idiom, by-participant basis; based on the post-experiment questionnaire). We used independent component analysis (ICA; using ICA weights from a version of the data which was downsampled to 300 Hz and high-pass filtered at 1 Hz) to filter artifacts resulting from eye movements and steady muscle activity. Last, we automatically rejected epochs in which the difference between the maximum and minimum voltage exceeded 150 μV. In total, we excluded 9.2% of the data (range of averages across conditions = 6.4%–12.1%). Each EEG segment was downsampled to 150 Hz to match the sampling rate of the speech envelopes. The preprocessed data were then subjected to mutual information analysis.

4.2.9 Mutual information analysis

To quantify cortical speech tracking in each frequency band, we computed MI between the band-limited Hilbert representations of the broadband speech envelope and the EEG signal (see Figure 4.3). In our experiment, MI measures the average reduction in uncertainty about the EEG signal given that the speech envelope (or annotation of syntax) is known, and can thus be used as a measure of the relatedness of the two signals (Ince et al., 2017). We followed the procedure described in Kaufeld et al. (2020), which involved the following steps for speech signals and EEG trials separately: first, each signal was band-pass filtered in the frequency bands of interest (2.i in Figure 4.3), using third-order Butterworth filters (forward and reverse). We then extracted the complex components from each filtered signal using a Hilbert transform (2.ii in Figure 4.3), whose real and imaginary parts were normalized separately using the copula normalization method developed by Ince et al. (2017). We derived instantaneous phase and power and concatenated the resulting signals from all trials (2.iii in Figure 4.3). MI was computed for each electrode, participant, and condition separately, in the following way:

$$MI(\text{EEG};\text{Speech}) = H(\text{EEG}) + H(\text{Speech}) - H(\text{EEG},\text{Speech})$$

Here, H(EEG) is the entropy of the (Hilbert representation of the) EEG signal, H(Speech) the entropy of the (Hilbert representation of the) broadband speech envelope, and H(EEG,Speech) their joint entropy. To accommodate speech-brain lag, we computed MI at five different lags, ranging from 60 to 140 ms, in steps of 20 ms. Statistical analysis was done on the average MI across all five lags.

The same steps were taken for the abstract stimuli, except that the band-pass filter was applied in the phrase frequency band only. MI was computed between the Hilbert representations of the abstract stimuli and the EEG signals corresponding to all forward conditions. For clarification, we use the term speech tracking to refer to MI computed between EEG and the speech envelopes, and syntax tracking to refer to MI computed between EEG and the abstract annotations of syntax.

4.2.10 Statistical analysis of MI values

We fitted linear mixed-effects models (Baayen et al., 2008) to the log-transformed, trimmed (2.5% at both tails of the distribution of each condition) MI

values in each frequency band and in a centroparietal cluster of electrodes (electrodes 1, 3, 4, 5, 8, 9, 10, 11, 28, 29, 30, 33, 35, 36, 37, 40, 41, 42, 43, based on Kaufeld et al., 2020; see Supplementary Information S4.2 for electrode layout) using lme4 (Bates et al., 2015) in R (R Core Team, 2021). In each frequency band we ran two separate models for the MI analysis between EEG and speech. The first model compared MI for Sentences to MI for Idioms, Syntactic prose, and Jabberwocky. This model contained the four-level factor Construction as fixed effect, which was treatment-coded with Sentence as the reference level. Participant was added as a random effect, which had a random intercept and Construction as random slope. Because we had backward versions of Sentences and Word lists, we compared Sentences to Word lists in a two-by-two analysis. This involved a second model with Structure (Sentence vs. Word list), Direction (Forward vs. Backward), and their interaction as fixed effects. Structure and Direction were deviation coded (-0.5, 0.5), and participant was added as random effect, with a random intercept and the interaction between Structure and Direction as random slope. This second model evaluates whether the MI difference between Sentences and Word lists in the forward version is different from the same difference in the backward version.

For the MI analyses in which the speech envelopes were replaced by abstract annotations, we ran a model with the five-level factor Construction (Sentence, Idiom, Syntactic prose, Jabberwocky, and Word list) as fixed effect. This model compared MI for Sentences to MI for Idioms, Syntactic prose, Jabberwocky, and Word lists. Construction was again treatment-coded with Sentence as the reference level. Participant was added as a random effect, which had a random intercept and Construction as random slope. In all analyses we evaluated whether adding a fixed effect increased predictive accuracy by comparing a model with that fixed effect to a model without that fixed effect using R's anova() function.

4.2.11 ERP preprocessing and analysis

To evaluate whether the different forward conditions were processed as intended, we compared the event-related potentials (ERPs) elicited by the sentence-final lexical verb in all syntactically structured conditions (Sentence, Idiom, Syntactic prose, and Jabberwocky). Word lists were not included because the lexical verbs in word lists were not sentence-final (see Table 4.1), due to which the ERP windows segmented around these verbs also contained activity evoked by the subsequent word. We were specifically interested in the N400, a negative-going ERP component that peaks between 300 and 500 ms after the onset of each

content word and is sensitive to predictability and semantic congruency (Baggio & Hagoort, 2011; Kutas & Federmeier, 2011).

Because the segments corresponding to the N400 for these sentence-final verbs lasted beyond the offset of the audio recordings, they were not captured in the segments we used for MI analysis. We therefore used a separate preprocessing pipeline for the ERP analysis, in which the data were low-pass filtered at 40 Hz (36 db/oct), re-referenced to the average of the left and right mastoid, and segmented into epochs ranging from –250 to 1500 ms relative to the onset of the sentence-final verb in each audio recording. All other preprocessing steps were identical to those reported in Section 4.2.8. In total, we excluded 4.8% of the data (range of averages across conditions = 4.1%–5.3%). Before statistical analysis, the EEG data were baseline-corrected using a 250 ms baseline window preceding the sentence-final verb.

For the N400 region of interest, we calculated the voltage in the centroposterior electrodes 3, 8, 9, 15, 27, 28, 35, 40, 41, 47 in a 300–500 ms time window after the onset of the sentence-final word, for each trial and each participant (based on Coopmans & Nieuwland, 2020; see Supplementary Information S4.2 for electrode layout). These voltage values were compared via a linear mixed-effects analysis in R. The mixed-effects model contained Construction as fixed effect, which was treatment-coded with Sentence as the reference level to which the conditions Idiom, Syntactic prose, and Jabberwocky were individually compared. We included participant as random effect, which had a random intercept and Construction as random slope. The models with and without Construction were compared with R's anova() function.

4.3 Results

4.3.1 Speech tracking

In the phrase frequency band, we ran two separate mixed-effects models. The first model evaluates whether MI is modulated by the type of Construction that was presented, comparing Sentences to the other syntactically structured conditions (Idioms, Syntactic prose, and Jabberwocky). Model comparison showed that Construction predicted MI ($\chi^2 = 15.30$, $p = .002$; see left panel of Figure 4.4). Specifically, MI was higher for Sentences than for both Jabberwocky and Syntactic prose, but not different from MI for Idioms (see Table 4.2 for the estimates of the fixed effects). The second model evaluated the interaction be-

tween Structure (Sentence vs. Word list) and Direction (Forward vs. Backward). Model comparison revealed that Sentences elicited higher MI than Word lists (χ^2 = 4.19, p = .041; see left panel of Figure 4.5), and that Forward stimuli elicited higher MI than Backward stimuli (χ^2 = 14.37, p < .001). The interaction was not significant (χ^2 = 0.72, p = .40), which means that the difference between Sentences and Word lists was not solely driven by the linguistic differences between their forward versions and thus (at least partially) also reflects differences in acoustics. The estimates of the fixed effects of this second model are presented in Table 4.3.

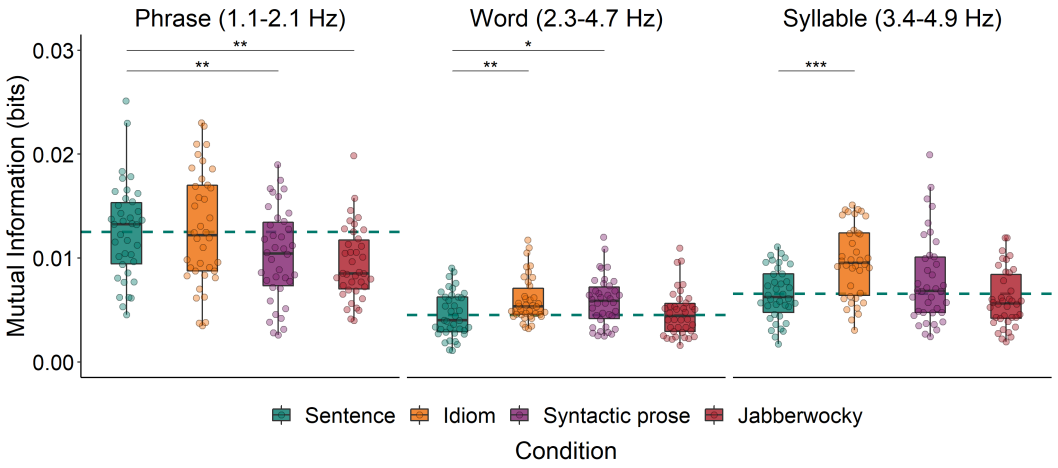


Figure 4.4: Mutual information between EEG and the speech envelopes of all syntactically structured conditions in the phrase, word, and syllable frequency bands. Drops reflect average per participant. The dashed horizontal line reflects the average of the Sentence condition. *p < .05, **p < .01, ***p < .001.

In the word frequency band, the first model showed that Construction predicted MI (χ^2 = 11.71, p = .008; see middle panel of Figure 4.4), but this effect was not driven by the same contrasts as the effect in the phrase frequency band. That is, MI was lower for Sentences than for both Idioms and Syntactic prose, but not different from MI for Jabberwocky (Table 4.2). The second model showed a marginal difference between Sentences and Word lists (χ^2 = 3.70, p = .054; see middle panel of Figure 4.5), and no difference between Forward and Backward stimuli (χ^2 = 0.33, p = .56). The interaction between Structure and Direction was not significant (χ^2 = 2.95, p = .086; Table 4.3).

In the syllable frequency band, the first model again showed that Construction predicted MI (χ^2 = 22.15, p < .001; see right panel of Figure 4.4). MI was lower for Sentences than for Idioms, but not different from MI for Syntactic prose and Jabberwocky (Table 4.2). The second model showed no difference

Table 4.2: Fixed effects of the models that compare speech tracking (i.e., speech-brain MI) for Sentences to speech tracking for Idioms, Syntactic prose, and Jabberwocky.

	Estimate	SE	df	t-value	p-value
Phrase frequency band					
Intercept	-4.77	0.07	37.5	-70.68	<.001
Sentence-Idiom	-0.05	0.10	38.0	-0.52	.61
Sentence-Syntactic prose	-0.30	0.10	38.0	-2.88	.007
Sentence-Jabberwocky	-0.38	0.11	38.0	-3.36	.002
Word frequency band					
Intercept	-5.82	0.09	37.6	-65.05	<.001
Sentence-Idiom	0.31	0.10	37.9	3.08	.004
Sentence-Syntactic prose	0.26	0.10	38.0	2.61	.013
Sentence-Jabberwocky	0.02	0.13	38.0	0.15	.88
Syllable frequency band					
Intercept	-5.44	0.08	37.5	-69.83	<.001
Sentence-Idiom	0.40	0.09	38.3	4.58	<.001
Sentence-Syntactic prose	0.12	0.10	38.0	1.15	.26
Sentence-Jabberwocky	-0.11	0.08	38.0	-1.39	.17

Note. The estimates are from three different models, corresponding to the phrase, word, and syllable frequency bands. *SE* = standard error; *df* = degrees of freedom.

between Sentences and Word lists ($\chi^2 = 0.98$, p = .32; see right panel of Figure 4.5), nor between Forward and Backward stimuli ($\chi^2 = 0.10$, p = .76). The interaction between Structure and Direction was also not significant ($\chi^2 = 2.33$, p = .13; Table 4.3).

4.3.2 Syntax tracking

We then evaluated whether Construction (i.e., Sentence, Idiom, Syntactic prose, Jabberwocky, and Word list) predicted MI between the EEG signal and the abstract annotations of syntactic structure. When these annotations reflected bracket presence, Construction indeed predicted MI ($\chi^2 = 35.98$, p < .001; Figure 4.6A). MI was higher for Sentences than for both Jabberwocky and Word lists, but not different from MI for Idioms or Syntactic prose (see Table 4.4).

The same pattern of results was found for the analysis of MI between the EEG signal and the annotations of bracket count, which differed across Constructions ($\chi^2 = 34.12$, p < .001; Figure 4.6B). MI was higher for Sentences than for Jabberwocky as well as for Word lists, but not different from MI for Idioms or Syntactic prose (Table 4.4).

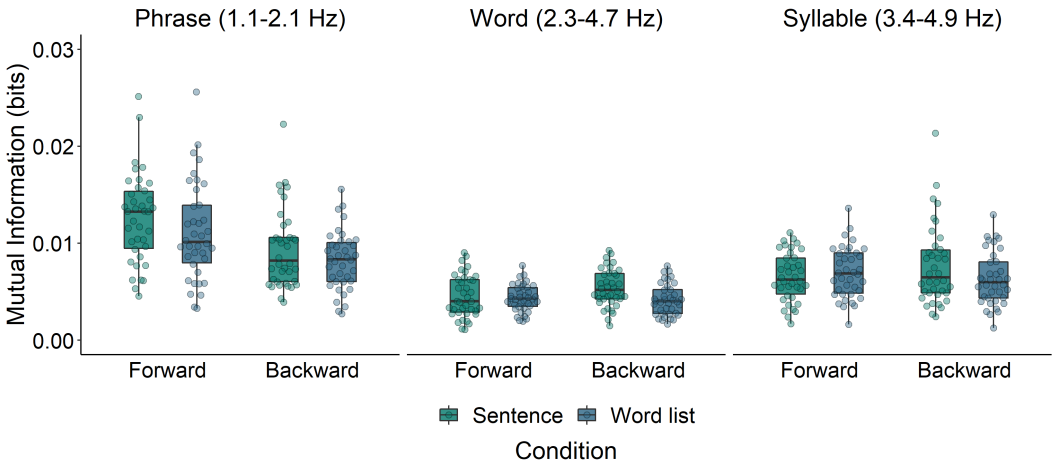


Figure 4.5: Mutual information between EEG and the speech envelopes of both forward and backward versions of Sentences and Word lists in the phrase, word, and syllable frequency bands. Drops reflect average per participant.

Table 4.3: Fixed effects of the interaction models, which evaluate the effects of Structure (Sentence vs. Word list) and Direction (Forward vs. Backward) on speech tracking (i.e., speech-brain MI).

	Estimate	SE	df	t-value	p-value
Phrase frequency band					
Intercept	-5.00	0.04	37.8	-123.55	<.001
Structure	-0.15	0.07	38.1	-2.07	.046
Direction	-0.28	0.07	38.0	-4.10	<.001
Structure*Direction	-0.12	0.14	38.2	-0.85	.40
Word frequency band					
Intercept	-5.77	0.04	37.6	-161.66	<.001
Structure	-0.11	0.06	38.2	-1.82	.077
Direction	0.09	0.08	37.9	1.08	.29
Structure*Direction	0.25	0.14	38.2	1.75	.088
Syllable frequency band					
Intercept	-5.43	0.04	38.0	-135.37	<.001
Structure	-0.13	0.09	37.8	-1.50	.14
Direction	0.03	0.08	37.6	0.45	.66
Structure*Direction	0.23	0.15	38.1	1.55	.13

Note. The estimates are from three different models, corresponding to the phrase, word, and syllable frequency bands. *SE* = standard error; *df* = degrees of freedom.

Overall, both analyses show that, at the frequency band corresponding to abstract phrase structure, the brain tracks the structure of sentences more strongly than the structure of both jabberwocky and word lists. It is interesting to note that the pattern of results is very similar for bracket presence and bracket count,

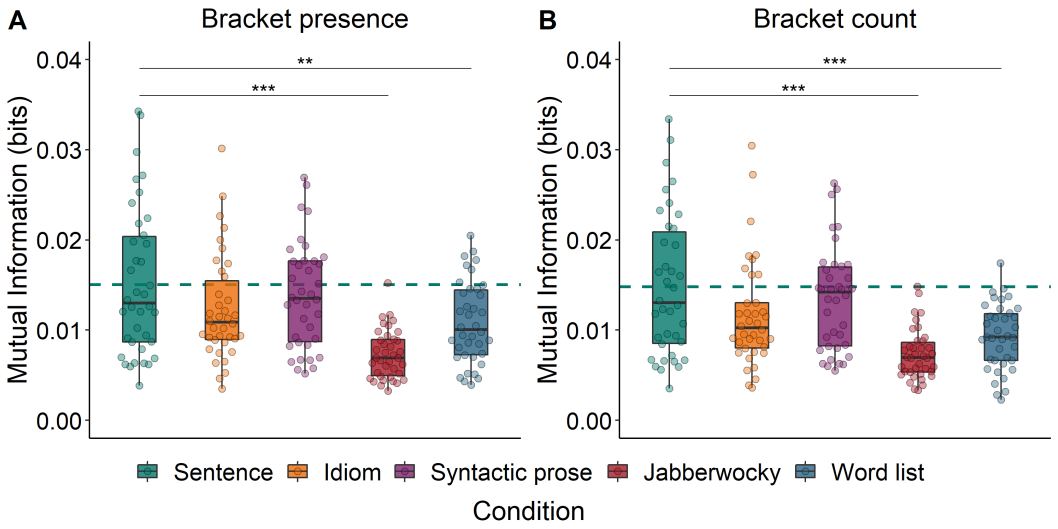


Figure 4.6: Mutual information between EEG and abstract annotations (bracket presence (A) and bracket count (B)) in the phrase frequency band. Drops reflect average per participant. The dashed horizontal line reflects the average of the Sentence condition. **p < .01, ***p < .001.

Table 4.4: Fixed effects of the models that compare syntax tracking (i.e., annotation-brain MI) for Sentences to syntax tracking for Idioms, Syntactic prose, Jabberwocky, and Word lists.

	Estimate	SE	df	t-value	p-value
Bracket presence					
Intercept	-4.62	0.10	37.6	-47.17	<.001
Sentence-Idiom	-0.19	0.13	37.5	-1.51	.14
Sentence-Syntactic prose	-0.12	0.10	38.3	-1.23	.23
Sentence-Jabberwocky	-0.74	0.13	37.8	-5.68	<.001
Sentence-Word list	-0.35	0.12	37.9	-3.02	.005
Bracket count					
Intercept	-4.64	0.10	37.6	-46.85	<.001
Sentence-Idiom	-0.24	0.13	37.6	-1.83	.076
Sentence-Syntactic prose	-0.13	0.11	38.1	-1.23	.23
Sentence-Jabberwocky	-0.71	0.13	37.8	-5.30	<.001
Sentence-Word list	-0.48	0.11	38.1	-4.25	<.001

Note. The estimates are from two different models, corresponding to annotations reflecting bracket presence and bracket count, respectively. *SE* = standard error; *df* = degrees of freedom.

suggesting that the more detailed syntactic information contained in the bracket count annotations does not add predictive accuracy with respect to phrase tracking (contrary to previous work, e.g., Brennan et al., 2016).

An anonymous reviewer rightly noted that the sentences in our conditions differ in co-occurrence frequency, with syntactic prose and jabberwocky sentences having lower transitional probabilities than regular sentences and idioms. However, we do not think this difference can account for our phrase-level effects, because it would predict a pattern of results that is different from what we found. First, it would predict no differences between syntactic prose and jabberwocky sentences, because they have similarly low transitional probabilities. Yet, these two conditions do elicit differences in cortical tracking of phrase structure. To show this, we repeated our linear mixed-effects analysis in the phrase frequency band, but with Syntactic prose as the reference level for the four-level factor Construction. When MI is computed between EEG and the abstract syntactic annotations, it is higher for Syntactic prose than for Jabberwocky, both when the annotations reflect bracket presence (β = -0.57, SE = 0.09, t = -6.32, p < .001; Figure 4.6A) and when they reflect bracket count (β = -0.62, SE = 0.09, t = -6.55, p < .001; Figure 4.6B). There were no differences between Syntactic prose and Jabberwocky in terms of speech tracking (β = -0.08, SE = 0.13, t = -0.59, p = .56). Second, it would predict differences between idioms and regular sentences, because the words in idioms are part of a fixed expression and therefore have high transitional probabilities. However, no such differences between sentences and idioms were found at the phrase level in either speech tracking (see Section 4.3.1) or syntax tracking.

4.3.3 ERPs to sentence-final verb

The MI analyses showed no consistent differences in phrase tracking between Sentences and Idioms, whereas the difference between Sentences and Syntactic prose was inconclusive (i.e., a difference in speech tracking but no difference in syntax tracking). This absence of expected differences might indicate either that the brain does not track the syntactic structure of these stimuli differently (i.e., the conditions are perceived as being different, but this does not affect phrase tracking), or that the conditions were not processed as being very different. To evaluate the latter possibility, we compared the ERPs elicited by the sentence-final lexical verb in all syntactically structured conditions. The results show that the stimuli in the different conditions were processed as expected.

As indicated by the sentence-final ERPs in Figure 4.7A, the variable Construction was associated with modulations of activity in the N400 region of interest (χ^2 = 45.9, p < .001). The ERP elicited by sentence-final verbs in Sentences was less negative than the ERP elicited by sentence-final verbs in both Syntactic

prose (β = -1.56, SE = 0.25, t = -6.37, $p < .001$) and Jabberwocky (β = -0.89, SE = 0.25, t = -3.51, $p < .001$), but more negative than the ERP elicited by sentence-final verbs in Idioms (β = 0.60, SE = 0.27, t = 2.26, p = .029). Note that the effects seem to start quite early, in particular for Syntactic prose (Figure 4.7A). This might have to do with the fact that the words preceding the verb in those stimuli are semantically odd (and thus elicit a strong N400), and with differences between conditions in the pre-verb parts in general. Figure 4.7B contains the topographical plots of the voltage differences in the 300–500 ms time window of interest.

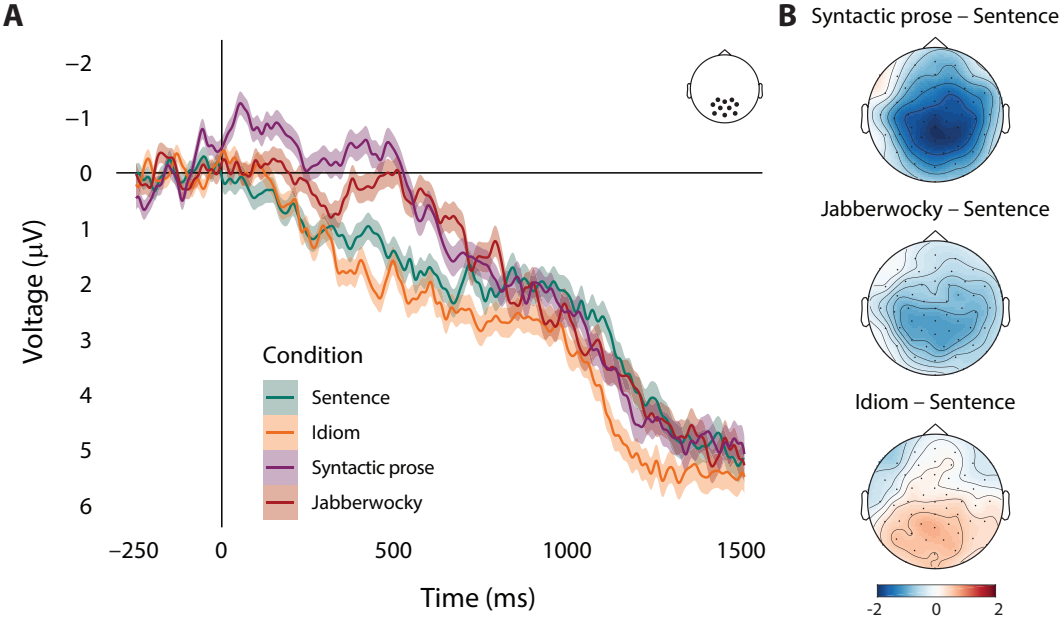


Figure 4.7: (A) Grand-average ERPs at the centroposterior cluster of electrodes, time-locked to the onset of the sentence-final verb in the four syntactically structured conditions. Negative voltage is plotted upwards, and color-shaded areas show the within-subjects standard error of the mean per time sample. (B) Topographical plots of the voltage differences between conditions in the 300–500 ms time window of interest.

4.4 Discussion

In this EEG study with naturally spoken stimuli, we investigated whether cortical tracking of phrase structure is modulated by the degree to which this structure is meaningful. Participants were presented with stimuli that contained different degrees of structural meaning. We measured tracking by computing mutual information between the EEG data and either the speech envelopes (speech

tracking) or abstract annotations of syntax (syntax tracking). Both signals were filtered in the frequency band corresponding to the occurrence of phrases. These analyses showed overall stronger tracking of phrases in regular sentences than in stimuli with reduced lexical-syntactic content (jabberwocky) or without syntactic structure (word lists), but no consistent differences in phrase-level tracking between sentences and divergent stimuli that contained a combination of both structure and lexical meaning (idioms, syntactic prose). As analyses of sentence-final ERPs showed clear differences between the conditions in terms of their sentence-level meaning, we take these findings to suggest that cortical tracking of linguistic structure reflects the internal generation of that structure, whether it transparently maps onto semantic meaning or not.

4.4.1 Effects of composition in processing idioms and syntactic prose

We contrasted regular sentences to two semi-compositional conditions: idioms and syntactic prose. We reasoned that compositional processes would be less engaged during the comprehension of idioms and syntactic prose (Canal et al., 2017; Peterson et al., 2001; Rommers et al., 2013; Vespignani et al., 2010), though there are several factors that likely influence the extent to which participants will try to derive a compositional interpretation from these sentences. Theories of idiom comprehension assume that before the idiom is recognized as being an idiomatic construction, standard literal processing is engaged (see Cacciari & Tabossi, 1988; Libben & Titone, 2008; Sprenger et al., 2006; Titone & Connine, 1999). As the idioms in our experiment were embedded in neutral carrier sentences, they could not be predicted from context, so the idiom recognition point might occur late. Sentence-final ERPs indicated that this was not too late to affect online comprehension: the N400 for sentence-final verbs was reduced in idioms compared to regular sentences, suggesting that the idiom was recognized and retrieved before sentence offset. Nevertheless, any effects of compositionality are likely restricted to late time points, reducing the overall effect of compositionality on cortical tracking on phrase structure.

In addition, there are several differences within the group of idioms that might affect compositional processing. Many of the idiomatic constructions have variable slots that can be filled by compositional information (Jackendoff, 2017). The initial NP of the idiom *NP door de vingers zien* (“to condone” NP), for instance, must be interpreted literally and will likely receive a compositional anal-

ysis. Relatedly, even in idioms that do not take variables, the literal meanings of the words are sometimes part of the idiom's figurative meaning. For example, the idiom *de regels aan je laars lappen*, which literally means "to patch the rules to your boot" (figuratively: "to ignore the rules") is actually about rules (though not about boots), so the NP *de regels* "the rules" will be processed literally. A question for future research is whether this variation within the class of idioms affects cortical tracking. We could not investigate this possibility, as it requires individual-item analyses that were impossible given the way MI was computed. Yet, it would be interesting to see whether cortical tracking of linguistic units is affected by idiomatic variation, such as the transparency, decomposability, and syntactic flexibility of the idiom (see Cacciari, 2014; Cacciari & Glucksberg, 1991; Gibbs et al., 1989; Libben & Titone, 2008).

At the phrase timescale, we did not find a consistent difference between regular sentences and syntactic prose. While speech tracking was stronger for regular sentences than for syntactic prose (left panel of Figure 4.4), this difference was absent in measures of syntax tracking (Figure 4.6). Before we give our interpretation of these results, it is important to emphasize that the computation of speech tracking involves the speech envelope. While we filtered the signals in narrow frequency bands that were based on manual annotations of linguistic information in our stimuli (see Section 4.2.3), and while we checked for acoustic differences via analysis of the modulation spectra (Figures 4.1 and S4.1), we cannot rule out the possibility that any difference between the conditions in terms of speech tracking is driven by acoustic differences between the speech recordings. This possibility is in line with the fact that sentences and syntactic prose did not show differences in terms of syntax tracking, which is based on abstract annotations of syntax without any acoustic differences.

A possible reason for the lack of a consistent effect is that the difference between regular sentences and syntactic prose in terms of compositional processing is not the same at the phrase and sentence levels. At the sentence level, these conditions were clearly differentiable, as indicated by the ERP results. The N400 for sentence-final verbs was larger in syntactic prose than in regular sentences, showing that participants noticed the sentence-level semantic incongruency of syntactic prose. At the phrase-level, however, these conditions might impose similar demands on compositional processing, in particular if the type of composition is mostly syntactic. In contrast to previous research (Brodbeck et al., 2018; Kaufeld et al., 2020), much of the phrase-level compositional processing in our stimuli can be conceived of as syntactic (i.e., containing a combination

of a determiner and a noun) rather than semantic or conceptual. This applies even more strongly to our measure of syntax tracking, which is based on abstract annotations of phrase structure. These annotations do not distinguish between combinations like *the boat* and *red boat*, even though the latter phrase involves more semantic and conceptual composition. It is not unlikely that the presence of semantically impoverished combinations reduces the overall degree of phrase-level tracking (i.e., making syntax tracking a less sensitive measure), especially in light of the finding that manipulations of lexical-semantic and conceptual composition affect measures of brain activity on top of the neural response to syntactic aspects of composition (e.g., Fedorenko et al., 2016; Kaufeld et al., 2020; Schell et al., 2017; Westerlund & Pylkkänen, 2014; Zhang & Pylkkänen, 2015).

4.4.2 Sentence vs. word lists: Structure and acoustics

We observed stronger phrase-level speech tracking for sentences than for word lists, and stronger tracking for forward stimuli than for backward stimuli (left panel of Figure 4.5). The difference between forward and backward speech has been reported before and is often related to differences in their intelligibility (Gross et al., 2013; Kaufeld et al., 2020; Park et al., 2015). However, other studies have failed to find such a relationship (Howard & Poeppel, 2010; Peña & Melloni, 2012; Zoefel & VanRullen, 2016), and it has been suggested that positive correlations between speech tracking and intelligibility are actually driven by the spectro-temporal properties of the unintelligible control condition (for discussion, see Kösem & van Wassenhove, 2017; Zou et al., 2019).

Contrasting with the results reported by Kaufeld et al. (2020), the MI difference between sentences and word lists did not differ across forward and backward versions of these conditions. We therefore cannot exclude the possibility that the difference between (forward) sentences and (forward) word lists reflects acoustic differences. Indeed, analysis of the modulation spectra shows that sentences were reliably different from word lists in terms of spectral properties (Figure 4.1B and Supplementary Information S4.1). However, the presence of acoustic differences does not necessarily negate the effect of syntactic differences. When the speech envelopes were replaced by abstract annotations of syntactic structure (bracket presence or bracket count), we found stronger MI for sentences than for word lists, presumably because these annotations do not reflect any syntactic information in word lists. This leaves open the possibility that syntactic structure did have an effect, but that it could not be detected in measures of speech tracking due to the masking effect of acoustic variance.

In order to find evidence for cortical tracking of phrase structure, we should find not only a difference in the forward condition (as a function of syntax), but also no difference in the backward condition. The latter might be quite difficult to obtain with our measures, because naturally produced sentences and word lists are acoustically quite different, and acoustic properties of the input can account for much of the variance in the neural response to speech (e.g., Brodbeck et al., 2018; Doelling et al., 2014). Moreover, the fact that syntactic information and suprasegmental modulations (e.g., prosodic phrases, intonation phrases) fluctuate at similar frequencies and both affect delta-band activity (Bourguignon et al., 2013; Ghitza, 2017; Meyer et al., 2017; Rimmele et al., 2021) makes it plausible that any structure-driven differences between sentences and word lists were partially masked by their acoustic differences. Supporting this possibility, analysis of spectral power in the phrase frequency band showed a bilateral distribution for all conditions (see Supplementary Information S4.3; see also e.g., Keitel et al., 2017; Molinaro & Lizarazu, 2018). This suggests that the neural signal in this frequency band is also affected by low-frequency modulations in suprasegmental information, which are present in both forward and backward recordings.

4.4.3 Cortical tracking of lexicalized structure

At the timescale of phrases, speech tracking was stronger for sentences than for jabberwocky and syntactic prose, but not different from idioms (left panel of Figure 4.4). In partial agreement with these results, syntax tracking was stronger for sentences than for jabberwocky and word lists, but not different from tracking for idioms and syntactic prose (Figure 4.6). Overall, cortical tracking of phrase structure seems to be enhanced for regular sentences compared to stimuli whose lexical-syntactic content is reduced (jabberwocky, word lists), but it is not consistently different from stimuli that contain both lexical content and syntactic structure (idioms, syntactic prose). This pattern of results is in line with the view that cortical tracking of syntactic structure reflects the generation of structure (Martin, 2020; Martin & Doumas, 2017, 2019; Meyer et al., 2020), whether this structure transparently maps onto a semantic interpretation or not.

Most current neurobiological models of language processing assume structure building to be a lexicalized process (Baggio, 2021; Hagoort, 2005, 2017; Martin, 2020; Matchin & Hickok, 2020). Words are associated with structures that are stored in the mental lexicon in the form of treelets. During language comprehension, these structures are combined to create the hierarchical structures of

phrases and sentences. In this lexical-syntactic conception of structure building, the syntax determines which words can be combined, but the words themselves are the units of combination. As such, this process is affected by its input, in terms of both structure and lexical content. Input-wise, both jabberwocky and word lists are markedly different from sentences. Word lists contain content and function words but lack (cues to) syntactic structure, which means that adjacent words cannot be combined into phrasal units. Jabberwocky sentences are structured sequences that contain both function words and inflectional morphology, but they lack content words and therefore miss the information carried by their argument structure (e.g., the different relations in "*saw* the book *on the table*" and "*put* the book *on the table*"). The lexical-syntactic difference between regular and jabberwocky sentences thus explains why they elicit different degrees of phrase-level speech tracking: structure-building processes are more weakly activated by lexically impoverished input. This idea is supported by evidence showing that lexical information affects neurocognitive measures of structure building (Burroughs et al., 2021; Fedorenko et al., 2016, 2020; Kaufeld et al., 2020; Matchin et al., 2019; Mollica et al., 2020).

Idioms and syntactic prose are similar to sentences in terms of lexical-syntactic structure, but they differ either in the extent to which their interpretation is compositionally derived from this structure (idioms) or in the extent to which a compositional interpretation of this structure makes sense (syntactic prose). Given the absence of consistent differences between these conditions, the interpretation most strongly supported by our data is that phrase-level tracking reflects the lexically-driven computation of syntactic structures in the service of semantic composition (Kaufeld et al., 2020; Martin, 2020; Martin & Doumas, 2017, 2019; Meyer et al., 2020). The computations involved in building hierarchical structure are most strongly activated by syntactically structured sequences of real words. Given the right input, these computations generate a compositional structure, whether the input can easily compose semantically or not.

A similar idea has been proposed for the segmentation of complex word forms into stems and grammatical affixes (Marslen-Wilson, 2007). Behavioral and neuroimaging evidence show that this morphophonological process is triggered by both real words and pseudowords, as long as they contain the diagnostic properties of inflectional affixes in English (Post et al., 2008; Tyler et al., 2005). To explain the lexical insensitivity of this process, Marslen-Wilson (2007, p. 180) argued that "without a decompositional analysis, the system cannot rule out the possibility that the pseudo-regular *trade* is actually the morpheme *tray* in the

past tense, or that *snade* is the past tense of the potential real stem *snay*." Analogously, without a decompositional analysis at the level of phrase structure, the system cannot rule out the possibility that "colorless green ideas sleep furiously" (Chomsky, 1957) actually involves sleeping ideas or that "to kick the bucket" involves buckets being kicked. The conclusion that the compositional meaning of these forms is either semantically incoherent (syntactic prose) or not identical to their intended figurative meaning (idioms) can only be drawn after a compositional analysis has taken place. In a sense, then, the structure-building processes know *how* to build structure (i.e., adhering to syntactic rules, subcategorization restrictions), but not *what* is being built (i.e., whether a compositional analysis yields an interpretation that makes sense). Supporting this functional distinction between generation and interpretation, both behavioral and neurobiological evidence show a difference between sentences and both idioms and syntactic prose in terms of compositional meaning, but not in terms of syntactic structure building (Bastiaansen & Hagoort, 2015; Canal et al., 2017; Konopka & Bock, 2009; Peterson et al., 2001; Rommers et al., 2013; Vespignani et al., 2010).

4.4.4 Effects of composition on word-level speech tracking

At the timescale of words, speech tracking was stronger for both idioms and syntactic prose than for sentences (middle panel of Figure 4.4). These word-level effects might be related to the differential predictability of words in these conditions. It has been shown that speech tracking is enhanced for unpredictable target words that are presented in low-constraining sentence contexts (Donhauser & Baillet, 2020; Molinaro et al., 2021). In these contexts, target words cannot be predicted by top-down mechanisms, so the brain relies more strongly on the bottom-up input to ensure successful comprehension (Donhauser & Baillet, 2020; Molinaro et al., 2021). Words in syntactic prose are semantically odd and thus very unpredictable. On this account, the brain attunes more strongly to unexpected words (in syntactic prose), whose content can only be derived from a bottom-up analysis, than to expected words (in sentences), whose content can be predicted by top-down mechanisms, explaining the difference between syntactic prose and regular sentences in terms of word-level speech tracking.

ERP analysis showed that sentence-final verbs elicited a reduced N400 in idioms compared to regular sentences, indexing facilitated activation or integration of the verb in idioms (e.g., Moreno et al., 2002; Rommers et al., 2013). This indicates that sentence-final verbs were more predictable in idioms than in

sentences, in line with the behavioral literature (Cacciari et al., 2007; Cacciari & Tabossi, 1988). This predictability-related ERP difference is opposite to the difference between sentences and syntactic prose, which suggests that the difference between idioms and sentences in terms of word-level speech tracking (middle panel of Figure 4.4) does not have the same origin as the difference between syntactic prose and sentences. It is unclear why words would be tracked more closely in idioms, but one possibility is that participants were relatively more attuned to words in idioms because words are the linguistic unit on which they have to rely to activate and retrieve the full idiom from memory (for discussion of the activation of properties of the individual words in idioms, see Cacciari, 2014; Cacciari & Tabossi, 1988; Hubers et al., 2021; Sprenger et al., 2006).

4.5 Conclusion

Despite a constantly growing literature on cortical speech tracking, it is still unclear which aspects of high-level linguistic content drive neural activity into alignment with the speech signal. In this EEG study with naturally spoken stimuli, we used an experimental design with a parametric modulation of linguistic information, comparing compositional sentences to stimuli that diverged in terms of their relationship between structure and meaning (idioms, syntactic prose, jabberwocky, word lists). We found that the brain tracks syntactic phrases more closely in regular sentences than in stimuli whose lexical-syntactic content is reduced, but we found no consistent differences in phrase tracking between sentences and stimuli that contained a combination of both syntactic structure and lexical content. These findings refine a recent account of cortical speech tracking, which holds that it indexes the generation of linguistic structure (Martin & Doumas, 2017, 2019; Meyer et al., 2020). Specifically, they suggest that phrase-level speech tracking is modulated by the lexical-syntactic properties of the input to structure building, not by the compositional interpretation of its output. This is in line with neurobiological models of language processing in which structure building is a lexicalized process.

S4 Supplementary Information

S4.1 Analysis of modulation spectra

Figure S4.1A shows the modulation spectra of all forward conditions, computed following the procedure described in Ding, Patel, et al. (2017). For each condition we concatenated all recordings, cut the long sound recording into segments of four seconds long, and then calculated the modulation spectrum of each segment separately. Figure S4.1A shows the modulation spectra after averaging over segments. The Word list condition visibly deviates from the other forward conditions, which are otherwise very similar. To quantify the difference between the modulation spectra of Sentences and the modulation spectra of the other conditions, we computed the area under the curve (AUC) of the modulation spectrum of each segment of each condition and compared the AUCs across conditions.

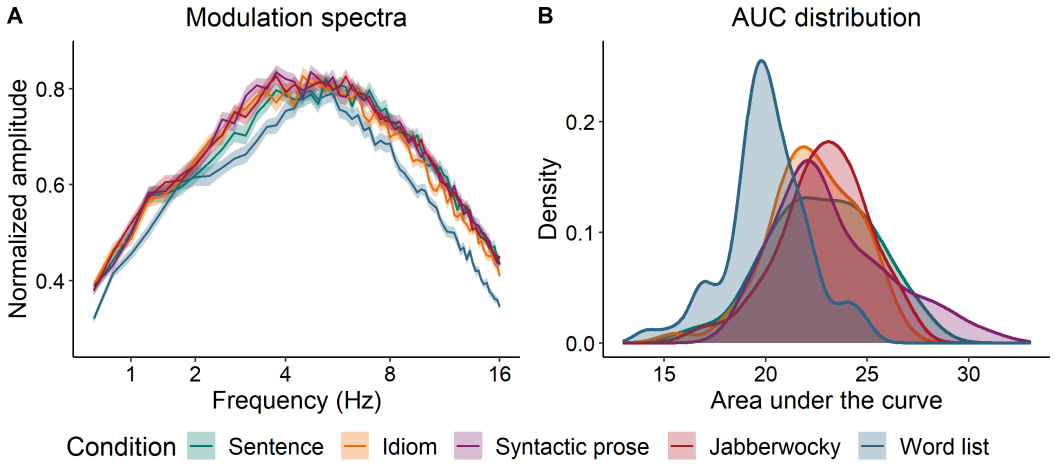


Figure S4.1: (A) Modulation spectra of the forward versions of all conditions. The Word list condition is visibly different from the Sentence condition. (B) Probability density plot representing the distribution of the areas under the curve (AUCs) of the modulation spectra of all forward recordings, which again show the difference between the Word list and the Sentence condition.

The distribution of the resulting AUCs are presented in the probability density plot in Figure S4.1B. Comparison of the five groups (Sentence, Idiom, Syntactic prose, Jabberwocky, Word list) through a one-way ANOVA in R (R Core Team, 2021) indeed reveals that the AUCs of the groups are different, $F(3,267) = 17.2$, $p < .001$. A post-hoc Tukey test shows that the AUCs for Sentences only differ from the AUCs for Word lists, Δ = -2.72, 95% CI [-3.94, -1.49], $p < .001$. They

do not differ from the AUCs for Idioms (Δ = -0.44, 95% CI [-1.71, 0.92], p = .87), Syntactic prose (Δ = 0.49, 95% CI [-0.79, 1.78], p = .83), or Jabberwocky (Δ = 0.067, 95% CI [-1.20 1.34], p = 1.00). Given the difference between the recordings of Sentences and Word lists, we included two conditions to control for these acoustic differences. One control condition contained backward versions of the Sentence recordings (i.e., each recording time-reversed), and the other contained backward versions of the Word list recordings.

S4.2 EEG electrode layout

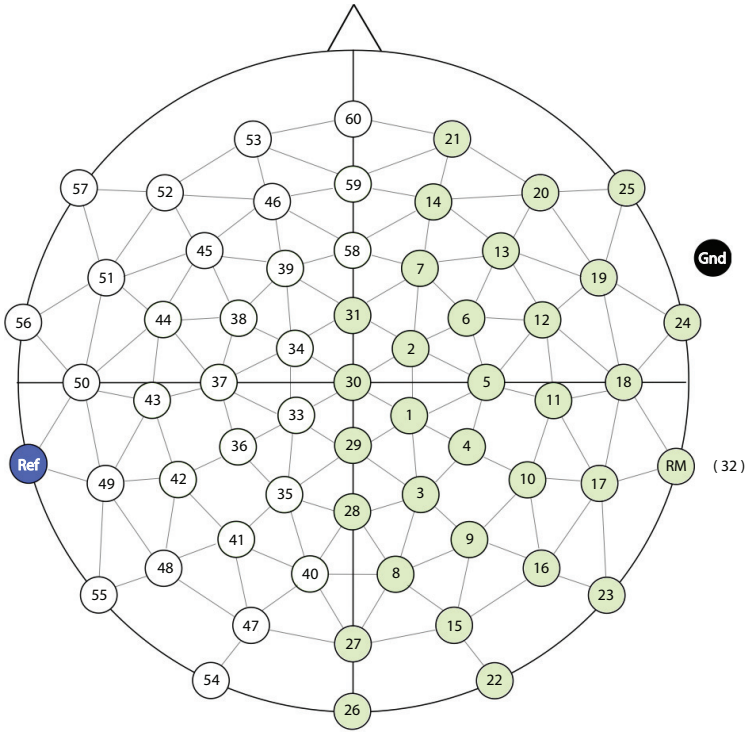


Figure S4.2: Schematic representation of the 59-electrode array layout.

S4.3 Spectral power analysis

We did not find differences in speech tracking between Sentences and Word lists, which might be related to the acoustic differences between their audio recordings (see Figure S4.1). As an exploratory analysis, we examined whether any differences between these conditions could be found in spectral power in the frequency band corresponding to phrases (see e.g., Bonhage et al., 2017; Ding et al., 2016). In contrast to the analysis of speech tracking, this analysis is not based on a comparison of the brain signal and the actual audio signal and might therefore be less affected by acoustic differences between these audio signals.

The topographies in Figure S4.3 reflect the grand average spectral power in the phrase frequency band (1-2 Hz), derived by a fast Fourier transform (Hanning window) of each EEG epoch (all zero-padded to four seconds). Delta power in all conditions, including the two backward conditions, has a bilateral distribution. We compared these effects using cluster-based random permutation tests (Maris & Oostenveld, 2007) in Fieldtrip (Oostenveld et al., 2011). These analyses showed effects of direction for both structures. That is, Forward Sentences elicited stronger delta power than Backward Sentences (one positive cluster, $p = .004$) and Forward Word lists elicited stronger delta power than Backward Word lists (one positive cluster, $p = .002$), but these effects of direction were not different across structures (i.e., no interaction).

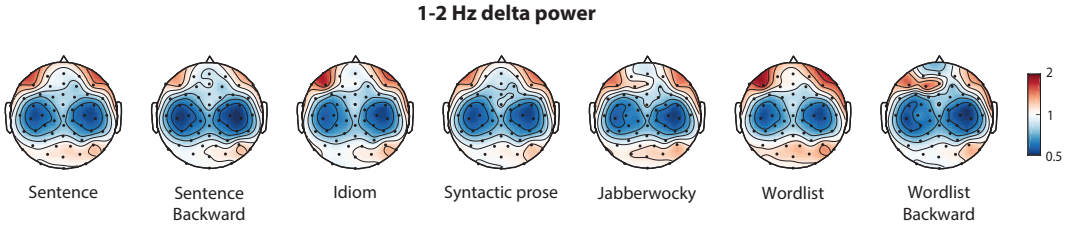


Figure S4.3: Topographical plots of 1-2 Hz delta power in each condition.

5 | Neural source dynamics of hierarchical structure building during natural story listening

Abstract

Neuro-computational language models have gained popularity as linguistically interpretable tools for studying language comprehension in naturalistic contexts. Here, we use this method to investigate to what extent three commonly used parsing strategies can account for neural activity related to Dutch language comprehension. In particular, we test how well the brain activity of people listening to Dutch audiobook stories is predicted by an integratory bottom-up parser, a predictive top-down parser, and a mildly-predictive left-corner parser. Each word in the audiobook was assigned a complexity metric corresponding to the number of nodes that would be visited by the three parsers when incrementally integrating the word into the hierarchical structure of the sentence. Using temporal response functions to map these metrics onto delta-band source activity acquired with magnetoencephalography, we found that the brain data most strongly reflect node counts derived by the top-down method, which consistently engages activity in left frontal and temporal regions. This finding suggests that predictive structure building is an important component of Dutch sentence comprehension. The absence of strong effects of the left-corner model further suggests that its mildly-predictive strategy does not represent Dutch language comprehension well, in contrast to what has been found for English. These findings underscore the need for more work on typologically diverse languages, whose structural properties are different from those of English and therefore invite different parsing strategies within the fronto-temporal language network.

5.1 Introduction

Integrating linguistic theory with cognitive neuroscience requires not only detailed knowledge of both linguistics and neural data, but also a linking hypothesis that specifies how linguistic information is to be connected with the observable neural signal (Embick & Poeppel, 2015; Poeppel, 2012). When it comes to studying syntactic structure building in the human brain, it is therefore important to be explicit about the structure of the syntactic representations, the algorithmic procedures for computing these representations in real time, and the linking theory that maps the output of these algorithms onto neural signals (Brennan, 2016; Demberg & Keller, 2019; Martin, 2016; Sprouse & Hornstein, 2016). One promising approach relies on neuro-computational language models, which are cognitive models of language processing that are computationally precise and have broad coverage (Brennan, 2016; Hale et al., 2022). Because of their broad coverage, they can be used to define a measure of linguistic processing difficulty for every word in a large dataset that reflects everyday language use. By determining whether this measure reliably predicts brain activity elicited by the words in that dataset, we can establish whether there is evidence for the hypothesized linguistic computations in the neural signal. Using neuro-computational language models is an important step towards integrating linguistics and cognitive neuroscience, because it allows for a characterization of neural activity in terms of computationally precise incremental parse states (Brennan et al., 2016; Stanojević et al., 2021).

In the current study, we compare different neuro-computational models in terms of their ability to predict brain activity of people listening to Dutch audiobook stories. We quantify the cognitive states implied by different parsing models that build syntactic structure either in a predictive or in a non-predictive (integratory) manner, and use magnetoencephalography (MEG) to localize responses reflecting syntactic structure building with high temporal precision.

5.1.1 Neuro-computational models of sentence comprehension

Brennan (2016) defines neuro-computational models as consisting of a parser $P_{G,A,O}$ that contains a grammar G with rules to construct representations, an algorithm A for incrementally applying the grammar word by word, and an oracle O that resolves indeterminacies. When applied to a sequence of words $w_1, w_2, ...w_n$, $P_{G,A,O}$ yields a sequence of mental states $m_1, m_2, ...m_n$, which corre-

spond to (partial) syntactic structures. These mental states can be quantified via an auxiliary hypothesis or linking rule, often referred to as complexity metric in psycholinguistics because of the way in which it quantifies language processing complexity (Hale et al., 2018). The complexity metric *C* thus represents the neural state in quantifiable cognitive terms; it stands for estimated brain states. The estimated brain states are linked to the observable neural signal via a response function *R*.

This neuro-computational method has the potential to evaluate the cognitive and neural relevance of linguistic constructs to the extent that there are appropriate linking hypotheses about how these constructs are manifested in neural activity. One assumption underlying this approach is that a model that captures linguistic competence should give rise to measures that are more predictive of experimental data (performance). It is promising that the results of recent neuroimaging studies point in this direction. For instance, grammars that compute hierarchical structure account for variance in brain activity above and beyond the variance accounted for by sequence-based models (Brennan et al., 2012, 2020; Brennan & Hale, 2019; Lopopolo et al., 2021; Martin & Doumas, 2017; Shain et al., 2020), and hierarchical grammars that naturally represent long-distance dependencies, which are ubiquitous in natural languages, uniquely predict activity in brain areas commonly linked to syntactic structure building (Brennan et al., 2016; Li & Hale, 2019; Nelson et al., 2017; Stanojević et al., 2021). These findings reinforce the view that grammars that are well-equipped to account for natural language structures (competence) are also required to adequately model the activity of the brain when it incrementally computes these structures (performance). In the following sections, we will discuss how neuro-computational language models are utilized in the current work. In particular, we discuss how varying the parsing algorithm used to build syntactic structure allows us to address a question about the universality of language processing strategies.

Three approaches to syntactic structure building

Parsing models specify how syntactic parse states unfold incrementally during language comprehension. In the current work, these parse states correspond to partial syntactic structures, which are generated via the rules of X-bar theory (the grammar *G*; Carnie, 2021; Jackendoff, 1977). We chose to construct X-bar tree structures because they are appropriately expressive to deal with natural language structures (e.g., long-distance dependencies, movement) and because recent work using neuro-computational models has shown that complexity met-

rics derived from X-bar structures predict brain activity above and beyond activity accounted for by both sequential models and models that are hierarchical but less expressive (Brennan et al., 2016; Li & Hale, 2019; Nelson et al., 2017). We further assume a perfect oracle O, which resolves temporary ambiguities in the right way (Bhattasali et al., 2019; Brennan et al., 2012, 2016; Brennan & Pylkkänen, 2017). This means that the parser builds the correct structure at any point in the parse, even when faced with locally ambiguous input.

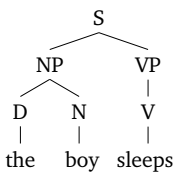
Any given syntactic structure may be built in different ways, depending on the parsing algorithm A that is adopted. We will be concerned with three algorithms for building structure: a top-down algorithm, a bottom-up algorithm and a left-corner algorithm (see Hale, 2014).[1] The top-down parsing method works via *expansion* of rewrite rules. Before each word, all rules necessary to attach the upcoming word to the structure are expanded (e.g., in VP → V NP, the VP is expanded as V and NP). As these rules are applied based on the left-hand side of the rule and in advance of each word, this method builds constituent structure entirely predictively. The bottom-up parsing method, instead, builds constituent structure in a non-predictive manner, as it postulates a constituent node only after all of its daughter nodes are available. This process is referred to as *reduction*: when all information on the right-hand side of a rewrite rule is available, the input is reduced to the constituent node (e.g., in VP → V NP, the input V and NP is reduced to VP). In between these two parsing methods is a mildly-predictive left-corner strategy, which works via *projection*. A constituent node is projected after the very first symbol on the right-hand side of the rewrite rule (its left corner) is seen (e.g., the VP node is projected when V is available). This strategy is only mildly predictive because, while it requires input to build structure (in contrast to the top-down strategy), the input can be incomplete (in contrast the bottom-up strategy).

To illustrate the difference between these strategies, consider the simplified structure corresponding to the sentence "the boy sleeps" in Table 5.1. The total number of operations (expand, reduce, project) for the three parsing methods is the same, but the timepoints at which they are applied differ. On the predictive top-down parsing strategy, three operations are applied before "the", corresponding to expansion of the S, NP, and D nodes. Only one operation is

[1] Note that the notions top and bottom in this context refer to the geometry of the syntactic tree structure. Top-down thus says something about the (vertical) direction of phrase structure building, and with 'top-down effects' we refer to the effects of node count derived by a top-down parser. In the predictive processing literature, the term 'top-down effects' is commonly used to describe how low-level processes are affected by high-level sources of information, but this is not our intended interpretation.

applied at "the" on the bottom-up method, because there is complete evidence for the determiner node D only. The next word "boy" is the second word of the noun phrase, so two operations are applied bottom-up, but only one is applied top-down. What this simple structure illustrates is that these parsing methods differ in the dynamics of structure building. They make different predictions as to when in the sentence processing complexity is high, and these differences will only be magnified when the sentences become longer and more complex.

Table 5.1: Parser actions for top-down, bottom-up and left-corner parsers for the sentence "the boy sleeps".



Node count			
TD	3	1	2
BU	1	2	3
LC	2	2	2

Top-down		**Bottom-up**		**Left-corner**	
`expand by`	S → VP NP	`shift`	the	`shift`	the
`expand by`	NP → D N	`reduce by`	D → the	`project`	D → the
`expand by`	D → the	`shift`	boy	`project`	NP → D N
`scan`	the	`reduce by`	N → boy	`shift`	boy
`expand by`	N → boy	`reduce by`	NP → D N	`project`	N → boy
`scan`	boy	`shift`	sleeps	`project`	S → NP VP
`expand by`	VP → V	`reduce by`	V → sleeps	`shift`	sleeps
`expand by`	V → sleeps	`reduce by`	VP → V	`project`	V → sleeps
`scan`	sleeps	`reduce by`	S → NP VP	`project`	VP → V

Note. The number of parser actions required at each word yields an incremental measure of processing complexity, here termed node count. The `scan`/`shift` action corresponds to processing or moving to the next word in the sentence and is not a structural operation. The other parser actions are explained in the text.

Here, we represent the number of parser actions at each word in the form of incremental node count, which corresponds to the number of new nodes in a partial syntactic structure that are visited by the parser when incrementally integrating a word into the structure (complexity metric *C*; Brennan et al., 2012; Frazier, 1985; Miller & Chomsky, 1963). Depending on the parsing algorithm that is used, node count reflects the number of `expand` (top-down), `reduce` (bottom-up) or `project` (left-corner) actions between successive words (Table 5.1), all of which can be taken as roughly corresponding to the syntactic load or complexity of those words. Previous neuroimaging work has shown that node count effectively quantifies syntactic complexity in cognitive terms (Bhattasali et al., 2019; Brennan & Pylkkänen, 2017; Brennan et al., 2012, 2016; Coopmans, de Hoop, Hagoort, & Martin, 2022; Giglio, Ostarek, Sharoh, & Hagoort, 2022; Li & Hale, 2019; Lopopolo et al., 2021).

The left-corner parsing strategy is thought to be cognitively plausible as a model of human language processing. It correctly predicts processing difficulty for center-embedded constructions (Abney & Johnson, 1991; Johnson-Laird,

1983; Resnik, 1992), is compatible with a range of findings from the sentence processing literature (Hale, 2014), and accounts for brain activity during language processing (Brennan & Pylkkänen, 2017; Nelson et al., 2017). However, much of the relevant psycholinguistic work has been done in English. A possible reason for the observation that left-corner parsing works well for English is that English phrases are strictly head-initial. The left corner of a phrase will therefore most often be its head, whose argument structure allows phrase structure to be built predictively (Arai & Keller, 2013; Boland & Blodgett, 2006; Schütze & Gibson, 1999). Dutch, in contrast, exhibits mixed headedness, with the verbal projections VP and IP being head-final. The left corner of a Dutch head-final VP will often be a multi-word constituent, such as in sentence (1) below:

(1) De student heeft een essay over syntaxis geschreven.
the student has an essay on syntax written
"The student has written an essay on syntax."

Notice the difference in word order between the Dutch example and its English translation. In English, the non-finite verb precedes its complement ("written – an essay on syntax"), thus yielding the head-initial order that is characteristic of English phrases. The reverse order is found in Dutch, where the non-finite verb follows its complement (*een essay over syntax – geschreven*, "an essay on syntax – written"), giving the head-final order. The left-corner method predicts that the VP constituent in Dutch will be projected only after the entire preverbal NP complement has been processed. This might be unrealistically late, in particular if speakers of head-final languages adopt predictive parsing strategies (Coopmans & Schoenmakers, 2020; Vasishth et al., 2010). It might thus very well be that left-corner parsing is not the best strategy for Dutch structures, and that brain activity linked to parsing by English listeners will differ from that found for speakers of Dutch.

From cognitive to brain states via temporal response functions

The syntactic complexity metrics derived from different parsing models can be mapped onto electrophysiological brain activity through temporal response functions (TRFs). TRFs are linear kernels that describe how the brain responds to a representation of a (linguistic) feature (Brodbeck et al., 2018, 2022). This approach is similar to that of recent neuro-computational fMRI studies, which

use the canonical hemodynamic response function to fit syntactic predictors onto brain activity in a given region of interest (Bhattasali et al., 2019; Brennan et al., 2012, 2016; Giglio, Ostarek, Sharoh, & Hagoort, 2022; Li & Hale, 2019). But rather than assuming the shape of the response function, with the TRF method a response function can be estimated for each predictor separately. This allows us to determine not only whether a brain region is sensitive to the information encoded in a predictor, but also when that information is processed by the brain. Moreover, by using TRFs, we can explicitly model acoustic properties of the auditory stimulus. This is important because high-level linguistic features can be correlated with low-level stimulus properties, such that neural effects attributed to linguistic processing can also be explained as the brain's response to non-linguistic, acoustic information (Daube et al., 2019). And because the response to acoustic properties is orders of magnitude larger than that to linguistic features (Brodbeck et al., 2022; Gillis et al., 2021), acoustic variance could mask the subtle effects of linguistic information that are hiding in the data (Coopmans, de Hoop, Hagoort, & Martin, 2022). In sum, the benefit of the TRF method for the current study is that it allows us to evaluate the brain's sensitivity to syntactic information in natural speech with high temporal precision, while appropriately controlling for lower-level factors.

5.1.2 The present study

In the current study, we compare different neuro-computational models in terms of their ability to predict MEG activity of people listening to Dutch audiobook stories. The three models we evaluate rely on the same grammatical assumptions (X-bar theory), linking hypothesis (node count), and type of response function (TRF), but they differ in the parsing algorithm by which they build syntactic structure. Left-corner parsing has been argued to be suitable for English sentence processing, but it is an open question whether the strategy adopted by the left-corner method also accounts for people's brain activity during the incremental comprehension of Dutch. Dutch is typologically related to English, but the head-finality of its verb phrases might invite different processing strategies (Bornkessel-Schlesewsky & Schlesewsky, 2016; Vasishth et al., 2010). Based on recent results linking syntactic processing to delta-band activity (Bai et al., 2022; Brennan & Martin, 2020; Coopmans, de Hoop, Hagoort, & Martin, 2022; Ding et al., 2016; Kaufeld et al., 2020; Martin, 2020), and in line with the idea that the timescale of syntactic processing overlaps with the delta frequency range (Henke & Meyer, 2021; Meyer et al., 2020), we focus on MEG activity in the delta band.

Previous work has identified several brain areas in the left hemisphere that are responsive to complexity metrics derived from incremental parse steps, including the inferior frontal lobe (Bhattasali et al., 2019; Giglio, Ostarek, Sharoh, & Hagoort, 2022; Nelson et al., 2017), the anterior temporal lobe (Bhattasali et al., 2019; Brennan & Pylkkänen, 2017; Brennan et al., 2012, 2016; Nelson et al., 2017; Stanojević et al., 2021), and posterior superior and middle temporal areas (Brennan et al., 2016; Giglio, Ostarek, Sharoh, & Hagoort, 2022; Li & Hale, 2019; Nelson et al., 2017; Lopopolo et al., 2021). On the assumptions that these effects reflect the operations involved in syntactic structure building and that people employ similar parsing strategies when processing English and Dutch sentences, we expect effects to show up in the same brain regions. Because the majority of this work has used fMRI, the timing of the expected effects is less clear, though the results of a few electrophysiological studies suggest that effects of structure building are reflected in brain activity within the first 500 ms after word onset (Brennan & Pylkkänen, 2017; Hale et al., 2018). In sum, the predictive accuracy of the different parsing models can give us insight into the (potentially language-specific) processing strategies Dutch parsers use (i.e., *how* they build structure), and the spatial-temporal properties of the effects provide clues about *when* and *where* in the brain these processes are implemented.

5.2 Methods

5.2.1 Participants

24 right-handed native speakers of Dutch (18 female, mean age = 33.4 years, age range = 20–58 years) were recruited via the SONA system of Radboud University Nijmegen. They all reported normal hearing, had normal or corrected-to-normal vision, and did not have a history of language-related impairments. Participants gave written informed consent to take part in the experiment, which was approved by the Ethics Committee for human research Arnhem/Nijmegen (project number CMO2014/288). This project is part of a larger data collection project, for which the pre-registered sample size is 32 ($d = 0.5$, $\alpha = 0.05$, $\beta = 0.2$). At the start of the analyses reported in this chapter, data collection was finished for 24 participants. Previous analyses of the same dataset have shown that this sample is large enough to reconstruct the neural response to linguistic features using TRF analysis (Tezcan et al., 2022).

5.2.2 Stimuli

The stimuli consisted of stories from three fairy tales in Dutch: one story by Hans Christian Andersen and two by the Brothers Grimm. All stories contain a rich variety of naturally occurring sentence structures, varying in syntactic complexity. In total, there are 8551 words in 791 sentences, which are on average 10.8 words long (range = 1-35, SD = 5.95). They were auditorily presented in nine separate segments, all of which were between 289 (4 min, 49 sec) and 400 seconds long (6 min, 40 sec; see Supplementary Information S5.1), for a total of 49 minutes and 17 seconds. The loudness of each audio segment was normalized using the FFmpeg software (EBU R128 standard). The transcripts of each story were automatically aligned with their corresponding audio recording using the WebMAUS segmentation software (Kisler et al., 2017).

5.2.3 Syntactic annotations

We manually created syntactic structures for all sentences in the audiobooks following an adapted version of X-bar theory (Carnie, 2021; Jackendoff, 1977). To be specific, we consistently created an X-bar structure for NPs and VPs, whereas intermediate projections for all other phrases were drawn only if they were needed to attach adjacent words to the structure (e.g., APs were unbranched unless they were modified by an adverb or prepositional phrase). The X-bar template for NPs and VPs was strictly enforced in order to make a structural distinction between arguments and adjuncts; arguments were attached as sister of the head, while adjuncts were attached to an intermediate projection (Jackendoff, 1977). All phrases except for VPs and IPs are head-initial. An example of a hierarchical tree structure is given in Figure 5.1.

Node counts were computed for each word in every sentence in three different ways. On the bottom-up strategy, a constituent node is posited when all daughter nodes have been encountered. This amounts to counting the number of closing brackets directly following a given word in a bracket notation.[2] On the top-down strategy, a node is posited right before it is needed to attach the upcoming word to the structure. This amounts to counting the number of opening brackets directly preceding a given word in a bracket notation. And on the left-corner strategy, a node is posited when its left-most daughter node has been encountered. Terminal nodes were not included in the node count calculation.

[2]Note that bottom-up node count reflects exactly the same as the bracket count measure used in Chapter 4.

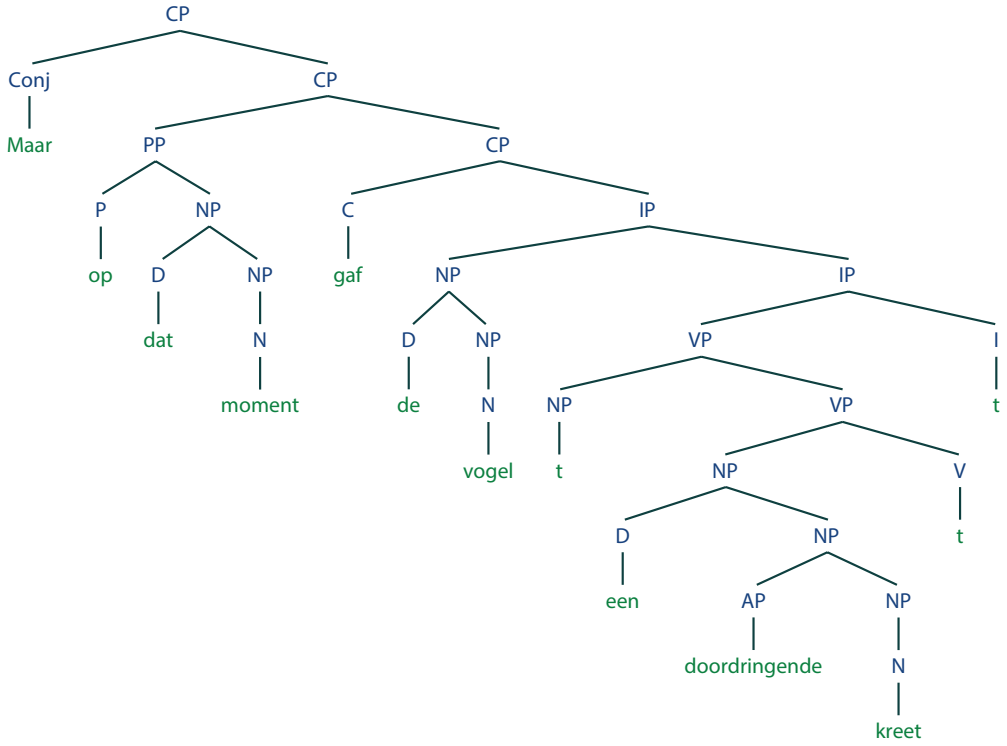


Figure 5.1: Syntactic structure of an example sentence from one of the audiobook stories. The *t* stands for trace and refers to the position at which the word or phrase to which it is related is interpreted. It does not have an acoustic correlate in the speech signal.

As can be seen in Figure 5.1, the X-bar structures contain traces of movement. As these empty elements do not have an acoustic correlate in the speech signal, we reassigned their node counts to another word in the same sentence. Our reasoning was that the precise location of these elements could not be predicted with certainty, though it could be inferred after their putative position. We therefore decided to add the node count of each empty element to the node count of the word following it. By doing so, we aimed to capture the syntactic processes associated with these elements (e.g., long-distance dependency resolution) around the times they occur. We will come back to this point in the discussion.

The resulting node count values were time-aligned with the onsets of the words in the audio recordings (see Figure 5.2 for an example). Each audio file could thus be represented as a vector with a node count value at the onset of each word and zeros everywhere else. The vectors for each of the three parsing strategies were the predictors in the TRF analysis (see Section 5.2.7).

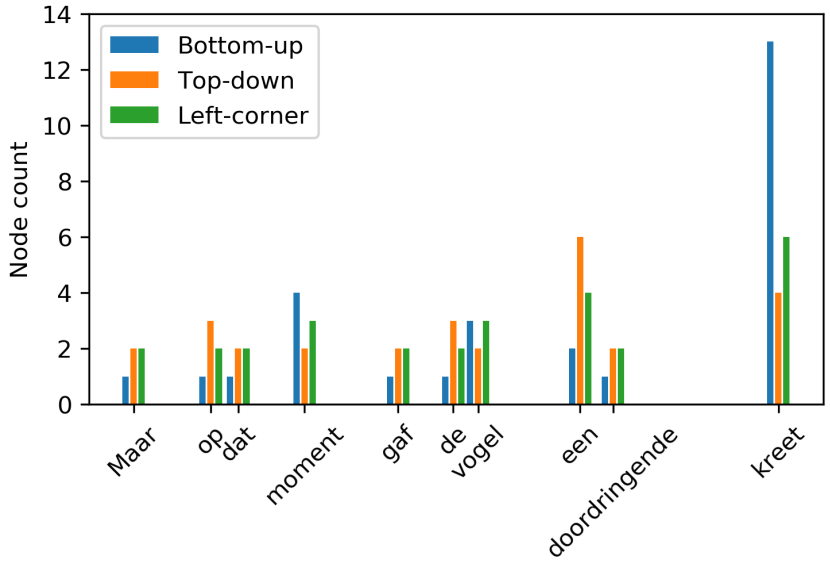


Figure 5.2: Node counts for the structure presented in Figure 5.1, time-aligned with the onsets of the words in the corresponding audio recording.

5.2.4 Procedure and data acquisition

Participants were individually tested in a magnetically shielded room. They were instructed to attentively listen to the nine audiobook stories while sitting still and looking at a fixation cross that was presented in the middle of the screen. After each of the nine story blocks, five multiple-choice comprehension questions (each with four options) were asked. On average, participants answered 88.1% of the questions correctly (SD = 7.52%), showing that they paid attention to and understood the content of the stories.

The MEG data were recorded with a 275-channel axial gradiometer CTF system at a sampling rate of 1200 Hz. The audio recordings were presented using Psychtoolbox in MATLAB (Brainard, 1997) via earphones inserted into the ear canal. Participants' eye movements and heartbeat were recorded with EOG and ECG electrodes, respectively. Throughout the recording session, their head position was monitored using three head localization coils, one placed in fitted earmolds in each ear and one placed at the nasion (Stolk et al., 2013). Each block started with a 10-second period during which resting state data were recorded. In the break between story blocks, participants were instructed to reposition their head location in order to correct for head movements. After the MEG session, each participant's head shape was digitized using a Polhemus 3D tracking device, and their T1-weighted anatomical MRI was acquired using a 3T Skyra system (Siemens).

5.2.5 MEG preprocessing

Preprocessing was done using MNE-Python (version 0.23.1). The MEG data were first down-sampled to 600 Hz and band-pass filtered at 0.5-40 Hz using a zero-phase FIR filter (MNE-Python default settings). We then interpolated channels that were considered bad using Maxwell filtering, and used Independent Component Analysis to filter artifacts resulting from eye movements (EOG) and heartbeats (ECG). We segmented the data into nine large epochs, whose onsets and offsets corresponded to those of the audio recordings. Source reconstruction was done for each epoch separately.

5.2.6 Source reconstruction

Individual head models were created for each participant with their structural MRI images using FreeSurfer (Dale et al., 1999). The MRI data were then co-registered to the MEG with MNE co-registration, using the head localization coils and the digitized head shape. We set up a bilateral surface-based source space for each individual participant using fourfold icosahedral subdivision, resulting in 2562 continuous source estimates in each hemisphere. The forward solution was computed using a BEM model with single layer conductivity. We low-pass filtered the signal at 4 Hz using a zero-phase FIR filter (corresponding to the 0.5-4 Hz delta band) and estimated sources using the dSPM method (noise-normalized minimum norm estimate), with source dipoles oriented perpendicularly to the cortical surface. The noise covariance matrix was calculated based on the resting state data that were recorded before each story (all concatenated). Before TRF analysis, each source estimate was downsampled to 100 Hz to speed up further computations.

5.2.7 Predictor variables

To control for brain responses to acoustic information, all models included two acoustic predictors: an eight-band gammatone spectrogram (i.e., envelope of the acoustic signal in different frequency bands) and an eight-band acoustic onset spectrogram. Both spectrograms covered frequencies from 20 to 5000 Hz in equivalent rectangular bandwidth space (Heeris, 2018), and were resampled to 100 Hz to match the sampling rate of the MEG data. The onset spectrogram was derived from the gammatone spectrogram using an auditory edge detection model (Brodbeck et al., 2020).

All models also included four word-based predictors that are all strongly linked to brain activity during naturalistic language processing (Brodbeck et al., 2018, 2022; Weissbart et al., 2020). These were word rate and the three statistical predictors word frequency, surprisal, and entropy. All predictors were modeled on the gammatone predictors in terms of length and sampling rate.

The word rate predictor is simply a one-dimensional array with the value 1 at word onsets and the value 0 everywhere else.

The frequency of word w was computed by the taking negative logarithm of the number of occurrences of w of per million words, extracted from the SUBTLEX-NL database of Dutch word frequencies (Keuleers et al., 2010):

$$Word\ frequency(w) = -log_2(frequency(w))$$

We chose to represent word frequency via the negative logarithm, because in this way infrequent words will get high values and frequent words will get low values, in line with the brain response to word frequency (i.e., a larger response to infrequent words; Brennan et al., 2016; Brodbeck et al., 2018). For some words we could not compute a frequency value because the word did not appear in the database. Manually checking them revealed that these were uncommon (and thus likely infrequent) words, so we assigned to them the value corresponding to the lowest frequency of all words present in the audiobook.

Surprisal is the conditional probability of a word given the preceding linguistic context, quantified as the negative logarithm of this probability (Hale, 2001, 2014). Thus, the surprisal of word w at position t is calculated via:

$$I(w_t) = -log_2(P(w_t|context))$$

Word surprisal was computed from conditional probabilities obtained with GPT-2 for Dutch (de Vries & Nissim, 2021). GPT-2 used the preceding 30 words as context, so *context* in the formula above refers to $(w_{t-30} \ldots w_{t-1})$.

Entropy at word position t is the uncertainty before observing the next word w_{t+1} given the preceding context. Context was again defined as the previous 30 words (including w_t), and conditional probabilities were again obtained with GPT-2 for Dutch (de Vries & Nissim, 2021). Entropy at word position t was then calculated as the sum of the conditional probabilities of each next word (within the set of possible upcoming words W), weighted by the negative logarithm of this probability:

$$H(t) = -\sum_{w_{t+1} \in W} P(w_{t+1}|context)\ log_2(P(w_{t+1}|context))$$

Our syntactic models included the syntactic predictors bottom-up node count, top-down node count and left-corner node count. We constructed a total of four models. Table 5.2 specifies which predictors are included in which model.

Table 5.2: Predictors included in each model.

Predictor(s) / Model name	Spectro-gram/ Onsets	Word onset	Word frequency/ Entropy/ Surprisal	Bottom-up	Top-down	Left-corner
Bottom-up + Top-down	X	X	X	X	X	
Bottom-up + Left-corner	X	X	X	X		X
Top-down + Left-corner	X	X	X		X	X
Full	X	X	X	X	X	X

5.2.8 Model estimation

TRFs were estimated for each subject and MEG source point separately using Eelbrain (Version 0.37.3; Brodbeck et al., 2021). The MEG response at time t, denoted as $\hat{y}(t)$, was predicted jointly by convolving each TRF with a predictor time series shifted by K time delays (Brodbeck et al., 2021, 2022):

$$\hat{y}(t) = \sum_{f=1}^{F} \sum_{k=1}^{K} \beta_f(\tau_k) x_f(t - \tau_k)$$

Here, x_f is the predictor time series and $\beta_f(\tau_k)$ is the TRF of the corresponding predictor at delay τ_k. The coefficient of the TRF at delay τ thus indicates how a change in the predictor affects the predicted MEG response τ milliseconds later. To generate each TRF, we used 50-ms wide Hamming windows and shifted the predictor time series between -100 and 1000 ms at a sampling rate of 100 Hz, thus yielding $K = 110$ different delays. The length of the TRF was chosen based on the latency of syntactic effects in naturalistic paradigms (Brennan & Hale, 2019; Hale et al., 2018). Before estimating the TRF, all predictors were mean-centered and then normalized by dividing by the mean absolute value.

TRFs were estimated using a five-fold cross-validation procedure. We first concatenated the data for each subject along the time axis, and then split them up into five equally long segments. During each cross-validation run, three segments were used for training, one for validation, and one for testing. For each test segment, there were four training runs with each of the remaining segments serving as the validation segment once. Using a boosting algorithm (David et al., 2007) to minimize the $\ell 1$ error, one TRF was estimated for each of the four training runs (selective stopping based on $\ell 1$ error increase). The resulting four TRFs were averaged to predict responses in the test segment. This analysis yields an average TRF for each predictor in each model, as well as a measure of reconstruction accuracy for the whole model. Reconstruction (or predictive) accuracy refers to the fit between the predicted and the observed MEG signal at each source point, quantified in terms of explained variance in R^2. Reconstruction accuracy can be seen as a measure of cortical tracking: the larger the reconstruction accuracy for a given model, the more closely the brain tracks the predictors in that model.

5.2.9 Model comparison

We tested the unique contribution of each syntactic predictor by comparing the reconstruction accuracy of the full model to the reconstruction accuracy of a null model in which only one of the predictors was omitted. The three null models we evaluated were Bottom-up + Top-down, Bottom-up + Left-corner, and Top-down + Left-corner (see Table 5.2). Comparing their reconstruction accuracy to the accuracy of the full model yields an accuracy difference measure for the left-corner, top-down, and bottom-up predictors, respectively. This comparison thus tests whether a predictor explains variance in the brain signal above and beyond the variance explained by all other predictors.

To determine where in the brain the reconstruction accuracy of the full model was different from that of the null model, we smoothed the source points of both models separately (Gaussian window, SD = 14 mm) and tested for differences in their source maps using non-parametric cluster-based permutation tests (Maris & Oostenveld, 2007). For all contrasts between full and null model, we applied two-tailed paired-samples t-tests at each source point, clustered adjacent source points (minsource = 10) with an uncorrected p-value lower than 0.05, and evaluated clusters of activity by comparing their cluster-level test statistic (sum of individual t-values) to a permutation distribution. The permutation distribution was generated based on the maximum cluster-level t-value in each of 10,000

random permutations of the same data, in which the condition labels were shuffled within subjects. The significance of clusters was evaluated at an alpha value that was Bonferroni-corrected for the number of tests (alpha $= 0.05/n_{tests}$). As an estimation of the effect size of the significant clusters, we report t_{av}, which corresponds to the average t-value within the significant cluster (the cluster-level t-value divided by the number of significant source points). When multiple clusters are significant, we report the test statistic of the cluster with the largest number of significant source points.

The syntactic predictors were quite highly correlated ($r_{BU-TD} = 0.43$, $r_{LC-BU} = 0.79$, $r_{TD-LC} = 0.83$; see Supplementary Information S5.2), potentially leading to multicollinearity, which affects estimation of the TRF coefficients. To control for this possibility, we separately compared the Bottom-up, Top-down, and Left-corner models to a base model, which included no syntactic predictors at all (see Table S5.2 in the Supplementary Information S5.2). This analysis does not suffer from multicollinearity issues because the correlated predictors never appear in the same model. The results of this analysis are both qualitatively and quantitively very similar to those reported in the main analyses (see Supplementary Information S5.2), supporting the conclusion that the different predictors explain unique variance in the MEG data despite being positively correlated.

5.2.10 Evaluation of the response functions

In addition to analyzing the fit between the predicted and the observed signal, we also evaluated the estimated response function, which provides information about the temporal relationship between the predictor and the neural response. This analysis involves the coefficients of the TRF at each time and source point (sources smoothed by a Gaussian window, SD = 14 mm). If the coefficients for a given predictor are significantly non-zero, this indicates that the brain responds to the information encoded in that predictor. In a spatiotemporal cluster-based permutation analysis, we first applied two-tailed one-sample t-tests at each source-time point to determine whether the TRF coefficients deviate from zero. The t-values of adjacent source-time points (minsource = 10, mintime = 40 ms) with an uncorrected p-value lower than 0.05 were then summed, and their cluster-level test statistic was compared to a permutation distribution based on 10,000 random permutations of the same data. The significance of clusters was evaluated at an alpha value (of 0.05) that was Bonferroni-corrected for the number of tests.

5.3 Results

5.3.1 Model comparison

Using cluster-based permutation tests in source space, we tested where in the brain the reconstruction accuracies were modulated by each of the three syntactic predictors. Clusters of activity were evaluated at alpha = 0.0083 (n_{tests} = 3 accuracy differences * 2 hemispheres). All predictors significantly improve reconstruction accuracy, with clusters mostly in the left hemisphere (see Figure 5.3). The improvement is largest for the top-down predictor, which explains variance in many regions of the left hemisphere (t_{av} = 5.49, p = .0034), as well as in an anterior part of the right temporal lobe (t_{av} = 2.60, p = .0075). The strongest left-hemispheric effects are found in superior and middle temporal regions and in inferior and middle frontal regions. The bottom-up predictor similarly engages inferior frontal and temporal regions, but only one cluster around Heschl's gyrus is significant at the adjusted alpha level (t_{av} = 3.84, p = .0081). Last, the left-corner predictor improves reconstruction accuracy in an area at the border of the temporal and frontal lobe in both the left (t_{av} = 5.72, p = .0057) and the right hemisphere (t_{av} = 3.35, p = .0081). It is noteworthy that even though the three syntactic predictors are positively correlated with one another, each of them explains unique variability in the MEG data that could not be attributed to the other two syntactic predictors.

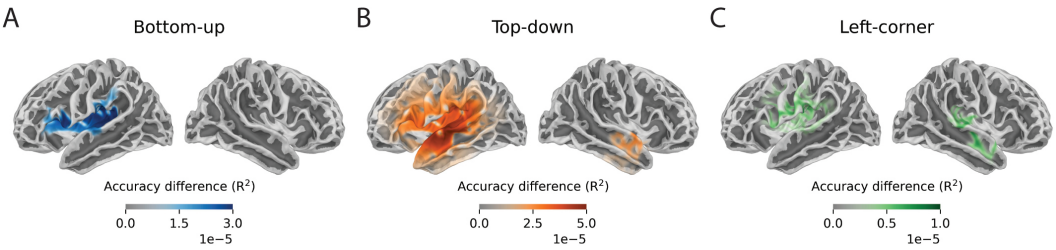


Figure 5.3: Sources of improved explained variance of the bottom-up (A), top-down (B) and left-corner (C) predictors, determined by comparing the reconstruction accuracy of the full model to the reconstruction accuracy of a null model from which the relevant predictor was omitted. All clusters that were significant at uncorrected alpha = 0.05 are displayed. Notice that the scales of the color bars are different across the plots.

5.3.2 Evaluation of the response functions

Given that all of the predictors increase reconstruction accuracy, we can examine their response functions, which reveal a more detailed picture of the time course of the neural response to syntactic information. Figure 5.4 shows the TRFs averaged over the significant regions from the cluster-based source analysis of reconstruction accuracies, in the left and the right hemisphere separately. These plots show the absolute value of the TRF coefficients and thus yield an estimate of the magnitude of the brain response to the syntactic predictor at each time point. All TRFs come from the full model, in which all predictors are competing for explaining variance, so the increases in amplitude reflect components of the neural response that are best explained by the respective predictor. In line with the reconstruction accuracy results, Figure 5.4 shows clearly that the neural response to the information encoded in top-down node counts is stronger than the response to node counts derived from the bottom-up or the left-corner method.

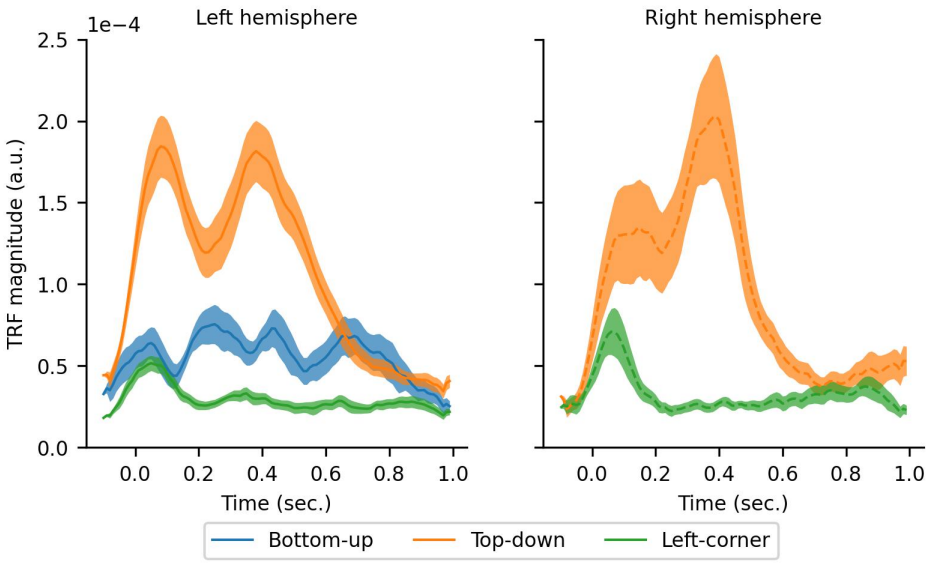


Figure 5.4: Temporal response functions for node count derived from bottom-up, top-down, and left-corner parsers in the full model, averaged over sources in the left and right hemisphere that showed a significant improvement in reconstruction accuracy. The error bars reflect the standard error of the mean per time sample. The plot on the right does not have a TRF for bottom-up because none of the clusters in the right hemisphere showed a significant improvement in reconstruction accuracy after the addition of the bottom-up predictor.

While these results are informative about the strength and timing of the response, they provide limited information about the direction of the effects. In order to analyze when the brain responds most strongly to the information encoded in a predictor, it is necessary to take the absolute value of the TRF coefficient at each time and source point, as this prevents TRF coefficients with opposite signs from cancelling when they are averaged across different sources. This is often justified because the sign of the TRF coefficient is not directly functionally interpretable (in the same way that the negative polarity of an ERP is not interpretable). Yet, despite this interpretability problem, differences in the sign of the TRF coefficients for the three parsing models within the same brain region indicate time-dependent differences in the functional involvement of that region.

Figure 5.5 therefore shows the source t-values (based on two-tailed, one-sample t-tests) of the TRFs for the three syntactic predictors, split up into four time windows (corresponding to different delays in TRF estimation). Non-parametric cluster-based permutation tests were used to determine when and where the TRF coefficients of each syntactic predictor deviated from zero. Because this involves 6 comparisons (3 TRFs * 2 hemispheres), clusters were evaluated at alpha = 0.0083. Focusing on left-hemispheric sources, this analysis revealed a negative cluster for the top-down predictor, broadly distributed in frontal and temporal regions (t_{av} = -0.53, $p < .001$)[3], and a positive cluster around the parahippocampal gyrus and anterior temporal lobe (t_{av} = 0.68, $p < .001$). The strongest effect of the bottom-up TRF was in a region centered around the left inferior frontal cortex (t_{av} = 0.43, $p < .001$), and the TRF of the left-corner predictor peaked in temporal (t_{av} = -0.38, $p = .0062$) and frontal regions (t_{av} = 0.48, $p < .001$). In sum, these results show that the relationship between the predictor and certain brain regions is time-dependent, suggesting that these areas are differently involved across different delays.

5.3.3 Region of interest analysis

To further explore the spatiotemporal differences between the response functions of the syntactic predictors, we analyzed the TRFs in three specific regions of interest (ROIs) that have been linked to syntactic structure building in naturalistic contexts. These ROIs are the inferior frontal gyrus (IFG), superior temporal

[3]These average cluster-level t-values are underestimated because they are corrected for the total number of significant source points across the whole duration of the cluster. As not all of these source points are significantly involved at all time points (and might at certain time points even have a sign opposing the sign of the cluster), the correction is overly strict.

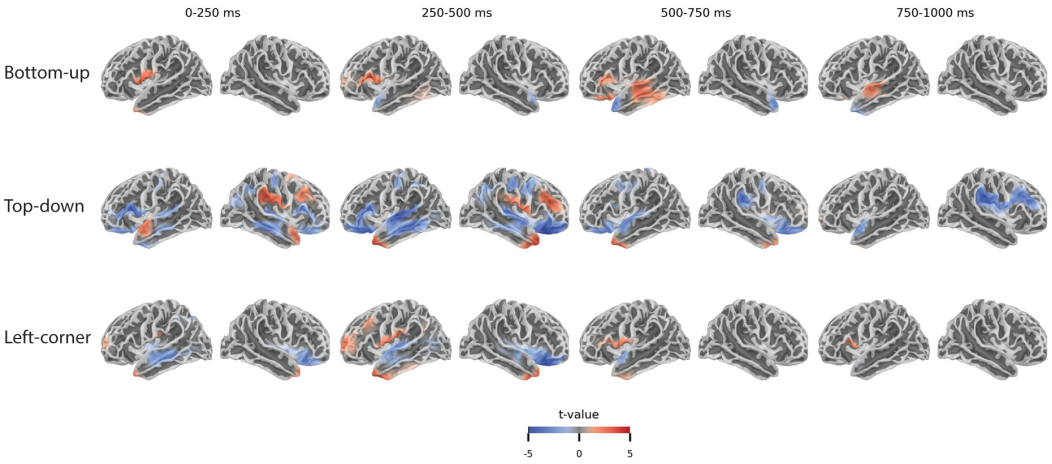


Figure 5.5: Sources of the TRFs for node count derived from bottom-up, top-down, and left-corner parsers, representing early to late responses. The colors represent positive (red) or negative (blue) t-values in sources that were significantly responsive (at corrected alpha = 0.0083) to the predictor in the indicated time windows.

gyrus (STG) and anterior temporal lobe (ATL) in the left hemisphere (Figure 5.6B), all of which also showed up in one or more of the contrasts in the accuracy analysis (Figure 5.3) and the TRF analysis (Figure 5.5). They were extracted using the 'aparc.a2009s' FreeSurfer parcellation (Destrieux et al., 2010).

The reconstruction accuracy results in Figure 5.6A show for each ROI the improvement in reconstruction accuracy when the syntactic predictors are separately added to the null model. Effects of the top-down predictor are strongest in general, and in particular in the STG, where all subjects showed evidence of responses associated with top-down node counts. In each of the three ROIs, we used cluster-based permutation tests to determine when the TRFs of each syntactic predictor deviated from zero. Because this involves 9 comparisons (3 TRFs * 3 ROIs), clusters were evaluated at alpha = 0.0056. As shown in Figure 5.6C, the top-down predictor showed effects in the IFG (from -80 to 470 ms, t_{av} = -3.84, $p < .001$), STG (from -80 to 110 ms, t_{av} = 3.62, $p = .0017$), and ATL (two peaks with opposite signs, from -80 to 240 ms, t_{av} = -3.55, $p = .0013$; from 360 to 860 ms, t_{av} = 3.56, $p < .001$). Note that the TRF coefficients are significantly non-zero before the onset of the top-down predictor, which is not implausible because of the predictive nature of top-down node counts as well as coarticulation in the speech signal (Brodbeck et al., 2022). The bottom-up predictor showed a brief effect in the ATL (from 260 to 450 ms, t_{av} = 2.88, $p < .001$) and a longer one in the IFG (from 370 to 750 ms, t_{av} = 3.06, $p < .001$),

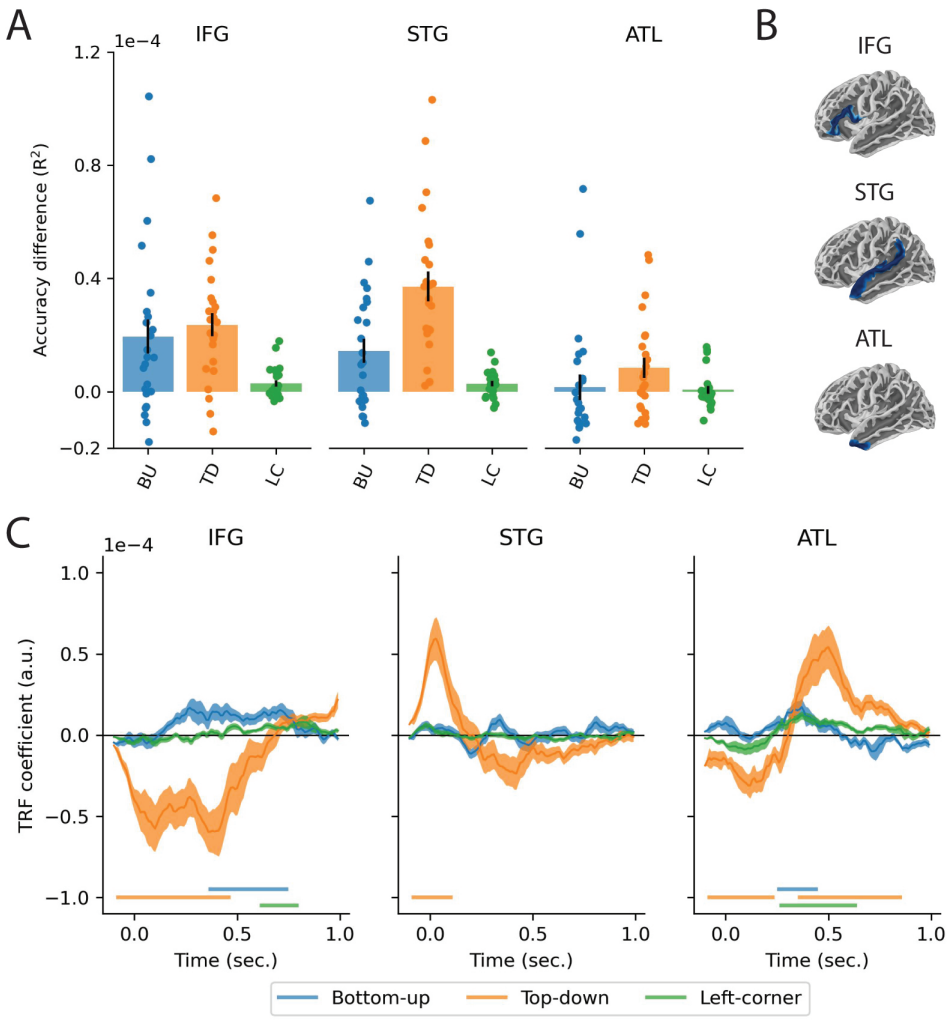


Figure 5.6: Region of interest analysis. (A) Difference in reconstruction accuracy with the full model, plotted for left IFG, STG, and ATL. The labels on the x-axis refer to the syntactic predictors that were taken out of the full model, so the height of each bar indicates the reduction in reconstruction accuracy compared to the full model when only that predictor is omitted. The drops represent the accuracy difference for individual participants, and the error bars represent the standard error of the mean across subjects. (B) Spatial extensions of the three regions of interest. (C) Temporal response functions for node count derived from bottom-up, top-down, and left-corner parsers in the full model. Error bars reflect the standard error of the mean per time sample. The horizontal bars below the TRFs reflect the time points at which the TRFs were significantly non-zero.

while the left-corner predictor showed a shorter effect in the IFG (from 620 to 800 ms, $t_{av} = 2.88$, $p = .0021$) and a longer one in the ATL (from 270 to 640 ms, $t_{av} = 3.69$, $p < .001$). It is noteworthy that the TRFs within the same region are not consistent across different predictors. While their sign is not directly inter-

pretable, the differences in Figure 5.6C (e.g., the consistently positive TRF for bottom-up vs. the negative TRF for top-down in the IFG) indicate that the processes these regions are involved in are not the same for the different syntactic predictors.

5.4 Discussion

In this study, we aimed to investigate how people build syntactic structure during naturalistic comprehension of Dutch. Using a forward modeling approach to map syntactic complexity metrics onto delta-band source activity, we compared three parsing models that differ in the dynamics of structure building. A key finding of these analyses is that neural source dynamics most strongly reflect complexity metrics derived from a top-down parsing model, suggesting that predictive structure building plays an important role in Dutch sentence comprehension. However, the additional (weaker) effects of the bottom-up and left-corner predictors indicate that other parsing strategies also play a role, and suggest that people's parsing strategy might be flexibly adapted to the specific properties of the linguistic input.

5.4.1 Predictive structure building in the brain

Node counts derived from all three parsing models explained unique variance in delta-band MEG activity, consistent with recent studies showing a relationship between delta-band activity and syntactic processing (Bai et al., 2022; Brennan & Martin, 2020; Coopmans, de Hoop, Hagoort, & Martin, 2022; Ding et al., 2016; Kaufeld et al., 2020). Of all syntactic predictors, node counts derived from a top-down parsing method were the strongest syntactic predictor of brain activity in language-relevant areas. These effects peaked twice within the first 500 ms after word onset and encompassed mostly inferior and middle frontal, and superior and middle temporal areas in the left hemisphere. The predictiveness of top-down node counts is at odds with previous studies that have looked at different parsing models in naturalistic comprehension, which either find that top-down methods are less predictive of brain activity (Giglio, Ostarek, Sharoh, & Hagoort, 2022; Nelson et al., 2017) or that they do not differ from other parsing methods (Brennan et al., 2016). What could account for the strong top-down effects? One explanation is that top-down node count values capture the predictive nature of language processing well. There is substantial evidence from

psycholinguistics that people generate structural predictions across a variety of syntactic constructions (Arai & Keller, 2013; Ferreira & Qiu, 2021; Lau et al., 2006; Staub & Clifton, 2006; Yoshida et al., 2013), and they do so in naturalistic contexts as well (Brennan & Hale, 2019; Hale et al., 2018; Heilbron et al., 2022). These predictive structure-building processes are mostly associated with activity in the left posterior temporal lobe (PTL; Brennan et al., 2016; Matchin et al., 2017, 2019; Matar et al., 2021; Nelson et al., 2017) and left inferior frontal gyrus (IFG; Matchin et al., 2017), both of which were responsive to the top-down predictor in the current study. Matchin et al. (2017) suggest that the PTL is involved in predictive activation of sentence-level syntactic representations and/or increased maintenance of the syntactic representations associated with lexical items when they are presented in a sentential context (see also Matar et al., 2021). On both interpretations, the PTL encodes structural representations that can be activated in a predictive fashion and are later to be integrated with the sentence-level syntactic representation in IFG (Hagoort, 2005; Hagoort & Indefrey, 2014; Snijders et al., 2009). Importantly, this process does not proceed in a purely feedforward manner, but rather relies on recurrent connections between temporal and frontal regions (Hultén et al., 2019). The fact that the TRF for the top-down predictor is bimodal is consistent with this idea. The first peak likely reflects predictive structure building, which can occur in the PTL within the first 150 ms after word onset (Matchin et al., 2017). In terms of timing, the second peak is largely consistent with syntactic surprisal effects in recent naturalistic M/EEG studies (Brennan & Hale, 2019; Hale et al., 2018; Heilbron et al., 2022), and might reflect responses related to the (dis)confirmation of predicted structures based on incoming information. An additional observation is that the effects of the top-down predictor, and to a lesser extent also those of the left-corner predictor, are somewhat bilateral. A tentative explanation for this finding is that the prediction of hierarchical structure, as quantified by top-down and left-corner node counts, is demanding and therefore requires support from right-hemispheric regions. These areas are not the locus of the syntactic representations and computations themselves but might be activated when processing demands are increased.

The bottom-up TRF was consistently modulated in the left inferior frontal cortex (Figure 5.5), in line with previous studies that found this area to be responsive to parametric manipulations of constituent structure (Giglio, Ostarek, Weber, & Hagoort, 2022; Matchin et al., 2019; Pallier et al., 2011; Zaccarella et al., 2017). Compared to the top-down TRF, however, bottom-up effects were rela-

tively weak in amplitude. One possibility is that the bottom-up and top-down effects are inversely related, such that strong top-down effects, which indicate predictive processing, are accompanied by weak bottom-up effects, and vice versa. One of the advantages of prediction in language comprehension is that it can reduce the burden on future integration processes (Kuperberg & Jaeger, 2016). If a structural representation has already been pre-built or pre-activated, incoming words only have to be inserted into the existing structure, so integration costs for these words are low. Because integration costs are approximated via bottom-up node count, any effects of bottom-up node count should be reduced if people engage in predictive processing (indicated by top-down effects). This account would thus predict that when top-down metrics modulate brain activity for a given sentence, bottom-up metrics will not provide a good fit for that same sentence, and conversely, when top-down metrics do not provide a good fit, bottom-up metrics should be highly predictive. While it should be investigated in future work whether the effects of top-down and bottom-up metrics indeed go hand in hand in this anti-correlated way, the results of a recent naturalistic study are consistent with this possibility. In this fMRI study by Giglio and colleagues, English participants had to listen other people's verbal summaries of a tv episode (Giglio, Ostarek, Sharoh, & Hagoort, 2022). Contrasting with coherent audiobook narratives, these spontaneously produced summaries contained dysfluencies and corrections, which might make participants less inclined to rely on a predictive processing strategy. Giglio et al. (2022) found that in this comprehension scenario, integratory bottom-up node counts but not predictive top-down node counts modulated activity in language-relevant brain areas (in particular, LIFG and LPTL). This result suggests that parsing strategies can be flexibly adapted to the specific properties of the current linguistic input (e.g., grammatical properties, sentence complexity, reliability of predictive cues), such that people are less likely to engage in predictive structure building if the input contains ungrammatical sentences that make predicting ineffective (Brothers et al., 2019; Kuperberg & Jaeger, 2016).

Effects of the left-corner parser were also relatively weak compared to effects of the top-down parser, in terms of both reconstruction accuracy and TRF amplitude. Region-of-interest analysis focused on the left ATL did show a modulation of the left-corner TRF around 300-600 ms, in line with previous work (Brennan & Pylkkänen, 2017), but this effect was not accompanied by a consistent increase in ATL reconstruction accuracy, suggesting large variability across subjects and/or trials. The comparatively weak effects of the mildly-predictive left-corner

parser might indicate that the left-corner parser is insufficiently predictive to account for the comprehension of head-final constructions of Dutch. As noted in the introduction, the left corner of head-initial structures is very informative, which could explain why left-corner parsing metrics successfully predict brain activity of participants listening to English (Brennan & Pylkkänen, 2017; Nelson et al., 2017). These effects might be weaker in languages with head-final constructions, in particular if speakers of these languages adopt predictive parsing strategies. For instance, while experiments on head-initial structures commonly find that processing costs are increased when the distance between the head and its dependents is increased (Gibson, 1998; Bartek et al., 2011), studies that investigate head-final structures have reported the exact opposite. They find a facilitation in processing of words integrating longer dependencies, an effect that is typically attributed to an increase in the prediction for the upcoming head (Konieczny, 2000; Husain et al., 2014; Nakatani & Gibson, 2008). In addition, it is commonly mentioned that the left-corner strategy predicts processing breakdown for exactly those constructions that are difficult to process. Left-corner parsers have the property that their memory demands increase in proportion to the number of embeddings in center-embedded constructions, while they remain constant for both right- and left-branching structures (Abney & Johnson, 1991; Johnson-Laird, 1983; Resnik, 1992). Sentences with multiple levels of center-embedding indeed quickly over-tax working memory resources (Miller & Chomsky, 1963), supporting the cognitive plausibility of the left-corner method. Intriguingly, however, processing difficulty for center-embedded constructions is not consistent across languages. Vasishth et al. (2010) find that speakers of German (a language with head-final VPs, like Dutch) are hindered less than English speakers during the comprehension of multiply center-embedded sentences, again suggesting that people's ability to generate syntactic predictions might be dependent on the specific grammatical properties of the language.

In all, the fact that a top-down parser best explains brain activity of people listening to Dutch audiobooks might have to do with certain grammatical properties of Dutch, including its head-final VPs, which make left-corner prediction inadequate. This conclusion underscores the need for more work on typologically diverse languages, whose structural properties invite different parsing strategies that might rely on different brain regions to varying degrees (Bornkessel-Schlesewsky & Schlesewsky, 2016). Overall, the fronto-temporal language network is remarkably consistent across speakers of different languages (Malik-Moraleda et al., 2022), but structural differences within this network can be

induced by experience with sentence structures that elicit different processing behavior (Goucha et al., 2022).

5.4.2 Is node count the right linking hypothesis?

Two critical questions can be raised about our use of node count as the complexity metric to represent syntax-related cognitive states. First, the X-bar structures that we used to compute node counts contained empty elements, such as traces. As traces do not have an acoustic correlate in the physical stimulus, we assigned their node count values to the subsequent word. We reasoned that the existence and location of an element, whether covert or overt, can usually be inferred with absolute certainty only after it has been encountered, which would be at the subsequent word. However, this wait-and-see (or wait-and-infer) approach is somewhat inconsistent with the parsing strategy of both top-down and left-corner parsers, which build structure predictively. On the top-down method, for instance, a constituent node is postulated before there is any evidence for its existence. Given that the structure corresponding to overt elements is built predictively, it is inconsistent if the structure corresponding to covert elements is built in an integratory manner, in particular given the evidence for prediction of null forms (Lau et al., 2006; Yoshida et al., 2013).

In order to check whether the node counts for traces were reassigned correctly, we should test the reconstruction accuracy of syntactic predictors in which node counts for traces are assigned to the previous word. This will shift the structural complexity of the sentence to a different point in time, and will do so differently depending on both the location of the trace and the parsing method. For traces at the left corner of a constituent (e.g., the trace in the SpecVP in Figure 5.1), whose node counts are higher for top-down than for bottom-up parsers, the preceding word will be assigned a higher node count on the top-down method. The reverse is the case for traces at the right corner of a constituent (e.g., the trace corresponding to the position of the non-finite V in Figure 5.1), because their node counts are higher on the bottom-up method. This shows that the method of reassigning node counts to the word preceding or following the trace has important consequences.

It has been shown that the influence of the parsing algorithm on node count depends on the grammar that is used. For instance, node counts for top-down and bottom-up parsers are strongly correlated for minimalist grammars (Van Wagenen et al., 2014), supposedly showing that the influence of the parsing algorithm is small when an expressive grammatical formalism is used. However, it

appears that this is the case only when node counts for empty elements are ignored. When their nodes are counted and subsequently reassigned to another word, top-down and bottom-up parsers make different predictions about when in the sentence processing complexity is high.

Another relevant question is whether node count is the right measure to represent syntactic structure building. The node count metric we used is unlexicalized, which means that it does not take into account the label of the node counted. During language comprehension, however, syntactic processing is lexicalized (Coopmans, de Hoop, Hagoort, & Martin, 2022; Hagoort, 2005), and lexical information guides predictive structure building (Arai & Keller, 2013; Boland & Blodgett, 2006; Schütze & Gibson, 1999). Such lexically driven structural predictions are represented to some extent in other metrics, such as part-of-speech surprisal values derived from probabilistic context-free grammars (Brennan & Hale, 2019; Brennan et al., 2016; Shain et al., 2020) or recurrent neural network grammars (Brennan et al., 2020; Hale et al., 2018). Both types of grammars incrementally build hierarchical structure, which they use to conditionalize the probability of an upcoming word's part-of-speech. They can thus quantify how likely a given structural analysis is given a preceding context that includes lexical information.

There are at least two other metrics that are informative about syntactic processing in a way that node count is not. First, the 'distance' metric counts the total number of syntactic analyses that are considered by a parser at every individual word (Brennan et al., 2020; Hale et al., 2018). The larger the number of alternative analyses to be considered, the higher the effort in choosing the correct parse. In this way, it models the ambiguity resolution process that is much studied in psycholinguistics but that is not captured by node count. Second, 'incremental memory' reflects the number of phrases to be held in memory on a stack (Nelson et al., 2017; Van Wagenen et al., 2014) and is sometimes quantified as the number of open nodes (Nelson et al., 2017; Giglio, Ostarek, Sharoh, & Hagoort, 2022). Incremental memory is particularly relevant to evaluate left-corner parsing because one of the arguments in favor of the plausibility of the left-corner method relies on a complexity metric that reflects the number of unattached constituents to be held in working memory (Abney & Johnson, 1991). Node count instead quantifies syntactic complexity rather than memory load, and it need not be the case that these make exactly the same predictions with respect to comprehension difficulty.

5.5 Final remarks

Hale et al. (2022) argue that in order for neuro-computational models to be useful for the neurobiology of language, they must be linguistically interpretable, which means that "the models connect with or implement theoretical constructs from linguistics" (p. 439). In the current study, we linked brain activity to syntactic complexity metrics that were derived from syntactic structures, which were generated via an expressive grammar that is linguistically interpretable. In this sense, our findings can go beyond recent naturalistic studies which show that deep neural networks (DNNs) strongly predict people's brain activity during natural story listening (Heilbron et al., 2022; Schrimpf et al., 2021). Despite being very successful in predicting activity in the fronto-temporal language network, these results are not easily interpretable in terms of the functional properties of this network, because DNNs typically do not lend themselves to formal analysis. Neuro-computational models as used in the current work are linguistically interpretable, thus allowing us to conclude that inferior frontal and superior temporal regions are involved in building hierarchical constituent structure in a predictive manner. We do, however, have to remain agnostic with respect to the kinds of functions these areas are computing. This has to do with the fact that we computed node count based on derived tree structures rather than on the actual derivation trees that reveal the parsing steps by which a structure is incrementally constructed. In order to build neuro-computational parsing models that are not only linguistically interpretable but also psycholinguistically accurate, it is important that future work assumes a more transparent relation between grammar and parser, for instance by using the derivation steps directly implemented by the grammar (Brennan et al., 2020; Chesi, 2015; Hale et al., 2018; Stanojević et al., 2021).

S5 Supplementary Information

S5.1 Auditory stimuli

Table S5.1: Auditory stimuli.

Story part	Duration
Andersen s1 p1	4 min 58 sec
Andersen s1 p2	5 min 17 sec
Andersen s1 p3	4 min 49 sec
Andersen s1 p4	5 min 50 sec
Grimm s20 p1	6 min 6 sec
Grimm s20 p2	6 min 40 sec
Grimm s23 p1	5 min 3 sec
Grimm s23 p2	5 min 32 sec
Grimm s23 p3	5 min 2 sec

S5.2 Comparisons against the base model

The syntactic predictors are positively correlated. This is mainly the case for the left-corner predictor, whose Pearson correlation with the bottom-up and top-down predictors is 0.79 and 0.83, respectively (Figure S5.1). A high correlation between predictors in a regression analysis can lead to multicollinearity, which can in turn result in increased variance of their TRF coefficients (Weissbart et al., 2020). This issue typically emerges when the variance inflation factor (VIF) is above 5, which indicates that the variance in one predictor can be explained by a linear combination of the other predictors (Sheather, 2009). We computed the VIF for each predictor in the full model by taking the diagonal of the inverse of the correlation matrix shown in Figure S5.1. This showed that when all predictors are included, the VIF for both top-down ($VIF_{top-down} = 5.86$) and left-corner ($VIF_{left-corner} = 12.64$) is above 5, indicating that multicollinearity between the predictors might hinder TRF coefficient estimation.

As a control, we therefore repeated our analyses with TRF models in which the VIF is below 5 for all predictors. In this control analysis we evaluated whether the addition of each of the three syntactic predictors to a base model (Table S5.2) improves reconstruction accuracy. This analysis is conceptually similar to the analysis reported in the main chapter, but it does not take into account co-dependencies between the different syntactic predictors because these predictors

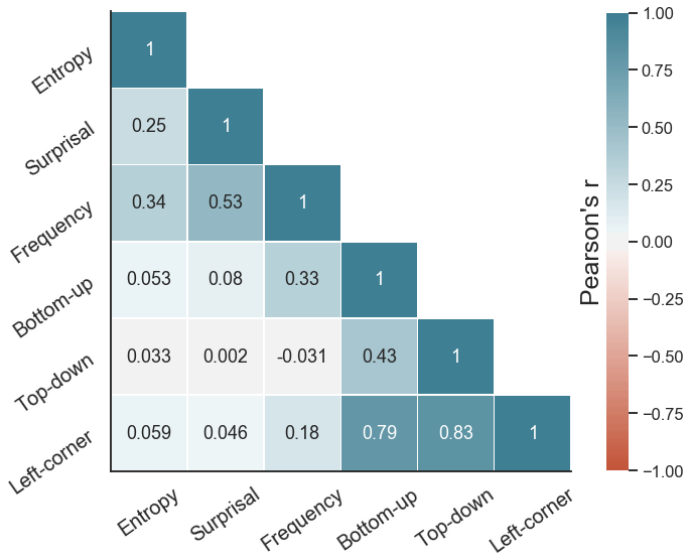


Figure S5.1: Correlation matrix showing the Pearson correlation between all word-based predictors.

never appear in the same model. So, while the analyses in the main chapter provide a conservative measure of reconstruction accuracy, the current analyses might overestimate the reconstruction accuracy of certain predictors that share variability.

Table S5.2: Predictors included in each model.

Predictor(s) / Model name	Spectro-gram/ Onsets	Word onset	Word frequency/ Entropy/ Surprisal	Bottom-up	Top-down	Left-corner
Base	X	X	X			
Bottom-up	X	X	X	X		
Top-down	X	X	X		X	
Left-corner	X	X	X			X

Model comparison

Figure S5.2A shows the sources in which the reconstruction accuracy of the base model was significantly non-zero. The spatial extent of these clusters is consistent with the location of the auditory cortex, likely reflecting the contributions of the two acoustic predictors in the base model. Figures S5.2B-D further show the sources of improvements in reconstruction accuracy when each of the three syntactic predictors are separately added to the base model. For comparison with

the results reported in Section 5.3.1 of the main chapter, the scales of the color bars in Figures S5.2B-D are identical to those in Figure 5.3. Spatial cluster-based permutation tests (cluster-level alpha = 0.0083, Bonferroni-corrected for 6 tests) revealed that the top-down predictor explains variance in frontal and temporal regions of the left hemisphere ($t_{av} = 5.08$, $p = .0036$) and to a lesser extent also the right hemisphere ($t_{av} = 3.16$, $p < .001$). The activation pattern for the left-corner predictor is rather similar, with sources of significant explained variance in the left superior temporal lobe ($t_{av} = 4.19$, $p = .0027$) and a spatially more extended area in the right frontal and temporal lobes ($t_{av} = 3.09$, $p < .001$). The bottom-up predictor engages a smaller region centered around Heschl's gyrus, but none of these clusters are significant at the adjusted alpha level ($t_{av} = 3.54$, $p = .0093$).

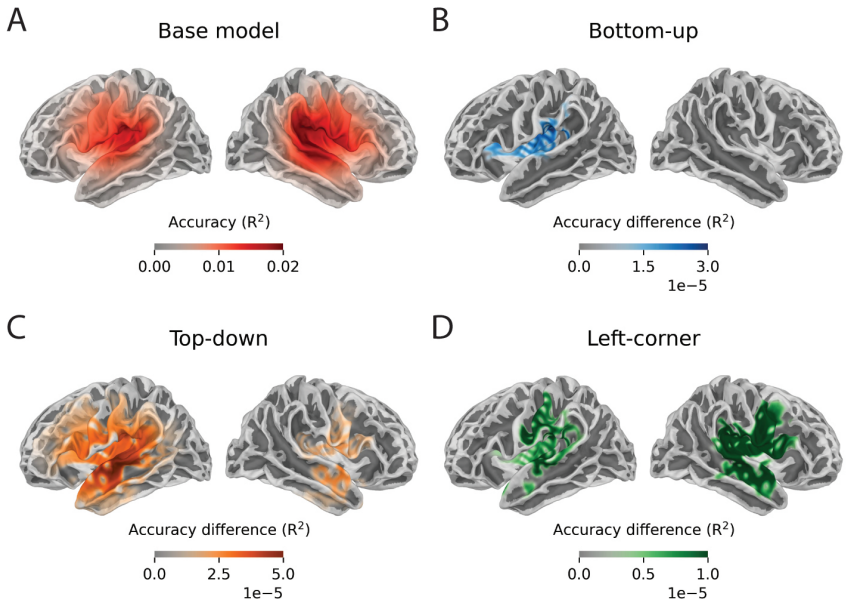


Figure S5.2: Sources of (improved) explained variance of the base model (A) and the predictors reflecting node count from the bottom-up (B), top-down (C) and left-corner (D) methods. All clusters that were significant at uncorrected alpha = 0.05 are displayed. Notice that the scales of the color bars are different across the plots.

Evaluation of the response functions

Figure S5.3 shows the TRFs averaged over the significant regions from the cluster-based source analysis of reconstruction accuries, in the left and the right hemisphere separately. The TRF for each syntactic predictor comes from a model which includes all base predictors as well as one syntactic predictor (e.g., the top-down TRF comes from the Top-down model; see Table S5.2). In line with

the reconstruction accuracy results, Figure S5.3 shows clearly that the neural response to the information encoded in top-down node counts is stronger than the response to node counts derived from a bottom-up or a left-corner parser.

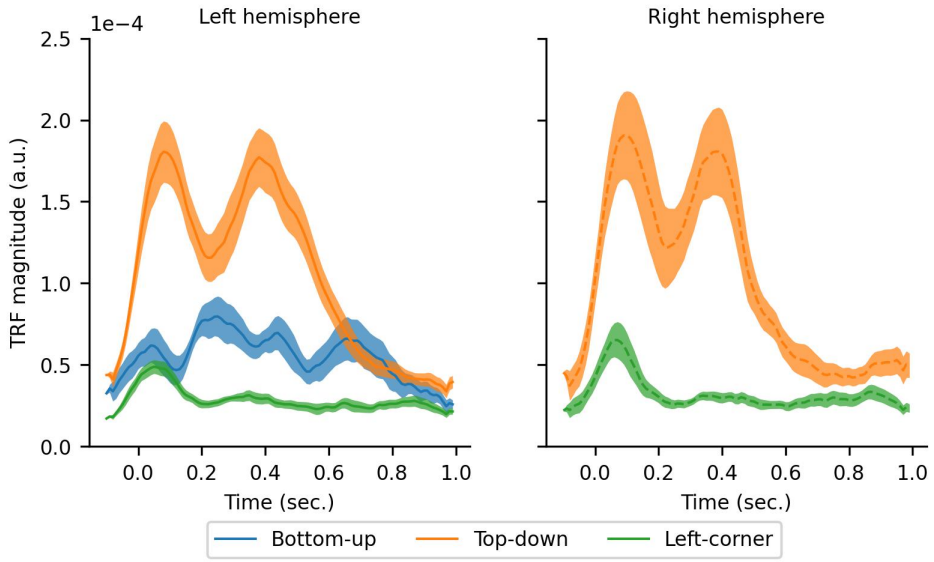


Figure S5.3: Temporal response functions for node count derived from bottom-up, top-down, and left-corner parsers in their respective models, averaged over sources in the left and right hemisphere that showed an improvement in reconstruction accuracy at uncorrected alpha = 0.05. The error bars reflect the standard error of the mean per time sample. The plot on the right does not have a TRF for bottom-up because none of the clusters in the right hemisphere showed a significant improvement in reconstruction accuracy after the addition of the bottom-up predictor to the base model.

Figure S5.4 shows the source t-values (based on two-tailed, one-sample t-tests) of the TRFs for the three syntactic predictors, split up into four time windows (corresponding to different delays in TRF estimation). We used spatiotemporal cluster-based permutation tests (cluster-level alpha = 0.0083, Bonferroni-corrected for 6 tests) to determine when and where the TRF coefficients of each syntactic predictor deviated from zero. Focusing on left-hemispheric sources, this analysis revealed a negative cluster for the top-down predictor in frontal and temporal regions (t_{av} = -0.45, $p < .001$), and a positive cluster in the middle frontal lobe ($t_{av} = 0.65$, $p < .001$). The strongest effect of the bottom-up TRF was again in a region centered around the inferior frontal cortex ($t_{av} = 0.44$, $p < .001$), and the TRF of the left-corner predictor peaked in temporal regions (t_{av} = -0.51, $p = .0073$).

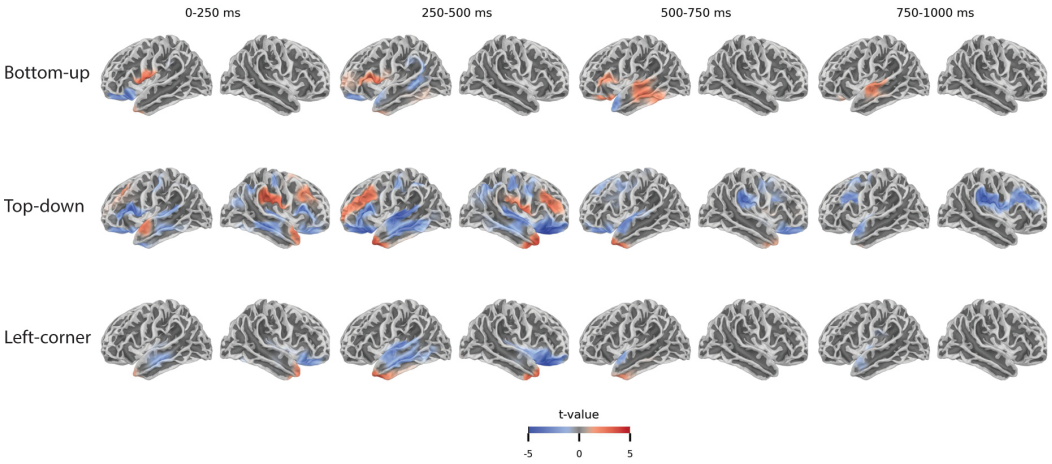


Figure S5.4: Sources of the TRFs for node count derived from bottom-up, top-down, and left-corner parsers, representing early to late responses. The colors represent positive (red) or negative (blue) t-values in sources that were significantly responsive (at corrected alpha = 0.0083) to the predictor in the indicated time windows.

Region of interest analysis

The following three regions of interest (ROIs) in the left hemisphere were extracted using the 'aparc.a2009s' FreeSurfer parcellation (Destrieux et al., 2010): the inferior frontal gyrus (IFG), superior temporal gyrus (STG) and anterior temporal lobe (ATL; see Figure S5.5B). Figure S5.5A shows for each ROI the improvement in reconstruction accuracy when the relevant predictor is added to the base model. Like in the main results, the effects of the top-down predictor are strongest in general, and in particular in the STG, where all subjects showed evidence of responses associated with top-down node counts. In each of the three ROIs, we used cluster-based permutation tests to determine when the TRFs of each syntactic predictor deviated from zero. This involves 9 comparisons (3 TRFs * 3 ROIs), so clusters were evaluated at alpha = 0.0056. As shown in Figure S5.5C, the top-down predictor showed effects in the IFG (from -80 to 460 ms, t_{av} = -3.82, $p < .001$), STG (from -80 to 120 ms, t_{av} = 3.60, $p = .0011$), and ATL (two peaks with opposite signs, from -80 to 240 ms, t_{av} = -3.35, $p = .0016$; from 350 to 830 ms, t_{av} = 3.63, $p < .001$). The bottom-up predictor showed a brief effect in the ATL (from 270 to 450 ms, t_{av} = 3.20, $p < .001$) and a longer one in the IFG (from 370 to 760 ms, t_{av} = 3.07, $p < .001$). Last, the left-corner predictor showed an early peak in the IFG (from -80 to 240 ms, t_{av} = -2.68, $p = .0024$), an early peak in the STG (from -60 to 240 ms, t_{av} = 3.18, $p = .0046$), and a later peak in the ATL (from 310 to 650 ms, t_{av} = 3.48, $p < .001$).

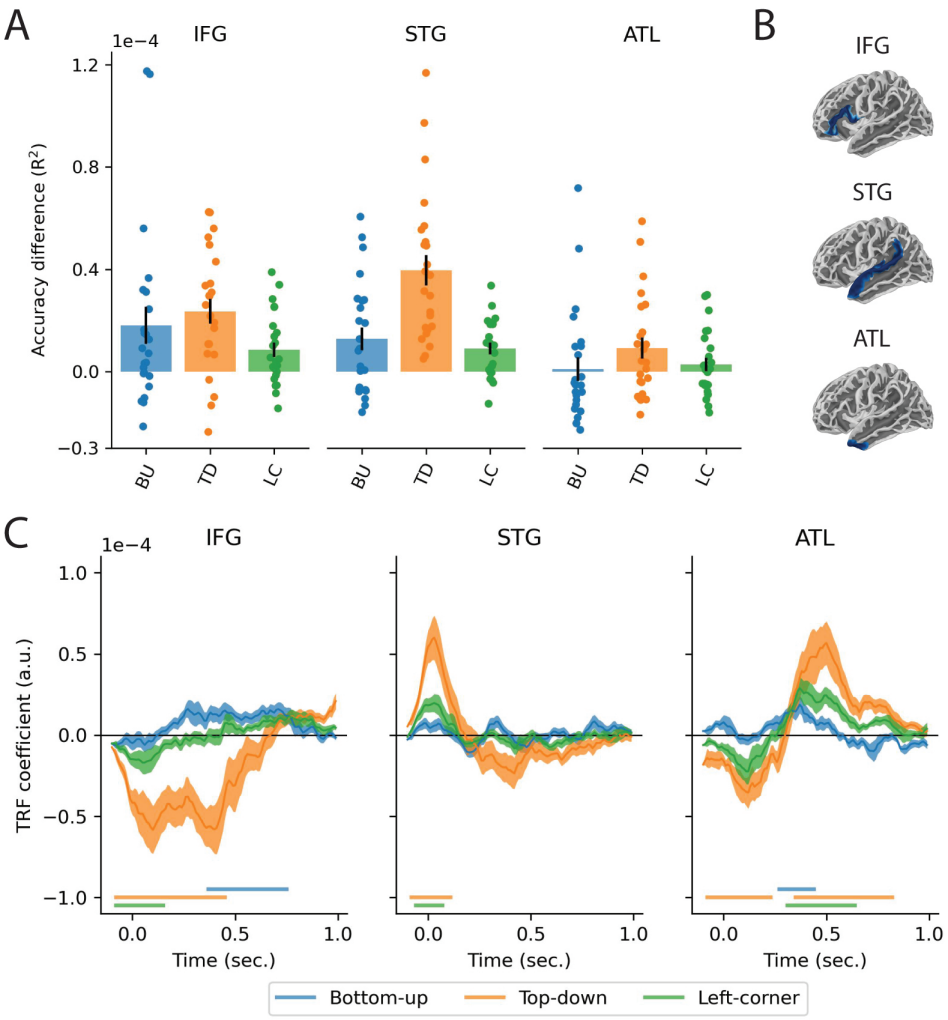


Figure S5.5: Region of interest analysis. (A) Difference in reconstruction accuracy with the base model, plotted for left IFG, STG, and ATL. The height of each bar indicates the improvement in reconstruction accuracy when only the relevant syntactic predictor was added to the base model. The drops represent the accuracy difference for individual participants, and the error bars represent the standard error of the mean across subjects. (B) Spatial extensions of the three regions of interest. (C) Temporal response functions for node count derived from bottom-up, top-down, and left-corner parsers in their respective models. Error bars reflect the standard error of the mean per time sample. The horizontal bars below the TRFs reflect the time points at which the TRFs were significantly non-zero.

In all, these results are largely consistent with the results reported in the main chapter, suggesting that multicollinearity between the predictors did not hinder estimation of the TRF coefficients. However, one notable difference has to do with the left-corner predictor, whose effects are weaker in the results reported in the main chapter. Regarding the response functions, it is noteworthy that

the TRFs for both bottom-up and top-down are stable and quite similar in the full model (Figure 5.6C in the main chapter) and in the simpler models (Figure S5.5C), which shows that they are unaffected by the presence of the other syntactic predictors in the full model. The left-corner TRFs, however, are smaller in size and less variable in the full model than in the simpler Left-corner model. In the latter, the left-corner TRFs are similar to the top-down TRFs (from the Top-down model; Figure S5.5C), suggesting that they are explaining the same variance. In terms of reconstruction accuracy, the explained variance in the Left-corner model attributed to the left-corner predictor (Figures S5.2 and S5.5A) is visibly reduced when the left-corner predictor is added to a null model that already contains both bottom-up and top-down as predictors (Figures 5.3 and 5.6A in the main chapter). The reverse does not happen, suggesting that some of the variance assigned to left-corner in the simpler Left-corner model was assigned incorrectly, 'belonging' to top-down rather than left-corner.

6 | Hierarchical structure in language and action: A formal comparison[1]

Abstract

Since the cognitive revolution, language and action have been compared as cognitive systems, with cross-domain convergent views recently gaining renewed interest in biology, neuroscience, and cognitive science. Language and action are both combinatorial systems whose mode of combination has been argued to be hierarchical, combining elements into constituents of increasingly larger size. This structural similarity has led to the suggestion that they rely on shared cognitive and neural resources. In this chapter, we compare the conceptual and formal properties of hierarchy in language and action using set theory. We show that the strong compositionality of language requires a particular formalism, a magma, to describe the algebraic structure corresponding to the set of hierarchical structures underlying sentences. When this formalism is applied to actions, it appears to be both too strong and too weak. To overcome these limitations, which are related to the weak compositionality and sequential nature of action structures, we formalize the algebraic structure corresponding to the set of actions as a trace monoid. We aim to capture the different system properties of language and action in terms of the distinction between hierarchical sets and hierarchical sequences, and discuss the implications for the way both systems could be represented in the brain.

[1] Adapted from Coopmans, C. W., Kaushik, K., & Martin, A. E. (2023). Hierarchical structure in language and action: A formal comparison. *Psychological Review*.

6.1 Introduction

It has long been recognized that both language and action are structurally organized in a way that is not immediately evident from their serial appearance. In the 1950s, Lashley (1951) and Chomsky (1959) separately showed that then dominant behaviorist 'chaining' theories based on contiguous stimulus-response associations could not account for serial behavior, such as language production and action execution. Instead, these behaviors appear to be controlled by internal, hierarchically organized plans, which allow human behavior to be creative, productive and flexible. Since then, similarities between language and action have often been noted (e.g., Greenfield, 1991; Holloway, 1969; Miller et al., 1960), and more recent studies propose that the two systems are analogous in their hierarchical organization (Fitch & Martins, 2014; Fujita, 2014; Jackendoff, 2007; Pulvermüller & Fadiga, 2010; Stout & Chaminade, 2009).

Such proposals about cross-domain convergence are desirable from an evolutionary perspective, in which one seeks to find a set of primitives that account for the distinguishing features of the human mind (Boeckx & Fujita, 2014; de Waal & Ferrari, 2010; Hauser et al., 2002; Marcus, 2006). However, arguments in favor of the analogy between language and action are formally underspecified. It is possible to draw a hierarchical tree structure over any sequence, but what is needed is independent empirical evidence that this structure describes or explains a phenomenon in the natural world (Berwick & Chomsky, 2017; Bloom, 1994; Fitch & Martins, 2014; Moro, 2014a). In other words, superficial resemblance is insufficient: "we cannot just observe that hierarchical structures are found in motor control (e.g., tool construction), and thereby claim that these are directly related to the hierarchical structures of language ... Rather, it is necessary to develop a functional description of the cognitive structures in question, parallel to that for language ... so we can look for finer-scale commonalities" (Jackendoff, 2002, p. 80).

While formal linguistics has provided many accounts of the specific properties of hierarchy in language, such a formal characterization in the domain of actions and action plans is lacking (but see Steedman, 2002 for an exception). To this end, the aim of this chapter is to characterize the similarities and differences between the hierarchical structures in language and action in both conceptual and formal terms. The chapter is structured as follows: in Section 6.2, we discuss the type of data that shows that the syntax of natural languages is organized hierarchically, after which we list the core properties of such hierarchical syntactic structure (Section 6.2.1). In Section 6.2.2, we formally describe these struc-

tures in a domain-neutral way using the mathematical language of set theory. We then show that this formalism is inadequate for describing the action system (Section 6.3.1) and suggest an alternative formalism to characterize its properties (Section 6.3.2). In Section 6.4, we conclude that the properties of syntactic hierarchy are not found in action structures (Section 6.4.2) and discuss this conclusion in light of the idea that syntactic representations are fundamentally hierarchical sets, while actions are better conceived of as hierarchical sequences (Section 6.4.3). We end by discussing the implications for how language and action might be represented in the brain.

6.2 Hierarchical structure in language

In linguistics, the term hierarchy refers to the format of linguistic representations. At all levels of organization (phrases, words and syllables), linguistic structure is organized hierarchically (see Everaert et al., 2015 for a recent overview). In the domain of syntax specifically, it refers to the fact that words are embedded into constituents, which are in turn recursively embedded into larger constituents, creating the hierarchically organized syntactic structures that are often visually denoted by means of tree structures. These tree structures are graphic representations of relations which are essentially set-theoretic (Lasnik, 2000).

A main source of evidence for constituency is the observation that the interpretation of phrases and sentences is often determined by structural relationships. For example, the sentence “the woman saw the man with binoculars” has two meanings. Either the woman has binoculars, which she uses to look at the man, or the man has binoculars. The sentence is ambiguous because it corresponds to two possible structures, which differ in terms of the attachment site of the prepositional phrase (PP) “with binoculars” (see Figure 6.1). If it attaches to “the man”, forming a complex noun phrase (NP) constituent (Figure 6.1A), the man has the binoculars, but if it attaches to the verb phrase (VP) “saw the man” (Figure 6.1B), the woman must be holding the binoculars. Here, it is the structural relationship between the PP and the other constituents that determines how the sentence is interpreted.

The structure dependence of meaning shows that language is compositional. To be able to compare combinatorial systems, such as language and action, we make a distinction between strong and weak compositionality (Pagin & Westerståhl, 2010). In a strongly compositional system, the meaning of a constructed unit is a function of the meanings of its constituents and the way in which these

are structurally combined (Partee et al., 1993; Partee, 1995). In a weakly compositional system, instead, the meaning of a constructed unit is a function of the meaning of the elements and the total construction (i.e., the result of an operation applied over the total construction of ordered elements; Pagin & Westerståhl, 2010). A weakly compositional system can thus distinguish the meanings of "John likes Mary" and "Mary likes John", because their total constructions differ. However, weakly compositional systems cannot capture structural ambiguity. Because they do not take into account the structural relationships between intermediate representations, such as between the different constituents in Figure 6.1, they are unable to distinguish the two interpretations of "the woman saw the man with binoculars".

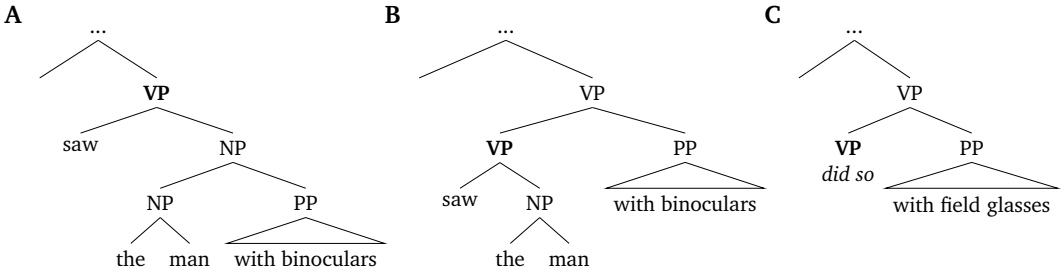


Figure 6.1: Hierarchical structures corresponding to the sentences "(the woman) saw the man with binoculars" (A and B) and "(the boy) did so with field glasses" (C).

A second source of evidence for constituent structure is that syntactic operations, such as deletion and substitution, target constituents rather than words or mere word sequences. For instance, the phrase *did so* can substitute for a verbal word sequence, such as "saw the man", if this sequence forms a constituent. Because the words "saw the man" form an isolated constituent only in the structure of Figure 6.1B, the sentence "the woman saw the man with binoculars and the boy did so with field glasses" (corresponding to Figure 6.1C) can only mean that the boy is holding the field glasses (analogous to the interpretation of Figure 6.1B), not the man. In sum, both semantic interpretation and syntactic operations are structure-dependent: they refer to hierarchical constituent structures rather than to linear sequences of words, with the result that word sequences that do not form constituents are not available to semantic interpretation nor to syntactic operations.

6.2.1 Properties of syntactic structure

To generate such hierarchical structure, (any theory of) the language faculty must include, at a minimum, a computational procedure for combining smaller elements into larger elements. The properties of this procedure are debated, but all linguistic frameworks assume it in one form or another: Merge in the Minimalist Program (Chomsky, 1995b), Unify in the Parallel Architecture (Jackendoff, 2002), Forward/Backward Application in Combinatory Categorial Grammar (Steedman, 2000), and Substitution in Tree-Adjoining Grammar (Joshi & Schabes, 1997). For our purposes, we do not need to select among these frameworks; all we assume is that the generation of syntactic structures relies on a combinatorial operation, whose properties must be formally defined, and which must be computationally general enough so that it could play a role in cognitive domains beyond language. Merge is one of the operators that meets these requirements, as it is formally defined as binary set formation: Merge(α, β) takes two elements α and β and forms the unordered set $\{\alpha, \beta\}$ (Chomsky, 2013; Collins, 2017). It can be applied recursively, such that it takes its own output as input: further combining the already formed set $\{\alpha, \beta\}$ with γ yields the set $\{\{\alpha, \beta\}, \gamma\}$. As should be clear, recursive application of this combinatorial operation yields a structure that is hierarchical: the smaller set is contained in the larger set. Because the generated set is unordered (i.e., $\{\{\alpha, \beta\}, \gamma\}$ is identical to $\{\gamma, \{\beta, \alpha\}\}$), the elements in the set cannot be described in terms of linear precedence. Rather, the relevant relationships are established with respect to structure: the element γ is higher in the structure and has a structurally more prominent position than the elements α and β.

In the remainder of this chapter, we will assume that the combinatorial procedure for generating syntactic structure is binary set formation. On this assumption, the hierarchical structure of syntax has the following properties:[2]

1. **Unbounded**. Human language use is creative: language users can produce and understand sentences that have never been produced before. Specifying such an open-ended capacity using finite means requires recursive procedures, such as the recursive combinatorial operation defined above. While this operation both generates hierarchical structure and ap-

[2] More properties of language can be derived from the minimal assumption that the structure-building procedure is binary set formation (see Hornstein, 2017 and Rizzi, 2013 for comprehensive lists of properties). However, many of these properties, such as displacement, do not have clear analogues in actions (Moro, 2015; Pulvermüller, 2014). Because our aim is to compare the (formal) properties that language and actions might share, we focus on the properties of hierarchy listed here.

plies recursively, hierarchy and recursion are two independent properties. Hierarchy is a property of the output generated by the combinatorial operation (i.e., a property of its extension). Recursion, instead, is a property of a function defined in intension. A recursive function is a function which can apply indefinitely to its own output, leading to structurally 'self-similar' output in which a unit of a specific type is contained in another unit of the same type (in linguistics, this is often called self-embedding: embedding of one thing into another thing *of the same kind*). This results in a hierarchical structure which displays similar properties across different levels of embedding, clearly visible in the repetition of complement clauses like "He said that she believes that he thought ...", which is a sentence within a sentence within a sentence. Note, however, that because a recursive function is defined in intension rather than in extension, the recursivity of a function should not be equated with its output. Absence of self-similar output therefore does not warrant the conclusion that the function generating the output is not recursive (Watumull et al., 2014; Hauser et al., 2014).

The independence of hierarchy and recursion is further illustrated by the fact that they doubly dissociate: not all hierarchical objects are generated by recursion and not all recursive functions generate hierarchical structure. For instance, artificial grammars that generate sequences of the type $(ab)^n$ and a^nb^n can be recursive, via respectively f: S $\rightarrow$ abS and f: S $\rightarrow$ aSb, but only the latter generates hierarchical structure.[3] Conversely, the syllable structure in phonology is hierarchical but not recursive. A syllable contains an onset and a rhyme, with the latter consisting of a nucleus and a coda. This hierarchy is not recursive: a syllable cannot be embedded in another syllable.

2. **Endocentric**. The categorial status of a constituent is determined by one of its elements (the 'head'): the set $\{\alpha, \beta\}$ can be of type α or β, but not of type γ. Endocentric structures are contrasted with exocentric structures, in which the label of a composed unit is not determined by one of its elements.[4] Labels allow phrases to be called upon by interpretive and formal

[3] Note that the grammars that generate $(ab)^n$ and a^nb^n sequences *can be* implemented recursively, though they do not have to be. These sequences can also be generated with iterative functions that are not recursive, i.e., do not call themselves (Fitch, 2010; Jackendoff, 2011). Iterative functions can also realize unboundedness, but they do so by creating sequences without internal structure.

[4] How it is determined which element defines the label of the phrase is still a much-debated question and is outside the scope of this chapter (see e.g., Boeckx, 2009; Chomsky, 2013; Fukui,

procedures, thereby determining their distributional behavior. To give an example, the set {eat, cookies} is a verb phrase, which has 'eat-like' (interpretive) semantic properties and 'verb-like' (formal) syntactic properties, both inherited from the verb "eat". That this is the case can be seen by the fact that "eat cookies" can take the place of the verb "eat" in "He likes to eat", yielding "He likes to eat cookies". It cannot, however, take the place of the noun "cookies" in "He likes chocolate cookies", as is clear from the ill-formedness of "He likes chocolate eat cookies". The label of a composed unit thus places a constraint on further computation, restricting the elements with which it can combine: given that {eat, cookies} is a verb phrase and not a noun phrase, it can combine with adverbs but not with adjectives.

Endocentricity is intricately linked to recursivity, because the combinatorial operation can only be said to apply recursively if its output is of the same type as its input (Boeckx, 2009; Hornstein, 2009; Watumull et al., 2014). Similar to recursivity, endocentricity is a distinctive property of syntactic hierarchy, as not all linguistic structures are endocentric.

3. **Unordered**. Because the combinatorial operation is defined as binary set formation, no order is imposed on the members of the combined set. While the unordered structure has to be linearized for spoken language production, differences in linear order do not feed differences in semantic interpretation, and syntactic operations do not refer to linear order. Different languages (and different modalities) can seem highly different in terms of the linear ordering of their words (e.g., whether heads precede or follow their dependents), which is a fundamental source of cross-linguistic variation (see Section 6.4). However, in terms of the compositional properties of the hierarchical structure generated by Merge, these languages show consistent similarities.

   Note that the assumption about unorderedness is specific to the definition of Merge as binary set formation, and might not be shared in other linguistic frameworks.[5] What these frameworks do agree on, however, is that syntactic operations are structure-dependent, not order-dependent.

2011) What is important here is not how phrases get their labels, but that they get them from one of their elements. Moreover, by using the term labels we only refer to the fact that the combined unit is of the same type as one of its elements. Whether these labels reflect phrasal projections from the syntactic category of a lexical item (as in X-bar theory; Jackendoff, 1977) or rather the lexical item itself (as in bare phrase structure; Chomsky, 1995a) is not critical for our purposes.

[5] See Saito and Fukui (1998) and Kayne (2011), who argue that Merge(α,β) forms the ordered pair $\langle \alpha, \beta \rangle$. This makes *immediate* precedence part of syntax.

This conception of structure building as binary set formation allows us to derive both compositionality and structure dependence. First, the structure of the input to the combinatorial operation is preserved in its output. Thus, if α and β are constituents (or sets) in the input, they are constituents (or sets) in the output as well: new elements can only be added on top of the already formed set, not inside it. Because the structure of every combination is retained at each level of the hierarchy, the hierarchical structure is strongly compositional. This can be shown with a structurally ambiguous phrase: {deep, {blue, sea}} is not the same as {{deep, blue}, sea}. Note that if the structure were not retained after recursive combination, it would be possible to derive from {blue, sea} not only {deep, {blue, sea}} but also {{deep, blue}, sea}. That would make it impossible to account for the ambiguity of the phrase.

Moreover, recursively generated sets describe hierarchical relations but not sequential relations. Therefore, syntactic operations that refer to these sets can only refer to its structure, and hence be structure-dependent, but not to its sequential order. Rules referring a word's linear (ordinal) position are also ruled out by recursion: because it is always possible to recursively insert material *between* two items and thereby change the linear position of the words (e.g., "the boy swims" → "the boy *with muscular arms* swims"), no operation can refer to the linear position of elements in a sequence.

We should note that the properties we described above are properties of a cognitive capacity, which can be expressed in varying degrees in natural languages (e.g., exocentricity might be found in certain subject-predicate relations). Moreover, the faculty of language is capable of assigning strongly compositional interpretations to most sentences, as is required to derive the multiple interpretations of structurally ambiguous sentences, but it can assign other interpretations as well (e.g., to non-decomposable idioms; Baggio, 2021; Jackendoff, 2002). In other words, we listed properties that a model of (the faculty of) language must have, even though these need not be found in all constructions in all languages. As we aim to illustrate how the action system differs from the language system, we will focus on the capacity for strong compositionality as a fundamental difference between both systems.

6.2.2 Formalizing linguistic structure

In order to be able to evaluate the similarities and differences between the hierarchical structure of language and action in a transparent way, we need a theory-neutral conceptual vocabulary to describe these structures. Ideally, this descrip-

tion should be accompanied by a formal analysis of the similarities and differences, as well as an evaluation of their implications (Guest & Martin, 2021a; Martin, 2016, 2020; O'Donnell et al., 2005; Partee et al., 1993; van Rooij & Blokpoel, 2020). To this end, the following paragraphs will present a formal model in which we incorporate the properties of syntactic structure as defined in Section 6.2.1.

Generating structures

Definition 1. $(M, \oplus, \varnothing)$ *is a unital, commutative magma generated from* W*, where:*

1. W *is the set of words that represent the lexicon of a language.*
2. M *is a set of elements that are generated from* W*, with* $W \subset M$*.*
3. $\oplus$ *is a binary set formation operation, such that for* $\forall a, b \in M$*,* $a \oplus b = \{a, b\} = b \oplus a \in M$*. Additionally,* $\oplus$ *is non-associative, so* $\forall a, b, c \in M, (a \oplus b) \oplus c \neq a \oplus (b \oplus c)$*.*
4. $\varnothing$ *is the identity element, such that* $\forall m \in M, m \oplus \varnothing = m = \varnothing \oplus m$*.*

A unital, commutative magma (henceforth referred to as a magma for conciseness) is an algebraic structure (see Box 1), whose operation we define as binary set formation following the formal definition of Merge described in Section 6.2.1. This allows us to derive a number of important properties. First, as the magma axiom states that for any two members $a, b \in M$, application of this operator to a and b generates a member of M, thus yielding unbounded generation. Second, $\oplus$ does not introduce labels, so the label of each set is derived from one of its elements (i.e., endocentricity; see Chomsky, 2013; Collins, 2017).[6] Third, all elements in M are unordered sets. And fourth, $\oplus$ is non-associative, which means that the order in which it is applied affects the structure that is generated (Fukui & Zushi, 2004). In other words, the structures that are generated are strongly compositional: their meaning is a function of the meanings of their parts and the way in which they are structurally combined.

[6] Labels are a convenient way to group together structures with identical formal properties. In our formal setup, constituent labels are simply part labels whose union produces the set of all grammatical structures. For example, with $W = \{dog, man, big\}$, the label N would be the part $\{man, dog\}$, A is $\{big\}$, and NP is $\{\{big, dog\}, \{big, man\}\}$. Therefore, $M = M_N \bigcup M_A \bigcup M_{NP}$.

Box 1. Algebraic structures

An algebraic structure consists of a nonempty set X (called the carrier set), a collection of finitary operations on X (typically binary operations), and a finite set of axioms that these operations must satisfy. To illustrate the relevant axioms for the current work, we consider X as the carrier set and $\odot$ as a binary operation acting on the elements of X.

(1.) $\forall x_1, x_2 \in X,\ x_1 \odot x_2 \in X$ *closed*

(2.) $\forall x \in X, \exists! i \in X$ such that $x \odot i = i \odot x = x$ *unital*

(3.) $\forall x_1, x_2, x_3 \in X,\ (x_1 \odot x_2) \odot x_3 = x_1 \odot (x_2 \odot x_3)$ *associative*

(4.) $\forall x_1, x_2 \in X,\ x_1 \odot x_2 = x_2 \odot x_1$ *commutative*

Depending upon the axioms they satisfy, the algebraic structures form a taxonomy. Presented below is a subset of this taxonomy, in which we highlight both the algebraic structures that are relevant for the current work as well as their corresponding axioms.

Magma	**Unital commutative magma**	**Monoid**	**Commutative monoid**	**Trace monoid**
closed	*closed* *unital* *commutative*	*closed* *unital* *associative*	*closed* *unital* *associative* *commutative*	*closed* *unital* *associative* *partially commutative*

Without further constraints, a freely generated magma would contain elements that should not be constituents, such as {{eat}, {happy}}. To avoid this without modifying the formal properties of $\oplus$, the lexical items themselves must determine which combinations are licensed and which are not. That is, the application of $\oplus$ is constrained by selectional restrictions on its input (i.e., which categories can(not) combine with which other categories). For instance, {{eat}, {happy}} is excluded because verbs do not combine with adjectives. The same restrictions apply when the output of $\oplus$ is recursively used as its input. For example, the set $\{_{\mathrm{V}}\{\text{eat}\}, \{\text{cookies}\}\}$ cannot combine with the adjective "happy" because the former is labeled as a type of verb rather than as a type of noun. Such illegitimate combinations are excluded by taking the grammatically licensed subset of the freely generated magma.

We make the relationship between these constituent structures explicit by defining a binary relationship between the elements of the magma, turning it into a partially ordered magma (see Box 2).

Definition 2. $(M, \oplus, \varnothing, \leq)$ *is a partially ordered magma, where* $\leq$ *is a* ***containment*** *relationship between the elements in* M *that is reflexive, transitive, and antisymmetric.*

The relation $\leq$ on the set M reflects containment or set-inclusion, which corresponds to the dominance relation commonly used in linguistics. Thus, $x_1 \leq x_2$ means that x_2 contains (and thus dominates) x_1. As a visualization of this partially ordered magma, consider the Hasse diagram in Figure 6.2, which displays the containment relationship for two structures that map onto the sequence "woman saw man with binoculars".

Box 2. Ordered sets

An ordered set X is a set ordered by a binary relation, denoted here with infix notation $\leq$, such that $\forall x, y, z \in X$, the following axioms hold (depending on the kind of order):

(1.) $x \leq x$ *reflexive*

(2.) if $x \leq y$ and $y \leq z$, then $x \leq z$ *transitive*

(3.) if $x \leq y$ and $y \leq x$, then $x = y$ *antisymmetric*

(4.) $x \leq y$ or $y \leq x$ *total*

When the binary relation is transitive and antisymmetric, the set is called partially ordered. A totally ordered set is an ordered set whose binary relation holds between all elements of the set. When a relationship is only total when restricted to X', which is a subset of X, we consider X' locally total (Kayne, 1994). We therefore say that $\forall x, y \in X' \subset X$, $x \leq y$ or $y \leq x$.

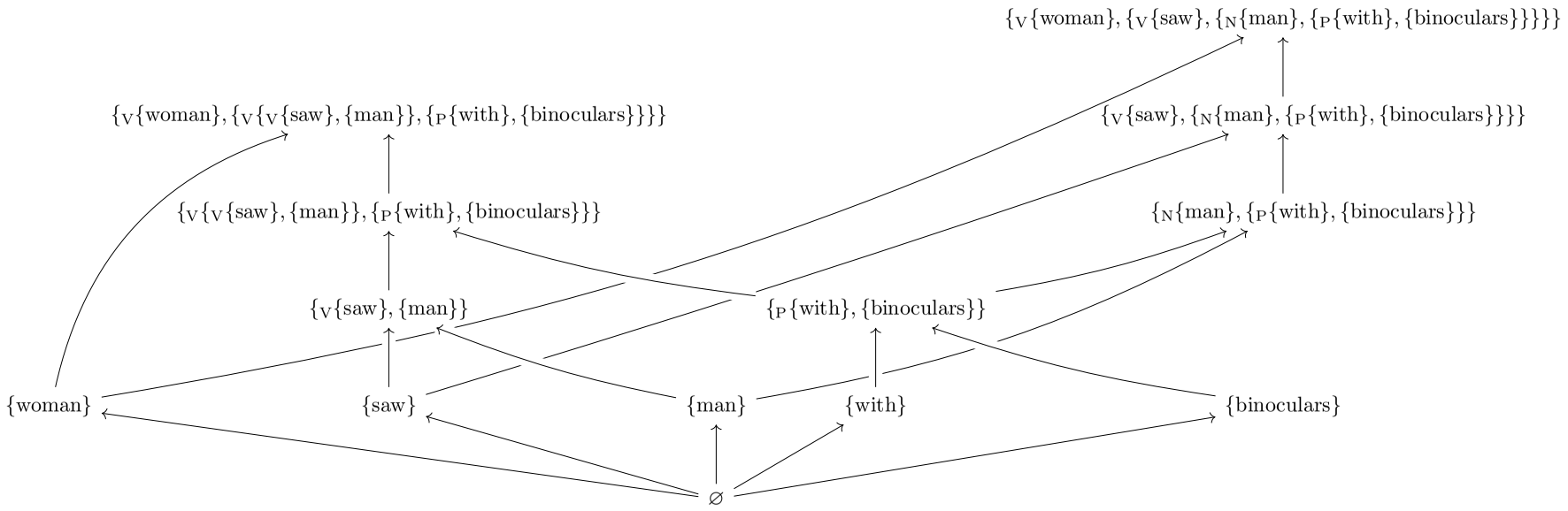


Figure 6.2: Hasse diagram of a subset of the partially ordered magma M, which displays two different structures that map onto the sequence "woman saw man with binoculars". Arrows indicate direct containment. The subscript at the opening curly bracket of each binary set indicates the label of that set.

Besides containment, there is another relevant structural relation between the elements in constituent structures. This relationship, called c-command (Reinhart, 1983), describes the scope domain of a node in the tree structure. Specifically, a node α is said to asymmetrically c-command a node β iff β is contained in the sister node of α (e.g., in Figure 6.1A "saw" asymmetrically c-commands every node contained in the higher NP).

Definition 3. *For $m_1, m_2 \in M$, m_1 c-commands m_2 (denoted $m_1 \gg m_2$) if $m_1 \not\leq m_2$, and $m_2 \not\leq m_1$, and $\exists! m = \{m_1, x\} \in M$, and $m_2 \leq x$. An **asymmetric c-command** relationship exists between m_1 and m_2 if $m_1 \gg m_2$ and $m_2 \not\gg m_1$. Asymmetric c-command is irreflexive, transitive, antisymmetric, and locally total.*

Given Definition 3, asymmetric c-command is a locally total relation on non-terminal nodes in the tree structure.[7] The Hasse diagram in Figure 6.3 visualizes the asymmetric c-command relationship for the two structures that map onto the sequence "woman saw man with binoculars".

Sequences

Definition 4. *$(S, *, ")$ is a monoid generated from W, where:*

1. *W is the set of words that represent the lexicon of a language.*
2. *S is the set of sequences generated from W, with $W \subset S$.*
3. *$*$ is the concatenation operation, which is unital and associative.*
4. *The empty sequence " is the identity element.*

Definition 5. *We define a binary relation ($\prec$) on the elements in $s = (x_1, x_2, ...x_n) \in S$, which we call **precedence**, where $x_1 \prec x_2 \prec ...x_n$. Precedence is irreflexive, transitive, antisymmetric, and locally total.*

Given Definition 5, precedence is a locally total relation on the set of elements in a sequence (i.e., corresponding to the terminal nodes in the tree structure). The Hasse diagram in Figure 6.4 visualizes the precedence relationship for the sequence "woman saw man with binoculars".

[7] Strictly speaking, the relation is left-locally total (Kayne, 1994). A left-locally total relation is total only on the elements to the left of the relation (e.g., for aRb, R is left-locally total for a).

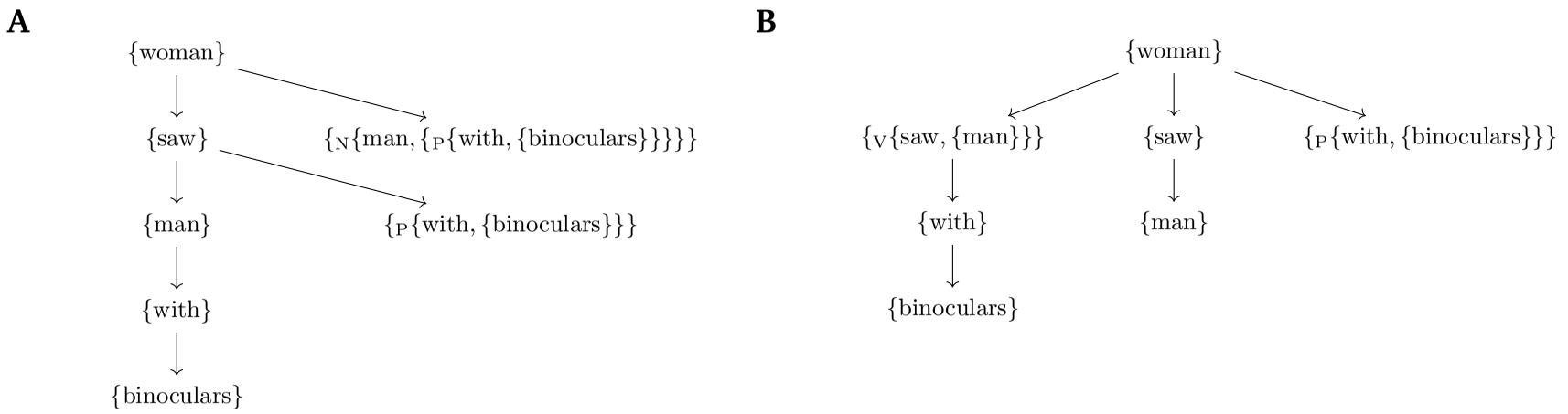


Figure 6.3: Hasse diagram of a subset of the partially ordered magma M, which displays two different structures that map onto the sequence "woman saw man with binoculars". Arrows indicate asymmetric c-command. The subscript at the opening curly bracket of each binary set indicates the label of that set. The diagram in (A) corresponds to the structure in Figure 6.1A, where the man has the binoculars. The diagram in (B) corresponds to the structure in Figure 6.1B, in which the woman is holding the binoculars.

woman ⟶ saw ⟶ man ⟶ with ⟶ binoculars

Figure 6.4: Hasse diagram of an element of the set of sequences S, which displays the sequential structure of the sequence "woman saw man with binoculars". Arrows indicate precedence.

Mapping structures to sequences

Following Kayne (1994), we assume that there exists a rigid mapping between hierarchical structure and linear order, such that only one linear sequence can be derived from a given hierarchical structure. As noted above, asymmetric c-command and precedence are locally total orders on the set of non-terminals and the set of terminals, respectively. Kayne (1994) formalizes the mapping between these two orders in the Linear Correspondence Axiom.

Linear Correspondence Axiom (LCA):
A lexical item α precedes a lexical item β iff
(i) α asymmetrically c-commands β or
(ii) an XP dominating α asymmetrically c-commands β

Definition 6. *We adopt the LCA as a surjective function $f : M \to S$, defining f for a pair of lexical items $\alpha, \beta \in m \in M$, which holds for all elements of the sequence by induction:*

α **asymmetrically c-commands** β

if $\alpha \gg \beta$ *and* $\beta \not\gg \alpha$, *or if*

$\alpha \leq \hat{m}$ *and* $\hat{m} \gg \beta$ *and* $\beta \not\gg \hat{m}$, *then*

$\alpha \prec \beta \in f(m)$

α **precedes** β

the mother of α **asymmetrically c-commands** β

In short, Definition 6 states that a word α precedes a word β if it asymmetrically c-commands β or if a node dominating α asymmetrically c-commands β. The result of this mapping is a full total ordering of the terminals of the hierarchical structure in question. It is important to note that this mapping can be defined as a proper function because, under the LCA, only one linear sequence can be derived from any given hierarchical structure. Conversely, multiple hierarchical structures can map onto the same linear sequence. For instance, the

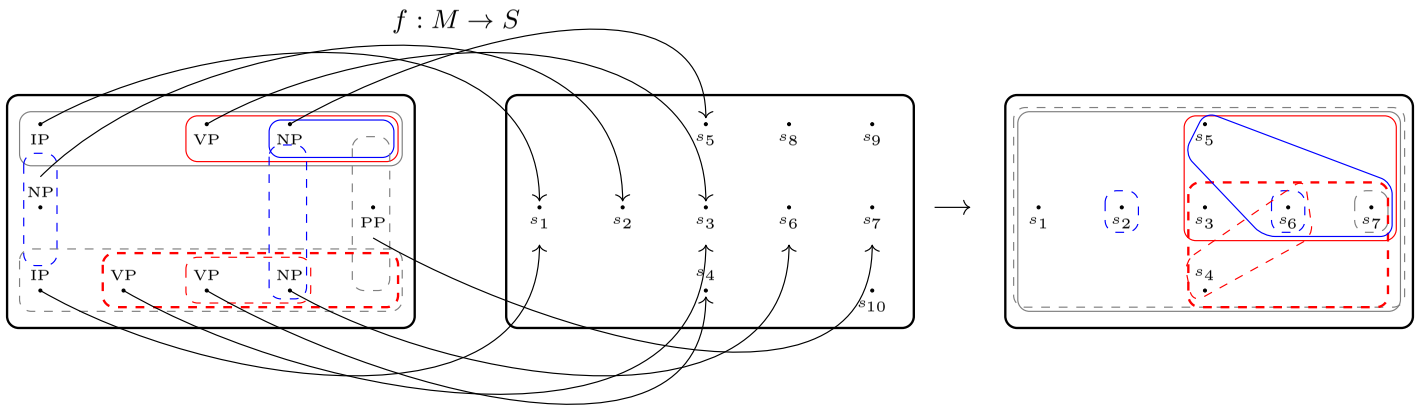
precedence relations that are derived from the asymmetric c-command relations in the two structures in Figure 6.3 are the same, which illustrates the fact that the corresponding sequence is structurally ambiguous.

Ordering sequences via structures

When the sets in M are mapped to sequences in S, these sequences are imbued with grammatical properties. What these grammatical properties are can be understood in terms of the ordering that is carried over from the containment relation in M. Consider Figure 6.5, where the constituent structures in M (left panel) are mapped to the sequences in S (middle panel) via the LCA. By virtue of the containment relation by which the elements of M are ordered, this mapping imposes structure on the set of sequences (right panel) that is not there if only their sequential properties are considered.

If we only consider the sequential properties of the elements in S, a partial ordering already exists. This partial ordering is based on string containment. For example, both "woman with" and "with binoculars" can be said to be contained in the sequence "woman with binoculars". Using the map $f : M \to S$, we impose a restriction on this ordering: for two elements $m_1, m_2 \in M$, $f(m_1) \leq f(m_2)$ iff $m_1 \leq m_2$. That is, two sequences in S are contained in one another only if their constituent structures in M are contained in one another. This imposed ordering restricts the initial ordering by excluding both ungrammatical sequences as well as containment relations that are not the result of a structural relationship. For example, in the middle panel of Figure 6.5, the subsequences s_8, s_9, and s_{10} do not appear in the imposed partial ordering. s_{10} is an ungrammatical sequence and therefore has no structural analog in M. s_8 and s_9 are subsequences of a grammatical sequence, yet they do not correspond to constituents and are therefore not retained in the ordering. Thus, only strings that correspond to constituents are retained in the partial ordering, and this partial ordering is based on constituent containment, as can be seen in the substructure in the right panel of Figure 6.5.

To sum up, we used the binary set formation operator $\oplus$ to generate hierarchical constituent structure. From the resulting structure, whose containment relationships are visualized in Figure 6.2, we derive all c-command relationships (see Figure 6.3). From these c-command relationships we derive a linear sequence with precedence relationships using the LCA. Using the containment relationship in the partially ordered magma (see Figure 6.2), we impose an ordering relation on the resulting set of sequences (see Figure 6.5). The latter is



s_1 : woman saw man with binoculars
s_2 : woman
s_3 : saw man with binoculars
s_4 : saw man
s_5 : man with binoculars
s_6 : man
s_7 : with binoculars
s_8 : man with
s_9 : saw man with
s_{10}: man man with with

Figure 6.5: An ordering relationship is imposed on S via the structure in M. The leftmost panel contains a subset of the partially ordered magma M, with the elements (denoted by their labels) ordered by containment. The labels IP, VP, NP and PP refer to the labels of the constituents. A subset of S is shown in the middle panel, with example sequences presented below the figure. The LCA function $f : M \to S$ maps elements in M to elements in S, thus imposing a structural ordering relation on the sequential elements in S (rightmost panel).

possible because we define the algebraic structure corresponding to the set of structures as a magma, whose combinatorial operator is non-associative. This allows us to generate strongly compositional structure, which is a necessary requirement for any description of (the faculty of) language.

6.3 Hierarchical structure in actions

Having defined and formalized the properties of hierarchical structure in syntax, we will now consider whether action hierarchies are analogous to syntactic hierarchies. Similar to the hierarchical structure underlying sentences, action sequences are thought to be governed by hierarchically organized action plans (Botvinick, 2008; Cooper & Shallice, 2000, 2006; Holloway, 1969; Koechlin & Jubault, 2006; Lashley, 1951; Miller et al., 1960; Rosenbaum et al., 2007).[8] This structural analogy between linguistic syntax and actions has received considerable attention from several corners of cognitive science (Boeckx & Fujita, 2014; Fadiga et al., 2009; Jackendoff, 2007, 2009; Moro, 2014a, 2014b; Stout & Chaminade, 2009), in which the hierarchical structure of actions is thought to be generated by an 'action syntax' (Fitch & Martins, 2014; Fujita, 2014; Maffongelli et al., 2019; Pulvermüller, 2014).

The idea is often illustrated using the example of tea- or coffee-making as a goal-directed behavioral routine (Cooper & Shallice, 2000; Fitch & Martins, 2014; Fischmeister et al., 2017; Humphreys & Forde, 1998; Jackendoff, 2007, 2009; Kuperberg, 2020). A multi-step action such as tea-making can be decomposed into discrete subsequences of actions, which in turn can be decomposed in sub-subsequences, and so on. Figure 6.6 shows a visual representation of the hierarchical part-whole structure of 'making tea'. The highest level in the hierarchy represents the complex, temporally extended and goal-directed action, middle levels represent short-term, less complex subactions with their own subgoals, and the lowest level (terminal nodes) contains atomic actions with immediate subgoals. Decomposing complex actions into these embedded subsequences is theoretically and empirically warranted because the subsequences may be used in different tasks, because they are sometimes omitted, repeated, or substituted as a whole, and because they all have their own subgoal, which

[8] Note that we are concerned with describing the structure of actions rather than with describing how action sequences come about in a processing system (Badre, 2008; Tettamanti & Moro, 2012). The latter question belongs to the study of motor control, which is also hierarchically organized but which has different properties: motor control is based on causal relations ('processing' hierarchy), while actions should be described in terms of part-whole relations ('representational' hierarchy; see Uithol et al., 2012 for discussion).

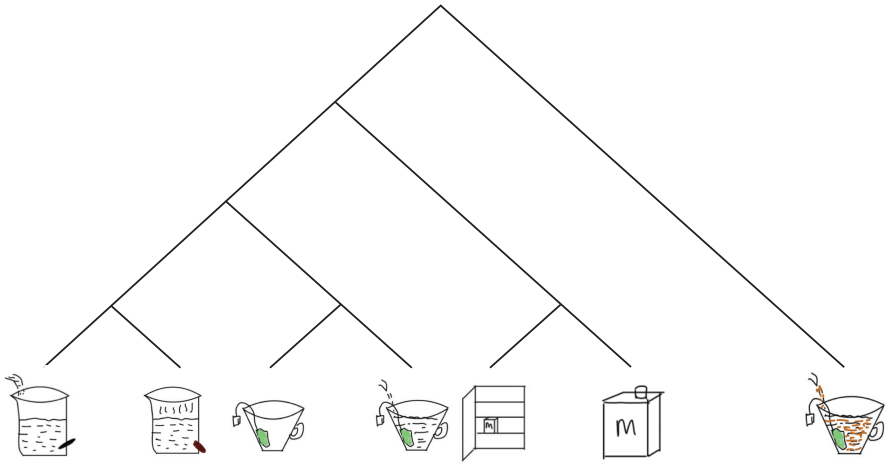


Figure 6.6: An example of a hierarchical decomposition of an action sequence, such as making tea. The terminal nodes correspond to the atomic actions in Figure 6.7.

must be fulfilled in order to achieve the overarching goal (Cooper & Shallice, 2000, 2006; Humphreys & Forde, 1998; Lashley, 1951; Norman, 1981; Reason, 1979; Rosenbaum et al., 2007; Schwartz, 2006).

6.3.1 Formalizing action structure (1)

The following sections describe the structure of actions using the same mathematical formalism used to describe language in Section 6.2.2. We first show that this formalism is inadequate for describing actions. Section 6.3.2 then proposes an alternative way to describe the structure of action sequences.

Definition 7. $(M, \oplus, \varnothing)$ *is a unital, commutative magma generated from A, where:*

1. *A is a set containing atomic actions, such as the examples presented in Figure 6.7.*
2. *M is a set of elements that are generated from A , with* $A \subset M$.
3. $\oplus$ *is a binary set formation operation, which is commutative, non-associative and closed.*
4. $\varnothing$ *is the identity element.*

By defining the same binary relationship as used in Definition 2, we derive a partially ordered magma in which the actions and action sets are partially ordered by containment. A subset of this partially ordered magma is visualized in

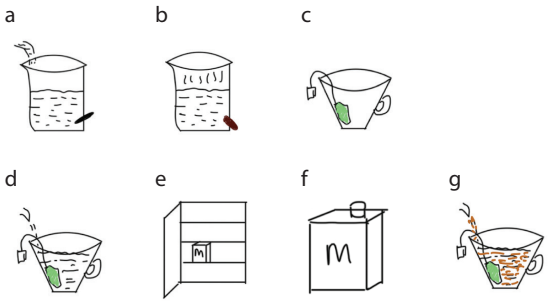


Figure 6.7: Atomic actions for tea-making. (a) fill kettle with water. (b) turn on kettle. (c) put teabag in cup. (d) pour hot water into cup. (e) open fridge. (f) grab milk. (g) pour milk into cup. In the context of Definition 7, $A = \{a, b, c, d, e, f, g\}$.

the Hasse diagram in Figure 6.8, which displays the containment relationship for two structures that map onto the same action sequence for making tea. Note that the two topmost action structures are derived in a different way. This figure illustrates a crucial point about the (ir)relevance of hierarchical structure in the interpretation of action sequences. That is, because the $\oplus$ operator is non-associative, the order in which actions are combined using $\oplus$ affects the structure that is generated. Therefore, if we were to interpret these structures in a strongly compositional way, we would have to conclude that they correspond to different actions. This is clearly an undesirable conclusion, because the two structures correspond to one and the same action sequence. In other words, adopting a non-associative combinatorial operator for generating action structures makes the model too strong: it will differentiate two action structures that should not be distinguished because they map onto the same action sequence and thus achieve the same goal in effectively the same way.

Compositionality in language and action

The fact that a strongly compositional formal model does not accurately describe actions indicates that the action system is not strongly compositional. If the action system is weakly compositional instead, it follows that one action sequence cannot be associated with multiple hierarchical structures. This prediction is borne out: structurally ambiguous actions, where one action sequence is associated with more than one structural representation and therefore more than one goal, do not seem to exist. This does not mean that actions cannot be ambiguous. Any given action may be characterized in terms of different goals, but these different goals are not a function of a decomposition of the action sequence in terms of hierarchically organized 'action constituents'. Whether the action's goal

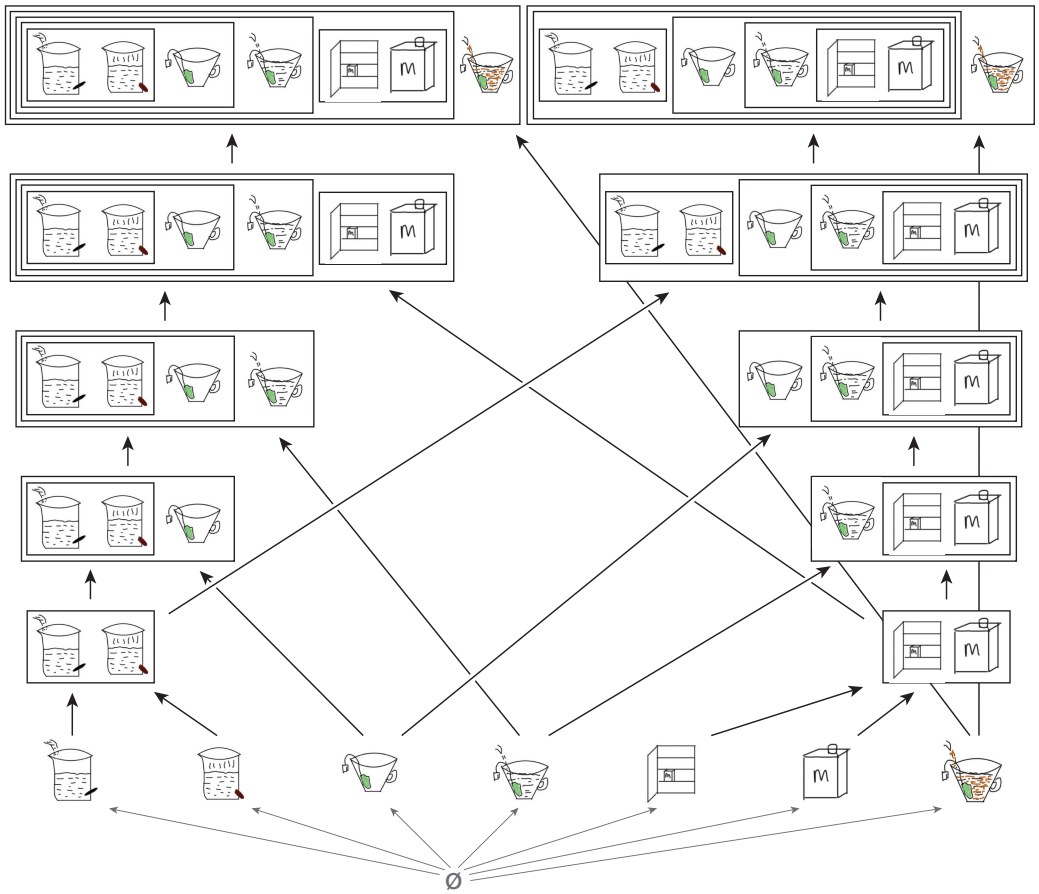


Figure 6.8: Hasse diagram of a partially ordered magma, which displays two different structures that map onto the same action sequence for making tea. The boxes around action combinations represent binary sets, and the arrows indicate direct containment.

is achieved depends on the temporal order of its constituent actions, not on their hierarchical organization.

6.3.2 Formalizing action structure (2)

In the previous section, we showed that when the language formalism is applied to actions, it appears to be too strong: it makes a distinction which should not be made. The model is also too weak: actions and action plans are structured by temporal (precedence) relations, but the model does not take temporal order into account. In the current section, we therefore propose an alternative way to describe the structure of actions. The operator used to generate action structures must meet at least two requirements. First, it must generate sequential structure, because actions are temporally ordered. Second, it must not be non-associative, because actions are not strongly compositional.

A set of partitioned sequences

We have already defined the set of sequences as S (see Definition 4). The elements in each sequence are in a total, transitive, and antisymmetric ordering (see Box 2). This set is partitioned according to the following criterion: sequences are deemed equivalent if they bring about a particular change in the environment (i.e., they achieve the same 'goal'). All equivalent sequences are part of a single equivalence class, whose label corresponds to the goal achieved by the sequences in it.

Definition 8. *Given the set S, a partition of S contains a set G, and for each $g \in G$, a non-empty subset $S_g \subseteq S$ exists, such that:*

the set of sequences is the union of all parts

$$S = \bigcup_{g \in G} S_g \quad \textit{and if } g \neq h, \; S_g \cap S_h = \varnothing$$

the parts do not overlap

Here, we take $g \in G$ as the set of part labels (i.e., the labels given to each element in the partition). Because all sequences in a given part are equivalent, we call every element $s \in S_g$ a representative sequence of that part S_g.

The partitioning of S yields a set of part labels that correspond to the set of goals they accomplish. These goals can be interpreted as abstractions over action sequences that have something in common, namely the change they bring about in the environment (see e.g., Cooper & Shallice, 2000).

Generating structured sequences

Definition 9. *We define action structure as $(G, \otimes, *, \varnothing)$, where:*

1. *The elements of G are part labels (see Definition 8) corresponding to action sequences that achieve a particular goal.*
2. *$\otimes$ and $*$ are two sequence-building operators that generate the elements of G.*
3. *$\varnothing$ is the identity element.*

Note that we include the set of atomic actions in G, because atomic actions achieve a particular change in the environment and thus have their own subgoal. Therefore, an atomic action is simply an equivalence class with only one element.

A goal can often be achieved in several ways. For example, given the actions in Figure 6.7, the goal 'make black tea' corresponds to the part S_b, where $S_b = \{(a,b,c,d),(a,c,b,d),(c,a,b,d)\}$. Here, a ('fill kettle with water') must precede b ('turn on kettle'), which in turn must precede d ('pour hot water into cup'), so the relative temporal ordering of (a,b,d) is fixed. However, the position of action c ('put teabag in cup') within this action sequence should be specified only in relation to d; it can be placed at any position before d within (a,b,d), thus yielding three action sequences. In other words, for a given goal to be achieved, the temporal ordering of some actions must be specified, whereas it need not be specified for other actions. We achieve this combination of the requirement of strict temporal ordering with temporal flexibility via the use of two sequence-building operators.

Definition 10. *$*$ is a sequence-building operation $* : G \times G \rightarrow G$. Let $a,b,c \in G$ be three part labels, and let $s_a \in S_a$, $s_b \in S_b$, $s_c \in S_c$ be three representative sequences, where $s_a = (a_1, a_2, \ldots a_n), s_b = (b_1, b_2, \ldots b_m)$.*

$*$ concatenates two sequences

$$s_a * s_b = (a_1, a_2, \ldots a_n, b_1, b_2, \ldots b_m) \quad where$$

$$\forall s_a, s_b, s_c \in S, \; (s_a * s_b) * s_c = s_a * (s_b * s_c)$$

$*$ is associative

Definition 11. *$\otimes$ is a sequence-building operation $\otimes : G \times G \rightarrow G$. Take s_a and s_b as defined in Definition 10. Then $s_a \otimes s_b = \{(c_1, c_2, \ldots c_{m+n})\}$, such that*

$\otimes$ retains precedence relations in its input

$$c_i = \begin{cases} a_j & \text{if } i > 1 \text{ and } a_{j-1} \in (c_1 \ldots c_{i-1}) \\ b_j & \text{if } i > 1 \text{ and } b_{j-1} \in (c_1 \ldots c_{i-1}) \\ a_1 \text{ or } b_1 & otherwise \end{cases}$$

The operator $*$ is simple concatenation. This operator is required because the temporal ordering of some actions must be specified. For example, if $h \in G$ represents the (sub)goal 'obtain hot water', then we must define $S_h = a * b$ due to the requirement that the kettle should be filled with water (action a) *before* it is turned on (action b). The temporal precedence relationship between these two

actions requires an operator that yields strict sequential orders. Clearly, concatenation is not commutative: the sequence generated by $s_a * s_b$ is different from the output of $s_b * s_a$. Moreover, because $*$ generates sequences whose only relationship is precedence, it is associative: $(s_a * s_b) * s_c = s_a * (s_b * s_c)$. In sum, $(G, *)$ forms a monoid, which is an algebraic structure consisting of a set equipped with an operation that is closed, unital and associative (see Box 1).

The operator $\otimes$ generates sets of sequences whose orders vary, with the only constraint that the relative ordering within its arguments is retained. For example, for two sequences $s_a = (a, b, d), s_b = (c)$, $s_a \otimes s_b = \{(a, b, d, c), (a, b, c, d), (a, c, b, d), (c, a, b, d)\}$. In each of the sequences generated by $s_a \otimes s_b$, a precedes b, which precedes d. So while $\otimes$ allows for flexibility in terms of the order of the actions in the sequences, the flexibility is constrained by the sequential properties of s_a, whose precedence relations must be retained in the output of $s_a \otimes s_b$. Because $\otimes$ is a sequence-building operator that is constrained only by the sequential properties of its input (i.e., the ordering *within* its input arguments), $\otimes$ is associative. But as it does not specify the ordering *among* its input arguments, $\otimes$ is also commutative. $(G, \otimes)$ therefore forms a commutative monoid (see Box 1).

The two notions of precedence and flexibility are combined in $(G, \otimes, *)$, which is an algebraic structure called a trace monoid (also called partially commutative monoid; see Box 1). A trace monoid is a monoid of traces, which are sets of sequences that form equivalence classes (Mazurkiewicz, 1995). In $(G, \otimes, *)$, the traces contain equivalent sequences generated by $\otimes$ and $*$. In a trace monoid, two sequences are equivalent if they only differ in the order of a pair of elements for which an independency relation is defined.[9] These independent elements are allowed to commute in the sequences of the equivalence class.[10] Consider the independency $I = \{(b, c), (c, b)\}$, which holds that the actions b and c are allowed to commute; no precedence relation between them is specified. Given $(b, c) \in I$, we say that two action sequences are equivalent if they differ only in the ordering of b and c. The trace monoid is then said to contain a trace where $acb \sim abc$.

[9] Independency relations are symmetric (i.e., if (a, b) is present, then so is (b, a)) and irreflexive (i.e., there are no relations of type (a, a)), and can be extended to relations between sequences (see Mazurkiewicz, 1995).

[10] Commutativity in the general sense is slightly different from the way it is used in the context of traces. In the general sense (as used in Box 1), it refers to an operation which produces the *same* output if the order of the operands is changed, such as in $a \otimes b = b \otimes a$. In the context of a trace monoid, the notion of sameness is replaced by equivalence, where $a \otimes b = \{ab, ba\}$, and $ab \neq ba$ but $ab \sim ba$.

To sum up, defining the trace monoid $(G, \otimes, *)$ allows us to achieve simultaneously temporal precedence *and* temporal flexibility. The operator $*$ is required to build sequences where temporal precedence is necessary (e.g., 'grab milk' must precede 'pour milk into cup'), and $\otimes$ is used to generate action sequences whose temporal relationship is not specified (e.g., the action 'grab milk' can precede or follow 'put teabag in cup'). The combined use of $*$ and $\otimes$ leads to equivalence classes of sequences that contain a mixing of intermediate goals that are temporally independent of other intermediate goals. The mixing procedure introduced by $\otimes$ might destroy immediate precedence (or temporal adjacency) relationships in the output of $*$, but this is unproblematic: while it makes sense to let 'open fridge' be directly followed by 'grab milk', this is not necessary. One could open the fridge, perform all other tea-making preparations, and then grab the milk.

Hierarchical relations between action sequences

While the output of the two associative operators are sets of sequences, these sequences contain underlying structure if we take their derivational history into account (cf. 'configurational properties' in Miller et al., 1960). For instance, given the atomic actions in Figure 6.7, the sequence (a, b, c, d) corresponds to the goal of making black tea. By itself, this sequence does not provide a lot of information about the precedence relations that might hold for the complex action; it could in principle have been generated via $(((a \otimes b) \otimes c) \otimes d)$. Such information can be inferred only if additional action sequences are observed that achieve the same goal (see Box 3). Knowing how the action sequence was derived allows us to specify the temporal constraints to which it must adhere, which in turn provides information about the causal structure of the action plan. Thus, by deriving $[(a, b, c, d)] \in (a * b) \otimes (c * d)$, we make temporal precedence relations concrete: a must precede b, and c must precede d.

The derivational history of the sequence provides information about which other sequences are also possible. From observing only (a, b, c, d), it would be impossible to know whether (c, a, b, d) is also a fine sequence. However, that knowledge can be deduced if we know how the sequence was derived, because $(a * b) \otimes (c * d)$ also generates the sequence (c, a, b, d). The derivational history thus provides information that is not present in the temporal structure of the sequence, including information about the relationship between the output sequence and its subsequences. In (c, a, b, d), it is still the case that c precedes d, even though they are not adjacent anymore. But by taking into account the derivational steps leading to (c, a, b, d), we can specify a hierarchical relation-

ship between (c,d) and (c,a,b,d), which states that $(c,d) \in (c,a,b,d)$ because (c,a,b,d) was generated via $(a,b) \otimes (c,d)$. This relationship holds even though (c,d) is not a subsequence of (c,a,b,d). Action sequences can be seen as hierarchical sequences, which are sequences with a derivational history that specifies how they relate to the action sequences from which they are derived. This allows us to go beyond the sequential structure of actions in a system that is still weakly compositional.

Box 3. Inferring plans from action sequences

In order to achieve a given goal, the relative order of some related actions must be specified, whereas that of some unrelated actions can be left undefined. Given a set of observed action sequences that successfully reach the same goal, the abstract plan to reach that goal can be extracted via the intersection of the sets of binary relations representing the sequences.

As a simple illustration, consider a sequence $x = (a,b,c,d)$ consisting of non-repeating atomic actions. The precedence relations for x are $a \prec b \prec c \prec d$. The sequence can be represented as a set of binary relations. If we take these binary relations to represent precedence, x will be represented as $\{(a,b),(b,c),(c,d),(a,d),(a,c),(b,d)\}$. By observing only x, it is not immediately clear which of these elements are dependent and which are not. However, observing the sequence $y = (c,a,b,d)$ (represented as $y = \{(c,a),(a,b),(b,d),(c,d),(c,b),(a,d)\}$), which achieves the same goal, provides more information. The plan to reach the goal is represented by the intersection of the sets of binary relations:

$$x \cap y = \{(a,b),(c,d),(b,d)\}$$

This intersection corresponds to the plan of making black tea (see Figure 6.7). Notice how this partial order is compatible with the previously unseen sequence (a,c,b,d), which reaches the same goal successfully as well.

6.4 Language vs. action

6.4.1 A formal comparison

In the previous sections we described the properties of hierarchical linguistic structure (generated by $\oplus$) using a magma. When this formalism was applied to

actions, it appeared to be too strong, deriving multiple 'interpretations' from unambiguous action sequences, and too weak, as it does not generate temporally ordered structures. To overcome these limitations, our alternative formalism described action structures as a trace monoid (generated by non-commutative $*$ and commutative $\otimes$). A crucial difference between these algebraic structures (see Box 1) is that the operation associated with magmas is non-associative, whereas that associated with monoids is associative. As a consequence, the structure generated by $\oplus$ and represented in the magma is strongly compositional: the constituent structure of the input to $\oplus$ is retained in its output. This is important because both syntactic operations and semantic interpretation are structure-dependent. If the internal structure of each combination would be lost, syntactic rules could not target constituents. Moreover, meaning could not be derived from constituent structure, and sentences could not be structurally ambiguous; the system would generate only one output for ((deep $\oplus$ blue) $\oplus$ sea) and (deep $\oplus$ (blue $\oplus$ sea)). In contrast, the action structure generated by the associative operators $*$ and $\otimes$ and represented in the trace monoid is weakly compositional. This weakly compositional, order-sensitive model can account for the relevant properties of action structures.

6.4.2 The nature of structure

Our formal characterizations of language and action show that their structural representations are different, in particular with respect to the relevance of constituency (for a similar conclusion from the neuroimaging literature, see Papitto et al., 2020 and Zaccarella et al., 2021). The same conclusion is reached when we compare language and action in terms of the properties of syntactic structure discussed in Section 6.2.1.

1. **Unbounded**. It has been argued that the combinatorial operation involved in building syntactic structures evolved from pre-existing systems for tool use, also called Action Merge (Fujita, 2014, 2017). This operation is thought to apply recursively (Fujita, 2017; Pulvermüller, 2014; Stout & Chaminade, 2009), even though Action Merge is bounded (Fujita, 2014). A distinctive feature of recursively generated hierarchical output is self-similarity across levels: recursively generated structures are characterized by self-embedding of tokens of the same type (Martins, 2012). One approach towards determining whether actions are recursively generated is therefore to examine whether their structures are self-similar. However,

that requires knowing what the types are. Consider the structure in Figure 6.6. One could combine 'open fridge' and 'grab milk' into an action constituent, which could be labeled 'get milk'. Here, it is unclear whether 'get milk' is of the same type as 'grab milk'. Moreover, it seems plausible that the action 'pour hot water into cup' is similar to 'pour milk into cup', but that is because the tokens are similar (both involve pouring), not necessarily because their types are. To determine whether action structures are recursively generated, a theoretical specification of the types of actions is needed.

Our primary goal is to evaluate the claim that the hierarchical structures found in language and action are analogous. The validity of this claim rests on positive evidence that actions, like language, are recursively generated. In the absence of such evidence (e.g., in the form of self-similar hierarchy), it is premature to conclude that actions are structurally analogous to language.

2. **Endocentric**. Some of the hierarchical representations of actions that are proposed in the literature contain action constituents with one key element, or 'head', which performs the core of the action and determines its (end)goal (e.g., Jackendoff, 2007, 2009, 2011; Fischmeister et al., 2017). While this makes the structures 'headed', it does not make them endocentric. That is, it seems that this head merely serves to describe the main action of the action sequence, rather than to provide a label for the constituent it is dominated by. In Figure 6.6, for instance, the action constituent formed by the combination of 'open fridge' and 'grab milk' is not a type of either of these actions. In line with the idea that endocentricity is unique to language (Boeckx, 2009; Hornstein, 2009), action hierarchies seem to be exocentric.

   A plausible reason for the difficulty in assigning labels to action constituents is that actions do not have clear conceptual units, such as words (Moro, 2014a; Berwick, Okanoya, et al., 2011), and that groups of actions do not obligatorily fall into a closed set of distinguishable categories, such as NP or VP (Jackendoff & Pinker, 2005). Without these categories, groupings of actions into constituents cannot be labeled or 'syntactically' named, which means that there are no grammatical constraints on how the resulting constituents can be used in further combinations.

3. **Unordered**. Representations of actions are intimately tied to the physical environment in which the actions are performed (Graves, 1994; Kuper-

berg, 2020; Moro, 2014a; Zaccarella et al., 2021). As such, they are not order-independent: some subactions must precede others in order for the action to achieve its goal (Fitch & Martins, 2014), and indeed, the output of Action Merge is inherently ordered (Fujita, 2014).[11] Comparing this to language, we see that the externalization of spoken language is also sequential, but that sequential order does not play a role in the representation of syntactic relations, which are invariably structure-dependent.

It has been proposed that closely related actions, which can be separated by arbitrarily many 'embedded' actions (e.g., [open door [switch on light [brush teeth] switch off light] close door]), are similar to long-distance dependencies in language (Pulvermüller & Fadiga, 2010; Pulvermüller, 2014). This analogy is incorrect, however, because long-distance dependencies are related to the hierarchical organization of constituent structure. These action dependencies, instead, have serial and temporal properties: you cannot close a door *before* having opened it (Dominey et al., 2003; Moro, 2015; Zaccarella et al., 2021). If they were truly hierarchical, the embedded action would be expected to adhere to structural restrictions on its distribution, which would be the case if the embedding of [brush teeth] at a different position, like in [open door [brush teeth] [switch on light switch off light] close door], were not allowed. Moreover, if the dependency between 'switch on light' and 'switch off light' were hierarchical, it should not be affected by linearly or temporally intervening actions, so whatever happens during 'brush teeth' should not be able to affect the action 'switch off light'. As neither appears to be the case, it is more appropriate to label the dependency between two actions temporal (or causal) rather than hierarchical (Moro, 2014b, 2015). Indeed, actions and events can be understood in terms of temporal (and causal) structure (McRae et al., 2019; Zacks & Tversky, 2001), and oddly ordered complex actions, which are thought of as ungrammatical actions (e.g., Maffongelli et al., 2019), reflect the violation of 'temporal rules' rather than phrase-structure rules (Zaccarella et al., 2021).

[11] Even under an analysis in which *immediate* precedence plays a role in syntax (as in Kayne, 2011), the crucial difference between language and actions remains: if two linguistic objects α and β are not adjacent in their base-generated position (i.e., they do not form the ordered pair $\langle \alpha, \beta \rangle$), their relationship is defined as a relationship that refers to the (hierarchical) constituents they are contained in, not as a relationship that refers to their linear or temporal order. There is no such constraint in actions, where some actions must precede (distant) others, regardless of the relationship between the action constituents in which they are contained.

A plausible reason for the observation that none of the properties of hierarchy in syntax are found in actions is that the analogy between language and action is not to be found in syntactic structure, but rather in conceptual structure (Jackendoff, 2007; Zaccarella et al., 2021). An important difference is that syntax is computationally autonomous, having its own principles and properties that cannot be reduced to other factors, such as meaning (Adger, 2018; Berwick, 2018; Chomsky, 1957). The application of these principles is constrained by economy conditions (e.g., locality, minimality; see Collins, 2001), but not by whether they generate interpretable output. Therefore, in language there is an independent notion of grammaticality: sentences are ungrammatical if their structures cannot be generated by the rules of syntax, or if they violate conditions on these rules. One way to illustrate this is by means of interpretable but nevertheless ungrammatical sentences. A sentence such as "which boy did they meet the girl who insulted?" is ungrammatical but can be interpreted (i.e., corresponding to the logical statement "for which x, x a boy, did they meet the girl who insulted x?"). Its deviance is due to the violation of a purely formal (locality) principle constraining the grammar, which is unrelated to its semantic interpretability. Conversely, the sentence "colorless green ideas sleep furiously" is semantically odd, yet fully grammatical, showing that grammaticality does not boil down to meaningfulness or interpretability.

In contrast, the validity of action sequences seems related to their coherence, in terms of both logical consistency and environmental appropriateness. It has been suggested that a complex action is 'ungrammatical' or 'ill-formed' if its subparts are ordered in such a way that the action's overall goal cannot be achieved (Jackendoff, 2007; Maffongelli et al., 2019). The 'grammaticality' of an action is thus intimately tied to the fulfillment of its goal, showing that the notion 'ungrammatical' is very different for action sequences and sentences. On this interpretation, an 'ungrammatical' action is similar to a sentence which does not convey the intended meaning, either because it is logically incoherent or because it is situationally inappropriate. The action equivalent of a logically incoherent sentence could be an action sequence in which a coffee grinder is turned on before the coffee beans are added. This is logically incoherent because it violates causality principles of the physical environment. An action like turning off the light when walking into your office during nighttime, instead, does not violate such constraints, but it would be situationally inappropriate because it would preclude you from seeing anything.

Because there is no autonomous action syntax, there is no independent notion of grammaticality, devoid of goal-dependent meaning. As a result, it is unclear how to evaluate whether a given structural decomposition of complex actions into constituents is veridical unless we know the goal or general conceptual content of the action (Berwick & Chomsky, 2017; Jackendoff, 2007). It seems that the decomposition of an action sequence into a hierarchical tree structure only works to the extent that the subactions are meaningful or coherent (i.e., represent subgoals).

6.4.3 Levels of abstraction

The difference between language and actions in terms of their dependence on hierarchical and sequential structure can be captured quite naturally under the distinction between *hierarchical sequences* and *hierarchical sets*, a terminological contrast adopted by Fitch and Martins (2014) to distinguish possible interpretations of the term hierarchy.[12] Fitch and Martins (2014) describe hierarchical sets as structures that specify the superior/inferior relation between their elements (i.e., specifying containment), but whose elements are unordered at any given level. Hierarchical sequences, instead, are hierarchical structures in which sequential order matters: at least some elements at any given level represent a sequence rather than a set.

We will argue that language needs to be described in terms of both hierarchical sequences and hierarchical sets, but that actions can be described as hierarchical sequences only (see Figure 6.9). Regarding language, the mapping between hierarchically organized constituent structure and sequentially ordered sentences represents the level of hierarchical sequences (i.e., the interface between hierarchical structure and linear order; see middle panel, top row in Figure 6.9). In this hierarchically structured sequence, we can describe a speaker's knowledge about linearized properties of their language (e.g., word order, morphosyntax), such as whether heads precede or follow their dependents (e.g., English vs. Japanese). Hierarchical sets are abstractions from these hierarchical sequences, which are not realized in the physical properties of the linguistic signal (left panel, top row in Figure 6.9). This level is explanatorily relevant for syntactic theory because it naturally captures the properties of syntax described in Section 6.2.1: hierarchical sets generated by Merge are unbounded, endocentric, and unordered

[12] A similar distinction is emphasized by Tettamanti and Moro (2012), who discuss the different meanings of hierarchical organization in terms of *sequential* vs. *internal* hierarchy, describing the computation of *sequential* hierarchical information (externalized) and the computation of *non-linear* hierarchical relations (mind-internal), respectively.

(Lasnik, 2000). Therefore, at this level, we can account for both structural relations within languages and structural generalizations between languages (e.g., the head-dependent relations in English and Japanese are identical at the level of hierarchical sets).

As we noted at the end of Section 6.3.2, actions might be seen as hierarchically generated yet ordered sequences of events (middle panel, bottom row of Figure 6.9). However, the structural properties of action sequences cannot be described in a way which is completely detached from the physical instantiation of the action sequence, most clearly because (the representations of) action sequences contain information about temporal order. Because the properties of syntactic structure are not found in actions, it is not necessary to postulate hierarchical sets as an explanatorily relevant level of abstraction for actions.

The distinction between hierarchical sets and sequences is useful in explaining why it has been found that the brain areas involved in language processing (in particular, BA44 in the left inferior frontal gyrus) are also activated in response to tasks involving hierarchically organized actions (Higuchi et al., 2009; Koechlin & Jubault, 2006). At first thought, these results support the idea that there is a supramodal hierarchical processor in the brain, which processes the hierarchical structures of cognitive systems such as language and action (Fadiga et al., 2009; Fazio et al., 2009; Fiebach & Schubotz, 2006; Higuchi et al., 2009; Jeon, 2014; Koechlin & Jubault, 2006; Tettamanti & Weniger, 2006). Crucially, however, instead of describing complex actions in terms of non-linear relations defined over hierarchical structures, these accounts refer to the processing of structured sequences that were hierarchically generated (for a related discussion, see Martins et al., 2019 and Zaccarella et al., 2021). The overlapping activation patterns for language and action might therefore point not to shared brain regions processing hierarchical, non-linear relations (operating over hierarchical sets), but rather to shared brain regions implicated in the linearization of hierarchically structured information (i.e., hierarchical sequences; see also Boeckx et al., 2014; Matchin & Hickok, 2020; Uddén & Bahlmann, 2012). We believe that a fruitful avenue for further investigation into the relationship between language and action concerns the externalization of hierarchically organized information into structured sequences rather than the (generation of the) hierarchical structure itself. The overlap between language and action then has to do with the fact that, externally, both are structured sequences, even though their internal structures are quite different.

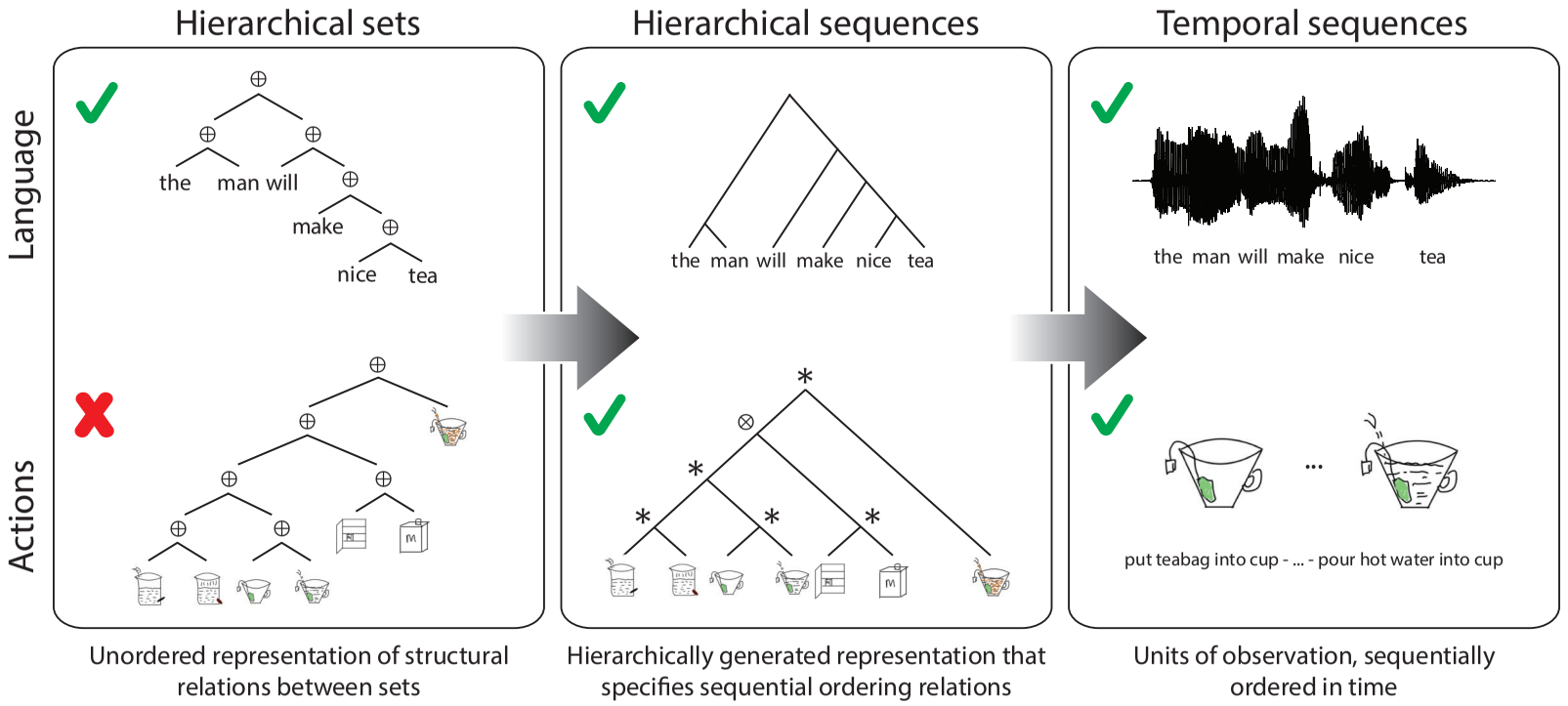


Figure 6.9: Three levels to describe hierarchically structured sequential information. The levels become increasingly concrete, from what is abstractly represented (on the left) to what is physically observed (on the right). The symbol at each node in the hierarchical structures indicates which operator was used to combine the elements, thus representing the derivational history of the sets in language (left) or the sequences in actions (middle).

6.5 Conclusions

In response to the claim that language and action are analogous because they are both organized hierarchically, we argued in this chapter that the formal properties of structure in both domains are fundamentally different. Our main argument is that the language system can embody strong compositionality, as both syntactic rules and semantic interpretation are structure-dependent. Structural analyses in language are thus concerned with non-terminal nodes in the hierarchical structure of syntax. Actions, instead, are weakly compositional: regularities in action structures are dependent on the temporal order of the atomic actions, not on their hierarchical organization into action goals. Analyses of actions are thus concerned with terminal nodes in the action hierarchy. Based on this difference, we argue that the structure of syntax is best described as a system of hierarchical sets, whereas action structures can be described as hierarchical sequences.

In order to formally capture the strong compositionality of language, we described the algebraic structure corresponding to the ordered set of hierarchical structures in language as a magma, whose non-associative combinatorial operator was defined as binary set formation. This set-based formalism integrates the three properties of syntactic structure (i.e., unboundedness, endocentricity, and unorderedness) with the description of syntax as a system of hierarchical sets and the fact that language exhibits strong compositionality. When this model was applied to actions, it appeared to be both too strong (i.e., it makes structural distinctions which should not be made) and too weak (i.e., it does not capture the importance of temporal precedence). We therefore proposed an alternative model for actions, which used two sequence-building operators that organize actions by sequential relations. This yielded an ordered set of action structures that could be described as a trace monoid. The associativity of the two operators formalizes the idea that actions exhibit a weaker form of compositionality, and aligns well with our argument that actions are best described in terms of hierarchical sequences. In sum, the formal tools needed to describe language are fundamentally different from those required to the describe the action system. We believe that this result has important implications not only for comparative cognitive science but also for cognitive neuroscience, as it points to differences in the ways in which hierarchies are represented in the brain.

7 | General discussion

This thesis started with the view that generalizations about *structure* and observations about *use* are different reflections of one and the same cognitive system. Assuming a single cognitive system for language yields the prediction that the type of structure postulated by linguists will be measurable in the type of data acquired by neuro- and psycholinguists. As an empirical test of this prediction, this thesis investigated the role of syntactic hierarchy in language, both in terms of its structural properties as well as in the way it is reflected in the behavior and brain activity of language users.

7.1 A discussion of the main findings

In five chapters, I aimed to shed light on different aspects of the role of hierarchy in language, asking questions such as: are people biased to interpret language hierarchically or linearly? Are artificial neural network models well-equipped to learn to behave like humans? How does the brain infer hierarchical structure during language comprehension? And is the hierarchical structure of language found in other domains of cognition? The following paragraphs provide a brief summary and discussion of each chapter.

In **Chapter 2**, I investigated whether people interpret ambiguous noun phrases such as *second blue ball* as a hierarchical structure or as a linear string. The results of two behavioral experiments were very clear: participants invariably interpreted *second blue ball* hierarchically. This finding directly falsifies the claim that hierarchy is not fundamental in language use (Frank et al., 2012). If language use were fundamentally sequential, we would expect people to consistently prefer the linear over the hierarchical option when the input is compatible with both. Note that all trials in the behavioral experiments were completely ambiguous; both interpretations were always available. The finding that people unanimously interpret *second blue ball* in line with its hierarchical structure therefore shows not just that they interpret language hierarchically if that is the only available option that yields a coherent interpretation, but that they are biased to do so. The strength of this bias is nicely illustrated by participants'

responses during the post-experiment debriefing, in which we asked them about the goal of the experiment. They all expressed their surprise about the simplicity of the task, and reported not even being aware of the ambiguity of *second blue ball*.

We then tested whether an artificial neural network (ANN) model could reproduce the behavior of the human participants if it was trained and tested on a similar task. To test whether it represented the phrases in a human-like way, we measured its performance in a range of train-test regimes. First, we established that it could be trained to give both linear and hierarchical answers if it was trained on unambiguously linear and hierarchical training data, respectively. We then tested the model's behavior after ambiguous training data, which were consistent with two generalizations, one of them being the linear generalization, the other one being the hierarchical generalization. As the test phase consisted of trials for which these two generalizations make different predictions (i.e., the divergent trials from the behavioral experiment), the model's answers on these trials reveal its inductive biases in this setup. This ambiguous train-test regime indirectly models natural language acquisition, in which the input is also compatible with multiple generalizations. The observation that humans consistently arrive at the same generalizations, despite the fact that these are underdetermined by the input, reflects the poverty-of-the-stimulus problem (Chomsky, 1980). In our simplified poverty-of-the-stimulus scenario, the model only gave linear answers on divergent test trials, in stark contrast to the human participants.

To be fair, the linguistic input humans receive is not ambiguous in the same way. For one thing, it contains cues to hierarchical structure in the form of n-gram statistics (Saffran, 2001; Takahashi & Lidz, 2007; Thompson & Newport, 2007). In a subsequent analysis we therefore added hierarchical information to the data by presenting an unbalanced ratio between ambiguous and unambiguously hierarchical trials, while still making sure that the hierarchical interpretation was the only generalization fully compatible with the data. Here, we found that even when only 10% of the training data are ambiguous, and the remaining 90% indicate that the linear interpretation is incorrect, the model still gives many linear responses on divergent test trials. The model's behavior is thus quite different from the way humans respond in similar scenarios (Culbertson & Adger, 2014; Ferrigno et al., 2020; Martin et al., 2020; Morgan & Ferreira, 2021). This mismatch likely arises from a difference in their inductive biases, which in humans favor structure-dependent generalizations. Without such a hi-

erarchical inductive bias, the network interprets ambiguous input in a way that is most in line with the simplest statistical mapping between input and output – that is, the linear interpretation. The linear interpretation is 'statistically simple' because it is invariant. In the context of our experiment, this means that the word *second* on target-present trials always maps to the same output. Hierarchy, instead, is not directly encoded in the sequential properties of the signal, which makes the hierarchical input-output mapping variable and difficult to learn for a purely data-driven system.

In **Chapter 3**, I continued the discussion about the role of ANN models in the scientific study of language. In particular, we discussed two cases of misalignment between the learning capacities of ANNs and humans, which show that ANNs do not meet the important demand of cognitive fidelity. On the one hand, they are too weak because they fail to learn structure-*dependent* relationships between form and meaning, as found in co-reference and binding. On the other hand, they are too strong because they succeed in learning structure-*independent* regularities that are not found in human language, i.e., impossible languages. Determining the source of these misalignments requires being explicit about what ANNs compute, how they structure data, and how they internally organize their states. Interestingly, attempts to answer these questions do exist. This work often relies on a probing technique, which experimentally manipulates the activation of the model's internal states to see how they relate to its linguistic behavior. But what is curious about this approach is that, when successful, it explains something about the model, not about language. That is, it ultimately yields knowledge about how the ANN represents the (statistical) properties of language, not about how those properties could be represented in the human mind or brain.

In order to use ANN language models in a way that aligns more closely with the scientific study of language, both in terms of goals and criteria for success, we propose two changes to the scientific practice in ANN language modeling. The first is a change to the modeling objective, which is currently narrowly focused on the generation of (probable) sequences. If the focus is shifted towards the interpretation of (possible) structures, we believe that ANNs would be better equipped to handle syntactic principles, whose effects on surface statistics are often very indirect. The second is a change to their cognitive architecture, which should incorporate constraints on possible structures. This will allow the model to learn from more realistic amounts of data and therefore present a fairer opportunity to test the learnability of syntactic principles. With these changes

in mind, computational language modeling research can become more strongly integrated in the scientific study of language.

Having established that hierarchy affects language use, I asked how the brain infers this type of structure during the comprehension of naturally spoken sentences. To this end, **Chapter 4** reported the results of an electroencephalography (EEG) experiment, which set out to test why syntactic phrases are tracked more strongly when they are embedded in regular, meaningful sentences than when the linguistic input in which they are embedded is less meaningful (Kaufeld et al., 2020; Keitel et al., 2018). To test which level of linguistic content modulates cortical tracking of syntax, we compared regular sentences to stimuli that differ from sentences in terms of either their compositional content (idioms, syntactic prose) or their lexical-syntactic content (word lists, jabberwocky). Tracking was quantified via mutual information between the EEG signal and either the speech envelopes or abstract annotations of phrase structure. To make sure that we measured phrase-level tracking, all signals were filtered in the narrow frequency band corresponding to the average presentation rate of phrases in the auditory stimuli (1.1-2.1 Hz). These analyses revealed stronger tracking of phrases in regular, meaningful sentences than in stimuli without either syntactic structure or lexical content (word lists, jabberwocky). However, when both structure and lexical content were preserved in the stimuli (idioms, syntactic prose), phrases were tracked as strongly as in sentences. Critically, we did find a modulation of the N400 component elicited by the sentence-final verb in the four syntactically structured conditions. Thus, while participants did notice differences between the stimuli in terms of their compositional meaning, this did not affect phrase-level speech tracking. In all, these findings refine a recent account of cortical tracking of linguistic structure, which holds that it reflects the internal generation of that structure (Martin & Doumas, 2017; Meyer et al., 2020). Our findings show that this structure-building process, as indexed by cortical tracking effects, is modulated by the lexical-syntactic properties of its input, and not by the compositional interpretation of its output. Any change to the input that results in a weakened activation of the structure-building process (e.g., structurally or lexically impoverished input) will therefore yield reduced phrase-level speech tracking, while modifying the compositional content of the output will have no such effect.

In **Chapter 5**, I used magnetoencephalography (MEG) to study the spatiotemporal correlates of hierarchical structure building in a naturalistic context. We compared three different neuro-computational language models in terms of their

ability to predict brain activity of people listening to a Dutch audiobook story. The models relied on different parsing strategies, which all build the same hierarchical structure but differ in the dynamics of structure building: the top-down parsing model is fully predictive, the left-corner model is mildly predictive, and the bottom-up model is non-predictive (integratory). For each word in the audiobook we calculated a syntactic complexity metric corresponding to the number of nodes that would be visited by the parser when integrating that word into the hierarchical structure of the sentence ('node count'). Using temporal response functions to map these metrics onto delta-band source activity, we found that activity was most accurately reconstructed by node counts derived from the top-down parsing method. These effects were particularly strong in left-hemispheric language regions, including the inferior frontal and superior temporal lobe. Both the bottom-up and the left-corner predictor also increased reconstruction accuracy, but their effects were relatively weak. While the predictiveness of top-down node counts is at odds with the results of previous studies, which were all done in English, it receives a plausible explanation in terms of the grammatical properties of Dutch. In contrast to English, which is head-initial, Dutch exhibits mixed headedness. The presence of head-final structures makes both left-corner and bottom-up parsing insufficiently predictive. The top-down method, instead, captures the predictive nature of language processing well. These findings thus underscore the need for neurobiological studies into languages whose grammatical properties and constructions differ from those of English (Bornkessel-Schlesewsky & Schlesewsky, 2016). These different constructions invite different parsing strategies, so typologically diverse work could point to functional differences within the neurobiological language network across speakers (or signers) of different languages.

In **Chapter 6**, I evaluated a commonly expressed view that the cognitive systems for language and action are related and rely on shared neural resources. We were particularly skeptical about the premise of this view, which holds that language and action are structurally analogous. This claim would require evidence that actions, like language, embody both strong compositionality and structure dependence (see **Chapter 1**, Section 1.2.1). Using a formalization based on set theory, we show that this is not the case. This formal approach reveals that the algebraic structure corresponding to the ordered set of hierarchical syntactic structures can be described as a magma. Its combinatorial operator, which we define as binary set formation, is non-associative, meaning that the order in which it is applied affects the structure that is generated (i.e., (a * b) * c $\neq$ a

* (b * c)). Non-associativity is necessary because sentences can be structurally ambiguous, so the order and manner of combination should be represented in the output of structure building. When we apply this formal model to the domain of actions, it appears to be both too strong and too weak. On the one hand, it can derive in a strongly compositional way two structures from one and the same action sequence. This result is undesirable, because structural ambiguity in the domain of actions does not seem to exist. On the other hand, the model relies on a set-based operator that does not take sequential order into account, yet actions are structured by temporal (precedence) relations. To address these limitations, we present an alternative model, which relies on two associative sequence-building operators that are combined in an algebraic structure called a trace monoid. In this way we account for the observations that actions are fundamentally sequential in nature and exhibit a weaker (structure-independent) form of compositionality. This result indicates that language and action are not structurally analogous, and that the explanation for the overlap in their neural resources should not be sought in the generation of hierarchical structure. More generally, the result of this endeavor underscores that claims about cross-domain convergence are strongest when they are based on theoretically informed comparisons and are accompanied by a formal evaluation of the putative similarity.

7.2 Are there triangles in the brain?

The empirical data and theoretical arguments presented in the preceding chapters show that hierarchy is an indispensable component of language use. If language use relies on hierarchical structure, that structure should somehow be represented in the human brain. While the EEG and MEG results from **Chapters 4** and **5** show that this is the case, they do not establish how exactly the structure is represented. Indeed, the neural representation of hierarchical structure remains a matter of debate. The idea that humans use symbolic tree structures as a form of mental representation has been challenged, even by strong advocates of the view that human cognition relies on the ability to represent symbols. Marcus (2009, 2013), for instance, presents several reasons for doubting the idea that our brains use tree structures for representing phrases and sentences. The first reason relies on evidence from people's behavior in psychological experiments, which, as discussed in **Chapter 1** (Section 1.2.2), presents possible cases of misalignment between the putative capacity to represent tree structures and the actual performance of doing so. Marcus notes that we have trouble remember-

ing sentences verbatim (Lombardi & Potter, 1992), that we experience difficulty in parsing multiply center-embedded sentences, such as "the rat that the cat that the dog chased killed ate the cheese" (Miller & Chomsky, 1963), and that we are prone to illusions of grammaticality. For instance, we sometimes derive globally incoherent and structure-independent interpretations from sentences if they contain subsequences that are locally coherent, as in "the coach smiled at the player tossed a frisbee" (Tabor et al., 2004) and "when the man hunted the deer ran into the woods" (Christianson et al., 2001). In these examples, the underlined sequences are coherent when presented in isolation (as an active main clause), but in the context of the overall sentence, their analysis is ruled out by the grammar. All three findings have been widely discussed, but none of them really challenge the view that hierarchical structure must be mentally represented. We indeed have difficulty remembering sentences verbatim, but failure to remember does not indicate failure to represent. The fact that a sentence is not remembered verbatim just means that its structure is not retained after its meaning has been derived, not that it was not represented accurately in the first place. And in fact, we have even more difficulty remembering sequences without any syntactic structure (Baddeley et al., 2009; Bonhage et al., 2014), showing that memories are also structure-sensitive. It is also true that we have difficulty parsing center-embedded sentences, but the difficulty of these constructions is likely caused by processing overload rather than the parser's inability to represent multiple levels of (center-)embedding (Lewis & Phillips, 2015). Similarly embedded sentences like "the reporter who everyone that I met trusts reported the coup" yield no comparable processing difficulty (Townsend & Bever, 2001; Warren & Gibson, 2002), suggesting that the explanation for the difficulty of center-embedding should not be sought at the level of representational capacity. And last, locally coherent but globally incoherent substrings sometimes yield processing uncertainty, but only under highly specific circumstances. Such effects are restricted mostly to syntactically complex sentences with an unusual analysis, such as when a relative clause is reduced and the form of the verb is ambiguous (Tabor et al., 2004) or when a transitive verb is used intransitively (Christianson et al., 2001), suggesting that these misalignments are the exception rather than the rule. Thus, rather than showing that people behave as if they cannot represent full tree structures (Marcus, 2009), these findings instead show that hierarchical structure is often accurately represented. This conclusion is in line with people's sensitivity to structural principles during processing (see the references cited in **Chapter 1**, Section 1.2.2) as well as with the predictive

accuracy of neuro-computational models that are based on expressive grammatical formalisms (see the results from and references cited in **Chapter 5**).

Marcus (2009, 2013) argues that the representation of hierarchical tree structures is also inconsistent with the architecture of human memory. In computer systems, trees are encoded within a location-addressable architecture, meaning that each node in the structure is assigned a particular location in memory. In human memory, instead, information is retrieved by content rather than location, suggesting that it is incapable of supporting the representation of tree structures (i.e., it does not yield tree-geometric traversability). This claim relies on the assumption that hierarchical tree structures must be represented explicitly, in the form of directed, acyclic graphs. That is not per se necessary though; the relevant properties of a structure can be read off the order in which its elements are combined (their 'derivational history'; see **Chapter 6**), so the tree structures do not need to be represented explicitly. As I discussed in **Chapter 6**, what is needed to capture the type of structure sensitivity found in language is a structure-building operator that is non-associative, meaning that the order in which it is applied affects the structure that is generated. With such an operator, the two interpretations of a structurally ambiguous sentence can be derived from their derivational history without the need to represent tree structures explicitly in the form of graphs. While graph-theoretic representations are commonly used to visualize hierarchical structure, this is done mostly for illustration purposes; the representations themselves are set-theoretic (Lasnik, 2000). Set-based systems are compatible with content-addressable memory (e.g., the label of each set or constituent provides the features used for retrieval), meaning that there is no inconsistency between the properties of human memory and the ability to represent the information encoded in hierarchical syntactic structures (Berwick & Chomsky, 2016; Franck & Wagers, 2020; Kush et al., 2015).

If hierarchical structure does not need to be represented explicitly, it might be represented indirectly in the form of instructions for applying parsing operations. When these operations are embedded in nested programs, they will yield output that is consistent with hierarchical structure, even though the structure itself is not explicitly represented in the neural code. The neurobiological results presented in this thesis are compatible with such an implicit representation of hierarchical structure. In **Chapter 5**, for instance, we quantified syntactic complexity in the form of node count, which we defined as the number of nodes that would be visited by a parser when integrating a word into the structure of the sentence. While this might initially suggest that node count reflects actual nodes

in the tree structure, that is not the case; the parser does not really 'visit' nodes (as in tree traversal). Rather, node count values reflect the number of expand, project, or reduce operations that are applied between successive words (i.e., the number of times these functions are called). On the bottom-up method, then, node count is a proxy for the number of times two words are combined. It thus indirectly reflects a merge- or unification-like *operation*. Likewise, the phrase-level tracking effects in **Chapter 4** are interpreted as reflecting the algorithmic processes that build structure incrementally, not as reflecting the (representations of those) structures themselves. The same holds for related findings in the neurolinguistics literature (Ding et al., 2016; Martin & Doumas, 2017; Nelson et al., 2017; Pallier et al., 2011).

Consistent with this idea, it has been shown in cognitive domains other than language that the brain represents structured sequences in terms of the mental programs that can generate these sequences. In a recent study by Amalric et al. (2017), participants had to look at and memorize visuo-spatial sequences that traced the vertices of a regular octagon in various orders. When the order in which these points are traced is unstructured, participants are required to remember all eight locations on the octagon. However, when the sequences comprise geometrical patterns, such as zig-zags or squares, it suffices to remember just the instructions that generate those patterns (e.g., trace right, rotate, repeat). Behavioral performance indeed shows that people represent spatial sequences in terms of these instructions rather than the independent locations. When the sequences were unstructured, performance was bad (error rate around 50%), plausibly because the number of locations that had to be remembered exceeded working memory capacity. However, when the patterns could be compressed into mental programs comprised of concatenations and embeddings of the instructions, memory performance substantially increased. In fact, performance for a given sequence was accurately predicted by the complexity of the shortest program that could reproduce that sequence, its minimum description length. The longer its description length, the more difficulty participants have in remembering the sequence (Amalric et al., 2017; Mathy & Feldman, 2012). This inverse relationship between program complexity and behavioral performance seems to be quite general, extending to people's ability to anticipate upcoming sequence locations with their eyes (Wang et al., 2019), to verbally describe abstract visual stimuli in compact form (Sun & Firestone, 2021), and to detect occasional outliers in visuo-spatial sequences (Al Roumi et al., 2021) and auditory sequences (Planton et al., 2021).

These findings show that abstract patterns can be represented in the form of generative mental programs, thus supporting the possibility that the hierarchical structure of sentences is similarly encoded via the program that generates that structure. If so, this would make the title of this thesis – "triangles in the brain" – doubly metaphorical. "Triangles" should not be interpreted literally; the triangles are not real, they reflect hierarchical syntactic structure. But to be precise, even the geometry of the tree structure is not real, it is represented in terms of generative mental programs.

7.3 Hierarchy beyond language

It is well-known that humans are quite unique in their ability to infer structure from sequences. While all animals have some pattern extraction ability, the types of structures humans readily infer from sequentially presented input are much more complex (Berwick, Okanoya, et al., 2011; Dehaene et al., 2015; Fitch & Hauser, 2004; ten Cate & Okanoya, 2012; Yang, 2013; Zhang et al., 2022). The hierarchical structure of language is a prime example, but humans recognize hierarchical structure in visuo-spatial patterns (Bahlmann et al., 2009; Fischmeister et al., 2017), mathematical expressions (Makuuchi et al., 2012; Schneider et al., 2012), music (Koelsch et al., 2013), and, as discussed in **Chapter 6**, perhaps also in action sequences (Koechlin & Jubault, 2006). Fitch (2014) coined the term dendrophilia to describe this seemingly human proclivity to attribute hierarchical tree-like structures to sensory patterns (*dendrophilia* means "tree-loving", in Greek). A key prediction of the dendrophilia hypothesis is that this proclivity biases learning in situations where the properties of the learned domain conform to hierarchical structure. Consistent with this prediction, the results of experimental paradigms that rely on a poverty-of-the-stimulus logic similar to the ambiguous train-test regime used in **Chapter 2** (see also the summary of that chapter in Section 7.1) indicate that humans have a propensity to induce rules that build hierarchical structures (formally known as supra-regular languages) rather than 'simpler' rules based on regular expressions (like linear adjacency; Culbertson & Adger, 2014; Ferrigno et al., 2020; McCoy et al., 2021; Morgan & Ferreira, 2021). I put the word *simpler* between inverted commas because while regular languages are seen as simpler and less powerful than supra-regular languages in formal language theory, what might seem simpler formally need not be easier for the human brain. In fact, hierarchical generalization might

be the more natural generalization over ambiguous sequences, as was suggested more than 50 years ago by George Miller in his Project Grammarama:

> "Constituent-structure languages [i.e., supra-regular languages] are more natural, easier to cope with, than regular languages ... The hierarchical structure of strings generated by constituent-structure grammars is characteristic of much other behavior that is sequentially organized; it seems plausible that it would be easier for people than would the left-to-right organization characteristic of strings generated by regular grammars."
>
> Miller (1967, p. 140)

That hierarchical generalization is a natural inference for humans is also suggested by a large literature on the acquisition of syntactic constructions, which shows that children converge on structure-dependent generalizations despite the evidence in line with linear generalizations (Crain & Nakayama, 1987; Fodor & Crowther, 2002; Kam & Fodor, 2012; Lidz et al., 2003; Legate & Yang, 2002; Shi et al., 2020). Regardless of whether the learning mechanisms underlying these generalizations are domain-general or language-specific (see Ambridge et al., 2014 and Crain et al., 2017 for two different perspectives), the observation that humans consistently generalize hierarchically speaks in favor of the idea that, for the human brain, linear generalizations are less natural.

If the foregoing discussion about the implicit representation of hierarchy in the form of generative mental programs (Section 7.2) is on the right track, then Fitch (2014)'s dendrophilia hypothesis might be reformulated as a proclivity to infer from sensory patterns the hierarchical mental programs that generate those patterns. This idea was recently proposed in a hypothesis about human singularity by Dehaene et al. (2022), who argue that humans possess multiple internal languages of thought (e.g., languages of mathematics, music, geometry), which use similar computational principles but rely on different cortical networks. Each language of thought contains compositional rules which are used to compress sequential input data into hierarchically nested mental programs that are capable of reproducing those data. As the cognitive complexity of the compression is determined by the size of the shortest possible mental program (its minimum description length; see Section 7.2), this type of hierarchical input compression enables efficient storage and control of information in memory. And as it is applicable across several cognitive domains beyond language, including mathematics, music, and geometry (Al Roumi et al., 2021; Amalric et al., 2017; Bor et al., 2003; Chater & Vitányi, 2003; Mathy & Feldman, 2012; Planton et al., 2021; Restle, 1970; Simon, 1972; Wang et al., 2019), it might indeed form the basis

of humans' proclivity to infer hierarchical structure from any type of sequential input.

Dehaene et al. (2022)'s proposal raises fundamental questions at the heart of cognitive (neuro)science. To what extent are the compositional rules that make up different languages of thought the same? In their proposal, all languages of thought follow similar computational principles (i.e., discrete symbols, composition, efficient compression), but that raises the question where differences between cognitive domains come from. One possibility is that they arise from the use of different compression schemes, which can detect different regularities. As Dehaene et al. (2022) note, many domains (mathematics, music, geometry) can generate symmetrical structures, but symmetry is typically avoided in syntactic structure (Kayne, 1994; Moro, 1997). The regularities underlying symmetrical and antisymmetrical patterns might be described more or less efficiently in different algorithms, and these differences could form the basis of representational domain specificity. Moreover, if all languages of thought rely on the same fundamental computational principles, do different domains also share the principles that underlie empirical generalizations? In other words, should one expect to find evidence of linguistic principles, such as locality constraints or binding principles, in other cognitive domains? While these principles are commonly formulated in domain-specific (syntactic) terms, the concepts underlying them (i.e., locality and binding) are fundamentally domain-general. Being able to explain architectural similarities and differences across cognitive domains using the same principles is one of the ultimate goals of an integrated theory of cognitive (neuro)science.

7.4 Concluding remarks

The aim of this thesis was to shed light on the question how we use and represent hierarchical syntactic structure during language comprehension. I hope to have shown that studying these syntactic triangles in the brain is a fundamentally interdisciplinary endeavor, which relies on insights and methodologies from many domains, including linguistics, psychology, cognitive neuroscience, and computer science. In five chapters, I investigated the role of hierarchical structure in language use, showing that people are biased to interpret phrases hierarchically **(Chapter 2**), and that such a hierarchical bias is unlikely to arise in currently popular artificial neural network models (**Chapters 2** and **3**). Two neuroscientific chapters showed that the tracking of hierarchical structure is

driven by word-level properties (**Chapter 4**), and that the hierarchical structure of Dutch sentences is largely built in a predictive manner (**Chapter 5**). While **Chapter 6** showed that the type of hierarchy found in language is fundamentally different from that found in the structure underlying actions, hierarchical organization might be a defining feature of human cognition. By studying cognitive domains in terms of their tendency to organize information hierarchically, we reach a better understanding of one of the fundamental properties of the human mind.

References

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, *20*(3), 233–250. doi: 10.1007/BF01067217

Adger, D. (2018). The autonomy of syntax. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic structures after 60 Years: The impact of the Chomskyan revolution in linguistics* (pp. 153–175). De Gruyter Mouton.

Adger, D. (2019). *Language unlimited*. Oxford: Oxford University Press.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367–13372. doi: 10.1073/pnas.201400998

Alexiadou, A., Haegeman, L., & Stavrou, M. (2007). *Noun phrase in the generative perspective*. De Gruyter Mouton.

Al Roumi, F., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, *109*(16), 2627-2639.e4. doi: 10.1016/j.neuron.2021.06.009

Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., & Dehaene, S. (2017). The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLOS Computational Biology*, *13*(1), e1005273. doi: 10.1371/journal.pcbi.1005273

Ambridge, B., Pine, J. M., & Lieven, E. V. M. (2014). Child language acquisition: Why Universal Grammar doesn't help. *Language*, *90*(3), e53-e90. doi: 10.1353/lan.2014.0051

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*(4), 525–560. doi: 10.1080/01690965.2012.658072

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. doi: 10.1016/j.jml.2009.09.005

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, *61*(3), 438–456. doi:

10.1016/j.jml.2009.05.004

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, *56*(1), 65–85. doi: 10.1016/j.jml.2006.08.004

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. doi: 10.1016/j.tics.2008.02.004

Baggio, G. (2021). Compositionality in a parallel architecture for language processing. *Cognitive Science*, *45*(5), e12949. doi: 10.1111/cogs.12949

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367. doi: 10.1080/01690965.2010.542671

Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial grammar processing engages Broca's area. *NeuroImage*, *42*(2), 525–534. doi: 10.1016/j.neuroimage.2008.04.249

Bahlmann, J., Schubotz, R. I., Mueller, J. L., Koester, D., & Friederici, A. D. (2009). Neural circuits of hierarchical visuo-spatial sequence processing. *Brain Research*, *1298*, 161–170. doi: 10.1016/j.brainres.2009.08.017

Bai, F., Meyer, A. S., & Martin, A. E. (2022). Neural dynamics differentially encode phrases and sentences during spoken language comprehension. *PLOS Biology*, *20*(7), e3001713. doi: 10.1371/journal.pbio.3001713

Baker, M. C. (2001). *The atoms of language: The mind's hidden rules of grammar*. Basic Books.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190307. doi: 10.1098/rstb.2019.0307

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In S. Lappin & J.-P. Bernardy (Eds.), *Algebraic structures in natural language* (pp. 1–16). Taylor & Francis.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of online locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1178. doi: 10.1037/a0024194

Bastiaansen, M. C. M., & Hagoort, P. (2015). Frequency-based segregation of syntactic and semantic unification during online sentence level language comprehension. *Journal of Cognitive Neuroscience*, *27*(11), 2095–2107.

doi: 10.1162/jocn_a_00829

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Baumann, P. (2014). Dependencies and hierarchical structure in sentence processing. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 152–157).

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463

Berent, I., & Marcus, G. (2019). No integration without structured representations: Response to Pater. *Language*, *95*(1), e75-e86. doi: 10.1353/lan.2019.0011

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2-3), 150–177.

Berwick, R. C. (2018). Revolutionary new ideas appear infrequently. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic structures after 60 Years: The impact of the Chomskyan revolution in linguistics* (pp. 177–194). De Gruyter Mouton.

Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and evolution*. Cambridge, MA: MIT Press.

Berwick, R. C., & Chomsky, N. (2017). Why only us: Recent questions and answers. *Journal of Neurolinguistics*, *43*, 166–177. doi: 10.1016/j.jneuroling.2016.12.002

Berwick, R. C., Okanoya, K., Beckers, G. J. L., & Bolhuis, J. J. (2011). Songs to syntax: The linguistics of birdsong. *Trends in Cognitive Sciences*, *15*(3), 113–121. doi: 10.1016/j.tics.2011.01.002

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*(7), 1207–1242. doi: 10.1111/j.1551-6709.2011.01189.x

Bhattasali, S., Fabre, M., Luh, W.-M., Saied, H. A., Constant, M., Pallier, C., ... Hale, J. (2019). Localising memory retrieval and syntactic composition: An fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, *34*(4), 491–510. doi: 10.1080/23273798.2018.1518533

Blanco-Elorrieta, E., Ding, N., Pylkkänen, L., & Poeppel, D. (2020). Under-

standing requires tracking: Noise and knowledge interact in bilingual comprehension. *Journal of Cognitive Neuroscience*, *32*(10), 1975–1983. doi: 10.1162/jocn_a_01610

Bloom, P. (1994). Generativity within language and other cognitive domains. *Cognition*, *51*(2), 177–189. doi: 10.1016/0010-0277(94)90014-0

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93. doi: 10.1016/0010-0285(91)90003-7

Boeckx, C. (2009). The nature of Merge: Consequences for language, mind, and biology. In M. Piattelli-Palmarini, J. Uriagereka, & P. Salaburu (Eds.), *Of minds and language: A dialogue with Noam Chomsky in the Basque country* (pp. 44–57). Oxford: Oxford University Press.

Boeckx, C., & Fujita, K. (2014). Syntax, action, comparative cognitive science, and Darwinian thinking. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00627

Boeckx, C., Martinez-Alvarez, A., & Leivada, E. (2014). The functional neuroanatomy of serial order in language. *Journal of Neurolinguistics*, *32*, 1–15. doi: 10.1016/j.jneuroling.2014.07.001

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer [Computer program].*

Boland, J. E., & Blodgett, A. (2006). Argument status and PP-attachment. *Journal of Psycholinguistic Research*, *35*(5), 385–403. doi: 10.1007/s10936-006-9021-z

Bonhage, C. E., Fiebach, C. J., Bahlmann, J., & Mueller, J. L. (2014). Brain signature of working memory for sentence structure: Enriched encoding and facilitated maintenance. *Journal of Cognitive Neuroscience*, *26*(8), 1654–1671. doi: 10.1162/jocn_a_00566

Bonhage, C. E., Meyer, L., Gruber, T., Friederici, A. D., & Mueller, J. L. (2017). Oscillatory EEG dynamics underlying automatic chunking during sentence processing. *NeuroImage*, *152*, 647–657. doi: 10.1016/j.neuroimage.2017.03.018

Bor, D., Duncan, J., Wiseman, R. J., & Owen, A. M. (2003). Encoding strategies dissociate prefrontal activity from working memory demand. *Neuron*, *37*(2), 361–367. doi: 10.1016/S0896-6273(02)01171-6

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2016). The importance of linguistic typology for the neurobiology of language. *Linguistic Typology*, *20*(3), 241–252. doi: 10.1515/lingty-2016-0032

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal

function. *Trends in Cognitive Sciences*, *12*(5), 201–208. doi: 10.1016/j.tics.2008.02.009

Bourguignon, M., Tiège, X. D., de Beeck, M. O., Ligot, N., Paquier, P., Bogaert, P. V., ... Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping*, *34*(2), 314–326. doi: 10.1002/hbm.21442

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

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

Brennan, J. R., Dyer, C., Kuncoro, A., & Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, *146*, 107479. doi: 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. doi: 10.1371/journal.pone.0207741

Brennan, J. R., & Martin, A. E. (2020). Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190305. doi: 10.1098/rstb.2019.0305

Brennan, J. R., 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. doi: 10.1016/j.bandl.2010.04.002

Brennan, J. R., & Pylkkänen, L. (2017). MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science*, *41*(S6), 1515–1531. doi: 10.1111/cogs.12445

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. doi: 10.1016/j.bandl.2016.04.008

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, *11*, e72056. doi: 10.7554/eLife.72056

Brodbeck, C., Brooks, T. L., Das, P., Reddigari, S., & Kulasingham, J. P. (2021). *Eelbrain 0.37*. Zenodo. doi: 10.5281/zenodo.4650416

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLOS Biology*, *18*(10), e3000883. doi: 10.1371/journal.pbio.3000883

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, *172*, 162–174. doi: 10.1016/j.neuroimage .2018.01.042

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *Journal of Neuroscience*, *39*(38), 7564–7575. doi: 10.1523/JNEUROSCI.0584-19.2019

Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia*, *135*, 107225. doi: 10 .1016/j.neuropsychologia.2019.107225

Burroughs, A., Kazanina, N., & Houghton, C. (2021). Grammatical category and the neural processing of phrases. *Scientific Reports*, *11*(1), 2446. doi: 10.1038/s41598-021-81901-5

Bybee, J. L. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 201–134). John Benjamins.

Cacciari, C. (2014). Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, *9*(2), 267–293. doi: 10.1075/ml.9.2.05cac

Cacciari, C., & Corradini, P. (2015). Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study. *Journal of Cognitive Psychology*, *27*(7), 797–811. doi: 10.1080/20445911.2015.1049178

Cacciari, C., & Glucksberg, S. (1991). Understanding idiomatic expressions: The contribution of word meanings. In G. B. Simpson (Ed.), *Understanding word and sentence* (pp. 217–240). Elsevier. doi: 10.1016/S0166-4115(08) 61535-6

Cacciari, C., Padovani, R., & Corradini, P. (2007). Exploring the relationship between individuals' speed of processing and their comprehension of spoken idioms. *European Journal of Cognitive Psychology*, *19*(3), 417–445. doi: 10.1080/09541440600763705

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, *27*(6), 668–683. doi: 10.1016/0749-596X(88) 90014-9

Canal, P., Pesciarelli, F., Vespignani, F., Molinaro, N., & Cacciari, C. (2017). Basic composition and enriched integration in idiom processing: An EEG study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(6), 928–943. doi: 10.1037/xlm0000351

Carnie, A. (2021). *Syntax: A generative introduction*. Wiley-Blackwell.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. doi: 10.1371/journal.pcbi.1000436

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22. doi: 10.1016/S1364-6613(02)00005-0

Chen, H., Huang, S., Chiang, D., & Chen, J. (2017). Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1936–1945). Vancouver, Canada: Association for Computational Linguistics. doi: 10.18653/v1/P17-1177

Chen, L., Goucha, T., Männel, C., Friederici, A. D., & Zaccarella, E. (2021). Hierarchical syntactic processing is beyond mere associating: Functional magnetic resonance imaging evidence from a novel artificial grammar. *Human Brain Mapping*, *42*(10), 3253–3268. doi: 10.1002/hbm.25432

Chesi, C. (2015). On directionality of phrase structure building. *Journal of Psycholinguistic Research*, *44*(1), 65–89. doi: 10.1007/s10936-014-9330-6

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and beyond: The cartography of syntactic structures* (Vol. 3, pp. 39–103). Oxford: Oxford University Press.

Chollet, F. (2015). *Keras*.

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

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1959). A review of B. F. Skinner's Verbal behavior. *Language*, *35*(1), 26–58. doi: 10.2307/411334

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1979). *Language and responsibility*. Pantheon.

Chomsky, N. (1980). *Rules and representations*. New York, NY: Columbia Uni-

versity Press.

Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York, NY: Praeger.

Chomsky, N. (1991). Linguistics and cognitive science: Problems and mysteries. In A. Kasher (Ed.), *The Chomskyan turn* (pp. 26–53). Blackwell.

Chomsky, N. (1995a). Bare phrase structure. In H. R. Campos & P. M. Kempchinsky (Eds.), *Evolution and revolution in linguistic theory: Essays in honor of Carlos Otero* (pp. 51–109). Washington, DC: Georgetown University Press.

Chomsky, N. (1995b). *The minimalist program*. Cambridge, MA: MIT Press.

Chomsky, N. (2013). Problems of projection. *Lingua*, *130*, 33–49. doi: 10.1016/j.lingua.2012.12.003

Chow, W.-Y., Lewis, S., & Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psychology*, *5*.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205. doi: 10.1207/s15516709cog2302_2

Christiansen, M. H., & Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01182

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*(s1), 126–161. doi: 10.1111/j.1467-9922.2009.00538.x

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407. doi: 10.1006/cogp.2001.0752

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. doi: 10.1016/j.tics.2019.01.009

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, *36*(3), 315–332. doi: 10.1162/0024389054396917

Cinque, G. (2013). Cognition, universal grammar, and typological generalizations. *Lingua*, *130*, 50–65. doi: 10.1016/j.lingua.2012.10.007

Cogan, G. B., & Poeppel, D. (2011). A mutual information analysis of neural coding of speech by low-frequency MEG phase information. *Journal of Neurophysiology*, *106*(2), 554–563. doi: 10.1152/jn.00075.2011

Collins, C. (2001). Economy conditions in syntax. In M. Baltin & C. Collins (Eds.), *The handbook of contemporary syntactic theory* (pp. 45–61). Oxford: Blackwell Publishers.

Collins, C. (2017). Merge(X,Y) = {X,Y}. In L. Bauke & A. Blümel (Eds.), *Labels and roots* (pp. 47–68). De Gruyter Mouton. doi: 10.1515/9781501502118-003

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297–338. doi: 10.1080/026432900380427

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887. doi: 10.1037/0033-295X.113.4.887

Coopmans, C. W., de Hoop, H., Hagoort, P., & Martin, A. E. (2022). Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiology of Language*, *3*(3), 386–412. doi: 10.1162/nol_a_00070

Coopmans, C. W., de Hoop, H., Kaushik, K., Hagoort, P., & Martin, A. E. (2022). Hierarchy in language interpretation: Evidence from behavioural experiments and computational modelling. *Language, Cognition and Neuroscience*, *37*(4), 420–439. doi: 10.1080/23273798.2021.1980595

Coopmans, C. W., & Nieuwland, M. S. (2020). Dissociating activation and integration of discourse referents: Evidence from ERPs and oscillations. *Cortex*, *126*, 83–106. doi: 10.1016/j.cortex.2019.12.028

Coopmans, C. W., & Schoenmakers, G.-J. (2020). Incremental structure building of preverbal PPs in Dutch. *Linguistics in the Netherlands*, *37*, 38–52. doi: 10.1075/avt.00036.coo

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*(4), 597–612. doi: 10.1017/S0140525X00071491

Crain, S., Koring, L., & Thornton, R. (2017). Language acquisition from a biolinguistic perspective. *Neuroscience & Biobehavioral Reviews*, *81*, 120–149. doi: 10.1016/j.neubiorev.2016.09.004

Crain, S., & McKee, C. (1985). Acquisition of structural restrictions on anaphora. In *North East Linguistics Society* (Vol. 16, pp. 94–110).

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, *63*(3), 522–543. doi: 10.2307/415004

Crain, S., & Pietroski, P. (2001). Nature, nurture and Universal Grammar. *Linguistics and Philosophy*, *24*(2), 139–186. doi: 10.1023/A:

1005694100138

Crain, S., & Thornton, R. (1991). Recharting the course of language acquisition: Studies in elicited production. In N. A. Krasnegor, D. M. Rumbaugh, R. L. Schiefelbusch, & M. Studdert-Kennedy (Eds.), *Biological and behavioral determinants of language development* (pp. 321–337). Psychology Press.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847. doi: 10.1073/pnas.1320525111

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139. doi: 10.1016/j.jml.2014.05.006

Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition*, *25*(1), 57–71. doi: 10.3758/BF03197285

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. doi: 10.1006/nimg.1998.0395

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*(12), 1924-1937.e9. doi: 10.1016/j.cub.2019.04.067

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems*, *18*(3), 191–212. doi: 10.1080/09548980701609235

Dawkins, R. (1976). Hierarchical organisation: A candidate principle for ethology. In P. Bateson & R. Hinde (Eds.), *Growing points in ethology* (pp. 7–54). Cambridge: Cambridge University Press.

de Vries, W., & Nissim, M. (2021). As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 836–846). doi: 10.18653/v1/2021.findings-acl.74

de Waal, F. B. M., & Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. *Trends in Cognitive Sciences*, *14*(5), 201–207. doi: 10.1016/j.tics.2010.03.003

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*(9), 751–766. doi: 10.1016/j.tics.2022.06

.010

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*(1), 2–19. doi: 10.1016/j.neuron.2015.09.019

Demberg, V., & Keller, F. (2019). Cognitive models of syntax and sentence processing. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 293–312). Cambridge, MA: MIT Press.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. doi: 10.1016/j.neuroimage.2010.06.010

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding* (No. arXiv:1810.04805). arXiv. doi: 10.48550/arXiv.1810.04805

Di Liberto, G. M., Lalor, E. C., & Millman, R. E. (2018). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *NeuroImage*, *166*, 247–258. doi: 10.1016/j.neuroimage.2017.10.066

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. doi: 10.1016/j.jml.2013.04.003

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience*, *11*. doi: 10.3389/fnhum.2017.00481

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. doi: 10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187. doi: 10.1016/j.neubiorev.2017.02.011

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*. doi: 10.3389/fnhum.2014.00311

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768. doi: 10.1016/

j.neuroimage.2013.06.035

Dominey, P. F., Hoen, M., Blanc, J.-M., & Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies. *Brain and Language*, *86*(2), 207–225. doi: 10.1016/S0093-934X(02)00529-1

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385-393.e9. doi: 10.1016/j.neuron.2019.10.019

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–45. doi: 10.1037/0033-295X.115.1.1

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife*, *9*, e55978. doi: 10.7554/eLife.55978

Dunbar, E. (2019). Generative grammar, neural networks, and the implementational mapping problem: Response to Pater. *Language*, *95*(1), e87-e98. doi: 10.1353/lan.2019.0013

Eisele, J., & Lust, B. (1996). Knowledge about pronouns: A developmental study using a truth-value judgment task. *Child Development*, *67*(6), 3086–3100. doi: 10.1111/j.1467-8624.1996.tb01904.x

Embick, D., & Poeppel, D. (2015). Towards a computational(ist) neurobiology of language: Correlational, integrated and explanatory neurolinguistics. *Language, Cognition and Neuroscience*, *30*(4), 357–366. doi: 10.1080/23273798.2014.980750

Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743. doi: 10.1016/j.tics.2015.09.008

Fadiga, L., Craighero, L., & D'Ausilio, A. (2009). Broca's area in language, action, and music. *Annals of the New York Academy of Sciences*, *1169*(1), 448–458. doi: 10.1111/j.1749-6632.2009.04582.x

Fazio, P., Cantagallo, A., Craighero, L., D'Ausilio, A., Roy, A. C., Pozzo, T., … Fadiga, L. (2009). Encoding of human action in Broca's area. *Brain*, *132*(7), 1980–1988. doi: 10.1093/brain/awp118

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. doi: 10.1016/j.cognition.2020.104348

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G.,

& Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256-E6262. doi: 10.1073/pnas.1612132113

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Ferreira, F., & Qiu, Z. (2021). Predicting syntactic structure. *Brain Research*, *1770*, 147632. doi: 10.1016/j.brainres.2021.147632

Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., & Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians. *Science Advances*, *6*(26), eaaz1002. doi: 10.1126/sciadv.aaz1002

Fiebach, C. J., & Schubotz, R. I. (2006). Dynamic anticipatory processing of hierarchical sequential events: A common role for Broca's area and ventral premotor cortex across domains? *Cortex*, *42*(4), 499–502. doi: 10.1016/S0010-9452(08)70386-1

Fischmeister, F. P., Martins, M. J. D., Beisteiner, R., & Fitch, W. T. (2017). Self-similarity and recursion as default modes in human cognition. *Cortex*, *97*, 183–201. doi: 10.1016/j.cortex.2016.08.016

Fitch, W. T. (2010). Three meanings of "recursion": Key distinctions for biolinguistics. In H. Yamakido, R. K. Larson, & V. Déprez (Eds.), *The evolution of human language: Biolinguistic perspectives* (pp. 73–90). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511817755.005

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*(3), 329–364. doi: 10.1016/j.plrev.2014.04.005

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*(5656), 377–380. doi: 10.1126/science.1089401

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*(1), 87–104. doi: 10.1111/nyas.12406

Flaherty, M., Hunsicker, D., & Goldin-Meadow, S. (2021). Structural biases that children bring to language learning: A cross-cultural look at gestural input to homesign. *Cognition*, *211*, 104608. doi: 10.1016/j.cognition.2021.104608

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture:

A critical analysis. *Cognition*, *28*(1-2), 3–71. doi: 10.1016/0010-0277(88)90031-5

Fodor, J. D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *The Linguistic Review*, *18*(1-2). doi: 10.1515/tlir.19.1-2.105

Fong, S., & Berwick, R. C. (2008). Treebank parsing and knowledge of language: A cognitive perspective. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 539–544). Austin, TX: Cognitive Science Society.

Fong, S., Malioutov, I., Yankama, B., & Berwick, R. C. (2013). Treebank parsing and knowledge of language. In A. Villavicencio, T. Poibeau, A. Korhonen, & A. Alishahi (Eds.), *Cognitive Aspects of Computational Language Acquisition* (pp. 133–172). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-31863-4

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)* (pp. 61–69).

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4), 371–404. doi: 10.1080/01690960143000254

Franck, J., & Wagers, M. (2020). Hierarchical structure and memory mechanisms in agreement attraction. *PLOS ONE*, *15*(5), e0232163. doi: 10.1371/journal.pone.0232163

Frank, R., Mathis, D., & Badecker, W. (2013). The acquisition of anaphora by simple recurrent networks. *Language Acquisition*, *20*(3), 181–227. doi: 10.1080/10489223.2013.796950

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. doi: 10.1177/0956797611409589

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522–4531. doi: 10.1098/rspb.2012.1741

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*, *33*(9), 1213–1218. doi: 10.1080/23273798.2018.1424347

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *PLOS ONE*, *13*(5), e0197304. doi:

10.1371/journal.pone.0197304

Frazier, L. (1985). Syntactic complexity. In A. M. Zwicky, D. R. Dowty, & L. Karttunen (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511597855.005

Frazier, L. (2015). Two interpretive systems for natural language? *Journal of Psycholinguistic Research*, *44*(1), 7–25. doi: 10.1007/s10936-014-9328-0

Freidin, R., & Lasnik, H. (1981). Disjoint reference and wh-trace. *Linguistic Inquiry*, *12*(1), 39–53.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, *103*(7), 2458–2463. doi: 10.1073/pnas.0509389103

Fujita, K. (2014). Recursive Merge and human language evolution. In T. Roeper & M. Speas (Eds.), *Recursion: Complexity in cognition* (pp. 243–264). Springer.

Fujita, K. (2017). On the parallel evolution of syntax and lexicon: A Merge-only view. *Journal of Neurolinguistics*, *43*, 178–192. doi: 10.1016/j.jneuroling.2016.05.001

Fukui, N. (2011). Merge and bare phrase structure. In C. Boeckx (Ed.), *The Oxford handbook of linguistic minimalism* (pp. 73–95). Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199549368.013.0004

Fukui, N., & Zushi, M. (2004). Introduction. In N. Chomsky (Ed.), *The generative enterprise revisited: Discussions with Riny Huybregts, Henk van Riemsdijk, Naoki Fukui and Mihoko Zushi* (pp. 1–25). Berlin: Walter de Gruyter.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). *Neural language models as psycholinguistic subjects: Representations of syntactic state* (No. arXiv:1903.03260). arXiv.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. doi: 10.1038/s42256-020-00257-z

Getz, H., Ding, N., Newport, E. L., & Poeppel, D. (2018). Cortical tracking of constituent structure in language acquisition. *Cognition*, *181*, 135–140. doi: 10.1016/j.cognition.2018.08.019

Ghitza, O. (2017). Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience*, *32*(5), 545–561. doi: 10.1080/23273798.2016.1232419

Gibbs, R. W., Nayak, N. P., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, *28*(5), 576–593. doi: 10.1016/0749-596X(89)90014-4

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. doi: 10.1016/S0010-0277(98)00034-1

Giglio, L., Ostarek, M., Sharoh, D., & Hagoort, P. (2022). *Diverging neural dynamics for syntactic structure building in naturalistic speaking and listening*. bioRxiv.

Giglio, L., Ostarek, M., Weber, K., & Hagoort, P. (2022). Commonalities and asymmetries in the neurobiological infrastructure for language production and comprehension. *Cerebral Cortex*, *32*(7), 1405–1418. doi: 10.1093/cercor/bhab287

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, *41*(50), 10316–10329. doi: 10.1523/JNEUROSCI.0812-21.2021

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. doi: 10.1038/nn.3063

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., & Zuidema, W. (2018). Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. doi: 10.48550/arXiv.1808.08079

Gleitman, L. R., & Newport, E. L. (1995). The invention of language by children: Environmental and biological influences. In L. R. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science, Vol. 1: Language* (p. 704). Cambridge, MA: MIT Press.

Goldberg, A. E., & Michaelis, L. A. (2017). One among many: Anaphoric one and its relationship with numeral one. *Cognitive Science*, *41*(S2), 233–258. doi: 10.1111/cogs.12339

Goldberg, Y. (2019). *Assessing BERT's syntactic abilities* (No. arXiv:1901.05287). arXiv.

Goldin-Meadow, S. (2003). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language.* Psychology Press.

Goucha, T., Anwander, A., Adamson, H., & Friederici, A. D. (2022). *Native language leaves distinctive traces in brain connections.* bioRxiv.

Graves, P. (1994). Flakes and ladders: What the archaeological record cannot tell us about the origins of language. *World Archaeology*, *26*(2), 158–171. doi: 10.1080/00438243.1994.9980270

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of language* (pp. 73–113). Cambridge, MA: MIT Press.

Greenfield, P. M. (1991). Language, tools and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and Brain Sciences*, *14*(4), 531–551. doi: 10.1017/S0140525X00071235

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biology*, *11*(12), e1001752. doi: 10.1371/journal.pbio.1001752

Guest, O., & Martin, A. E. (2021a). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. doi: 10.1177/1745691620970585

Guest, O., & Martin, A. E. (2021b). *On logical inference over brains, behaviour, and artificial neural networks.* PsyArXiv.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics. doi: 10.18653/v1/N18-1108

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

Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews*, *81*, 194–204. doi: 10.1016/j.neubiorev.2017.01.048

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, *37*(1), 347–362. doi: 10.1146/annurev-neuro-071013-013847

Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Pittsburgh, Pennsylvania: Association for Computational Linguistics. doi: 10.3115/1073336.1073357

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

Hale, J. T. (2014). *Automaton theories of human sentence comprehension*. CSLI Publications.

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. doi: 10.1146/annurev-linguistics-051421-020803

Hamburger, H., & Crain, S. (1984). Acquisition of cognitive compiling. *Cognition*, *17*(2), 85–136. doi: 10.1016/0010-0277(84)90015-5

Hauser, M. D. (2009). The possibility of impossible cultures. *Nature*, *460*(7252), 190–196. doi: 10.1038/460190a

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. doi: 10.1126/science.298.5598.1569

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., ... Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00401

Heaven, D. (2019). Why deep-learning AIs are so easy to fool. *Nature*, *574*(7777), 163–166. doi: 10.1038/d41586-019-03013-5

Heeris, J. (2018). *Gammatone Filterbank Toolkit*.

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. doi: 10.1073/pnas.2201968119

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Oxford: Blackwell.

Heinz, J., & Idsardi, W. (2011). Sentence and word complexity. *Science*, *333*(6040), 295–297. doi: 10.1126/science.1210358

Henke, L., & Meyer, L. (2021). Endogenous oscillations time-constrain linguistic segmentation: Cycling the garden path. *Cerebral Cortex*, *31*(9), 4289–4299. doi: 10.1093/cercor/bhab086

Higuchi, S., Chaminade, T., Imamizu, H., & Kawato, M. (2009). Shared neural correlates for language and tool use in Broca's area. *NeuroReport*, *20*(15), 1376–1381. doi: 10.1097/WNR.0b013e3283315570

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Holloway, R. L. (1969). Culture: A human domain. *Current Anthropology*, *10*, 395–412.

Holsinger, E., & Kaiser, E. (2013). Processing (non)compositional expressions: Mistakes and recovery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 866–878. doi: 10.1037/a0030410

Hornstein, N. (2009). *A theory of syntax: Minimal operations and Universal Grammar*. Cambridge: Cambridge University Press.

Hornstein, N. (2017). On Merge. In J. McGilvray (Ed.), *The Cambridge companion to Chomsky* (Second ed., pp. 69–86). Cambridge: Cambridge University Press. doi: 10.1017/9781316716694.004

Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology*, *104*(5), 2500–2511. doi: 10.1152/jn.00251.2010

Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2021). Individual word activation and word frequency effects during the processing of opaque idiomatic expressions. *Quarterly Journal of Experimental Psychology*. doi: 10.1177/17470218211047995

Hubers, F., Ginkel, W. V., Cucchiarini, C., Strik, H., & Dijkstra, A. (2018). *Normative data on Dutch idiomatic expressions: Native speakers.* Data Archiving and Networked Services (DANS). doi: 10.17026/DANS-ZJX-HNSK

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195. doi: 10.1080/15475441.2005.9684215

Hultén, A., Schoffelen, J.-M., Uddén, J., Lam, N. H., & Hagoort, P. (2019). How the brain makes sense beyond the processing of single words – An MEG study. *NeuroImage*, *186*, 586–594. doi: 10.1016/j.neuroimage.2018.11.035

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466. doi: 10.1037/0033-295X.104.3.427

Humphreys, G. W., & Forde, E. M. E. (1998). Disordered action schema and action disorganisation syndrome. *Cognitive Neuropsychology*, *15*(6/7/8), 771–811.

Hunsicker, D., & Goldin-Meadow, S. (2012). Hierarchical structure in a self-created communication system: Building nominal constituents in home-sign. *Language*, *88*(4), 732–763. doi: 10.1353/lan.2012.0092

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795. doi: 10.1613/jair.1.11674

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLOS ONE*, *9*(7), e100986. doi: 10.1371/journal.pone.0100986

Ince, R. A. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping*, *38*(3), 1541–1573. doi: 10.1002/hbm.23471

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Jackendoff, R. (1977). *X' syntax: A theory of phrase structure*. Cambridge, MA: MIT Press.

Jackendoff, R. (1988). Why are they saying these things about us? *Natural Language & Linguistic Theory*, *6*(3), 435–442.

Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Erlbaum.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jackendoff, R. (2007). *Language, consciousness, culture: Essays on mental structure*. Cambridge, MA: MIT Press.

Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Perception*, *26*(3), 195–204. doi: 10.1525/mp.2009.26.3.195

Jackendoff, R. (2011). What is the human language faculty? Two views. *Language*, *87*(3), 586–624.

Jackendoff, R. (2017). In defense of theory. *Cognitive Science*, *41*(S2), 185–212.

doi: 10.1111/cogs.12324

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, *97*(2), 211–225. doi: 10.1016/j.cognition.2005.04.006

Jeon, H.-A. (2014). Hierarchical processing in the prefrontal cortex in a variety of cognitive domains. *Frontiers in Systems Neuroscience*, *8*. doi: 10.3389/fnsys.2014.00223

Jin, P., Lu, Y., & Ding, N. (2020). Low-frequency neural activity reflects rule-based chunking during speech listening. *eLife*, *9*, e55613. doi: 10.7554/eLife.55613

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Joos, M. (1957). *Readings in linguistics*. Washington, DC: American Councel of Learned Societies.

Joshi, A. K., & Schabes, Y. (1997). Tree-Adjoining Grammars. In G. Rozenberg & A. Salomaa (Eds.), *Handbook of formal languages* (Vol. 3, pp. 69–123). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-59126-6_2

Kaan, E., & Swaab, T. Y. (2002). The brain circuitry of syntactic comprehension. *Trends in Cognitive Sciences*, *6*(8), 350–356. doi: 10.1016/S1364-6613(02)01947-2

Kam, X.-N. C., & Fodor, J. D. (2012). Children's acquisition of syntax: Simple models are too simple. In M. Piattelli-Palmarini & R. C. Berwick (Eds.), *Rich languages from poor inputs* (pp. 43–60). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199590339.003.0003

Kaufeld, G., Bosker, H. R., ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *The Journal of Neuroscience*, *40*(49), 9467–9475. doi: 10.1523/JNEUROSCI.0302-20.2020

Kayne, R. S. (1994). *The antisymmetry of syntax*. Cambridge, MA: MIT Press.

Kayne, R. S. (2011). Why are there no directionality parameters? In *Proceedings of WCCFL* (Vol. 28, pp. 1–23).

Kayser, S. J., Ince, R. A. A., Gross, J., & Kayser, C. (2015). Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *Journal of Neuroscience*, *35*(44), 14691–14701. doi: 10.1523/JNEUROSCI.2243-15.2015

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora.

Journal of Memory and Language, *56*(3), 384–409. doi: 10.1016/j.jml.2006.09.003

Kazanina, N., & Phillips, C. (2001). Coreference in child Russian: Distinguishing syntactic and discourse constraints. In *Proceedings of the 25th Boston University Conference on Language Development*. Somerville, MA.

Kazanina, N., & Tavano, A. (2022). What neural oscillations can and cannot do for syntactic structure building. *Nature Reviews Neuroscience*, 1–16. doi: 10.1038/s41583-022-00659-5

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biology*, *16*(3), e2004473. doi: 10.1371/journal.pbio.2004473

Keitel, A., Ince, R. A. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, *147*, 32–42. doi: 10.1016/j.neuroimage.2016.11.062

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. doi: 10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi: 10.3758/BRM.42.3.643

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602. doi: 10.1016/0749-596X(91)90027-H

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. doi: 10.1016/j.csl.2017.01.005

Koechlin, E., & Jubault, T. (2006). Broca's area and the hierarchical organization of human behavior. *Neuron*, *50*(6), 963–974. doi: 10.1016/j.neuron.2006.05.017

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443–15448. doi: 10.1073/pnas.1300272110

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645. doi: 10.1023/A:1026528912821

Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence

formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*(1), 68–101. doi: 10.1016/j.cogpsych.2008.05.002

Kösem, A., & van Wassenhove, V. (2017). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*(5), 536–544. doi: 10.1080/23273798.2016.1238495

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., & Blunsom, P. (2018). LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1426–1436). Melbourne, Australia: Association for Computational Linguistics. doi: 10.18653/v1/P18-1132

Kuperberg, G. R. (2020). Tea with milk? A hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science*. doi: 10.1111/tops.12518

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

Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, *82*, 18–40. doi: 10.1016/j.jml.2015.02.003

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. doi: 10.1146/annurev.psych.093008.131123

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2873–2882). PMLR.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*. doi: 10.1037/rev0000297

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, *213*, 104699. doi: 10.1016/j.cognition.2021.104699

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). New York, NY:

Wiley.

Lasnik, H. (1976). Remarks on coreference. *Linguistic Analysis*, *2*, 1–21. doi: 10.1007/978-94-009-2542-7_4

Lasnik, H. (2000). *Syntactic structures revisited: Contemporary lectures on classic transformational theory*. Cambridge, MA: MIT Press.

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language*, *98*(1), 74–88. doi: 10.1016/j.bandl.2006.02.003

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 151–162. doi: 10.1515/tlir.19.1-2.151

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*(1), 27–46. doi: 10.1007/s10936-014-9329-z

Li, J., & Hale, J. (2019). Grammatical predictors for fMRI time-courses. In R. C. Berwick & E. P. Stabler (Eds.), *Minimalist parsing* (pp. 159–173). Oxford: Oxford University Press. doi: 10.1093/oso/9780198795087.003.0007

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition*, *36*(6), 1103–1121. doi: 10.3758/MC.36.6.1103

Lidz, J. (2018). The explanatory power of linguistic theory. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic structures after 60 years: The impact of the Chomskyan revolution in linguistics* (pp. 225–239). De Gruyter Mouton.

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal Grammar and statistical learning. *Annual Review of Linguistics*, *1*(1), 333–353. doi: 10.1146/annurev-linguist-030514-125236

Lidz, J., Lukyanenko, C., & Sutton, M. (2021). The hunt for structure-dependent interpretation: The case of Principle C. *Cognition*, 104676. doi: 10.1016/j.cognition.2021.104676

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*(3), 295–303. doi: 10.1016/S0010-0277(03)00116-1

Lightfoot, D. (1982). *The language lottery: Toward a biology of grammars*. Cambridge, MA: MIT Press.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5210–5217).

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. doi: 10.1146/annurev-linguistics-032020-051035

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. doi: 10.1162/tacl_a_00115

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of Memory and Language*, *31*(6), 713–733. doi: 10.1016/0749-596X(92)90036-W

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. doi: 10.1162/nol_a_00029

Loula, J., Baroni, M., & Lake, B. M. (2018). Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 108–114).

Lukyanenko, C., Conroy, A., & Lidz, J. (2014). Is she patting Katie? Constraints on pronominal reference in 30-month-olds. *Language Learning and Development*, *10*(4), 328–344. doi: 10.1080/15475441.2013.853529

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010. doi: 10.1016/j.neuron.2007.06.004

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior* (No. arXiv:2005.02181). arXiv.

Maffongelli, L., D'Ausilio, A., Fadiga, L., & Daum, M. M. (2019). The ontogenesis of action syntax. *Collabra: Psychology*, *5*(1), 21. doi: 10.1525/collabra.215

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., & Golumbic, E. Z. (2017). Sleep disrupts high-level speech parsing despite significant basic auditory processing. *Journal of Neuroscience*, *37*(32), 7772–7781. doi: 10.1523/JNEUROSCI.0168-17.2017

Makuuchi, M., Bahlmann, J., & Friederici, A. D. (2012). An approach to separating the levels of hierarchical structure building in language and mathe-

matics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2033–2045. doi: 10.1098/rstb.2012.0095

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: Examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, *174*, 57–68. doi: 10.1016/j.visres.2020.04.013

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., ... Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019. doi: 10.1038/s41593-022-01114-5

Marcus, G. (2009). How does the mind work? Insights from biology. *Topics in Cognitive Science*, *1*, 145–172. doi: 10.1111/j.1756-8765.2008.01007.x

Marcus, G. (2013). Evolution, memory, and the nature of syntactic representation. In J. J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech, and language: Exploring the evolution of mind and brain* (pp. 27–44). Cambridge, MA: MIT Press.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282. doi: 10.1006/cogp.1998.0694

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, G. F. (2006). Cognitive architecture and descent with modification. *Cognition*, *101*(2), 443–465. doi: 10.1016/j.cognition.2006.04.009

Marcus, G. F. (2018). *The deepest problem with deep learning*. The Gradient.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Marslen-Wilson, W. D. (2007). Morphological processes in language comprehension. In M. G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 175–186). Oxford: Oxford University Press.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71. doi: 10.1016/0010-0277(80)90015-3

Martin, A., Holtz, A., Abels, K., Adger, D., & Culbertson, J. (2020). Experimental evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa: A Journal of General Linguistics*, *5*(1), 97. doi: 10.5334/gjgl.1085

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in*

Psychology, *7*. doi: 10.3389/fpsyg.2016.00120

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427. doi: 10.1162/jocn_a_01552

Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology*, *15*(3), e2000663. doi: 10.1371/journal.pbio.2000663

Martin, A. E., & Doumas, L. A. A. (2019). Predicate learning in neural systems: Using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences*, *29*, 77–83. doi: 10.1016/j.cobeha.2019.04.008

Martin, A. E., & Doumas, L. A. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190306. doi: 10.1098/rstb.2019.0306

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*(3), 879–906. doi: 10.1016/j.jml.2007.06.010

Martins, M. J. D. (2012). Distinctive signatures of recursion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2055–2064. doi: 10.1098/rstb.2012.0097

Martins, M. J. D., Bianco, R., Sammler, D., & Villringer, A. (2019). Recursion in action: An fMRI study on the generation of new hierarchical levels in motor sequences. *Human Brain Mapping*, *40*(9), 2623–2638. doi: 10.1002/hbm.24549

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202). Brussels, Belgium: Association for Computational Linguistics. doi: 10.18653/v1/D18-1151

Matar, S., Dirani, J., Marantz, A., & Pylkkänen, L. (2021). Left posterior temporal cortex is sensitive to syntax within conceptually matched Arabic expressions. *Scientific Reports*, *11*(1), 7181. doi: 10.1038/s41598-021-86474-x

Matchin, W., Brodbeck, C., Hammerly, C., & Lau, E. (2019). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human Brain Mapping*, *40*(2), 663–678. doi: 10.1002/hbm.24403

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123. doi: 10.1016/j.cortex.2016.12.010

Matchin, W., & Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, *30*(3), 1481–1498. doi: 10.1093/cercor/bhz180

Matchin, W., İlkbaşaran, D., Hatrak, M., Roth, A., Villwock, A., Halgren, E., & Mayberry, R. I. (2022). The cortical organization of syntactic processing Is supramodal: Evidence from American Sign Language. *Journal of Cognitive Neuroscience*, *34*(2), 224–235. doi: 10.1162/jocn_a_01790

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362. doi: 10.1016/j.cognition.2011.11.003

Matthei, E. H. (1982). The acquisition of prenominal modifier sequences. *Cognition*, *11*(3), 301–332. doi: 10.1016/0010-0277(82)90018-X

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., ... Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479. doi: 10.1162/jocn.1993.5.4.467

Mazurkiewicz, A. (1995). Introduction to Trace Theory. In V. Diekert & G. Rozenberg (Eds.), *The book of traces* (pp. 3–41). Singapore: World Scientific. doi: 10.1142/9789814261456_0001

McCoy, R. T., Culbertson, J., Smolensky, P., & Legendre, G. (2021). Infinite use of finite means? Evaluating the generalization of center embedding learned from an artificial grammar. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2093–2098).

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, *8*(0), 125–140.

McDaniel, D., & McKee, C. (1992). Which children did they show obey strong crossover? In H. Goodluck & M. Rochemont (Eds.), *Island constraints: Theory, acquisition and processing* (pp. 275–294). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-017-1980-3_10

McRae, K., Brown, K. S., & Elman, J. L. (2019). Prediction-based learning and processing of event knowledge. *Topics in Cognitive Science*, *13*(1), 206–223. doi: 10.1111/tops.12482

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., & Friederici, A. D. (2017).

Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, *27*(9), 4293–4302. doi: 10.1093/cercor/bhw228

Meyer, L., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089–1099. doi: 10.1080/23273798.2019.1693050

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (No. arXiv:1301.3781). arXiv. doi: 10.48550/arXiv.1301.3781

Miller, G. A. (1967). Project Grammarama. In G. A. Miller (Ed.), *Psychology of communication*. New York, NY: Basic Books.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology ii* (pp. 419–491). New York: Wiley.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York, NY: Holt, Rinehart and Winston.

Mitchell, J., & Bowers, J. (2020). Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5147–5158). Barcelona, Spain (Online): International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.451

Mitchell, J., Kazanina, N., Houghton, C., & Bowers, J. (2019). Do LSTMs know about Principle C? In *Conference on Cognitive Computational Neuroscience* (pp. 188–191). Cognitive Computational Neuroscience. doi: 10.32470/CCN.2019.1241-0

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, *48*(7), 2642–2650. doi: 10.1111/ejn.13811

Molinaro, N., Lizarazu, M., Baldin, V., Pérez-Navarro, J., Lallier, M., & Ríos-López, P. (2021). Speech-brain phase coupling is enhanced in low contextual semantic predictability conditions. *Neuropsychologia*, *156*, 107830. doi: 10.1016/j.neuropsychologia.2021.107830

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., ... Fedorenko, E. (2020). Composition is the core driver of the language-selective network. *Neurobiology of Language*, *1*(1), 104–134. doi: 10.1162/nol_a_00005

Moreno, A., Limousin, F., Dehaene, S., & Pallier, C. (2018). Brain correlates of

constituent structure in sign language comprehension. *NeuroImage*, *167*, 151–161. doi: 10.1016/j.neuroimage.2017.11.040

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language*, *80*(2), 188–207. doi: 10.1006/brln.2001.2588

Morgan, A. M., & Ferreira, V. S. (2021). Beyond input: Language learners produce novel relative clause types without exposure. *Journal of Cognitive Psychology*, *33*(5), 483–517. doi: 10.1080/20445911.2021.1928678

Moro, A. (1997). Dynamic antisymmetry: Movement as a symmetry-breaking phenomenon. *Studia Linguistica*, *51*(1), 50–76. doi: 10.1111/1467-9582.00017

Moro, A. (2014a). On the similarity between syntax and actions. *Trends in Cognitive Sciences*, *18*(3), 109–110. doi: 10.1016/j.tics.2013.11.006

Moro, A. (2014b). Response to Pulvermüller: The syntax of actions and other metaphors. *Trends in Cognitive Sciences*, *18*(5), 221. doi: 10.1016/j.tics.2014.01.012

Moro, A. (2015). *The boundaries of Babel: The brain and the enigma of impossible languages*. Cambridge, MA: MIT Press.

Moro, A. (2016). *Impossible languages*. Cambridge, MA: MIT Press.

Mulligan, K., Frank, R., & Linzen, T. (2021). Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. In *Proceedings of the Society for Computation in Linguistics* (Vol. 4, pp. 125–135).

Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*(7), 774–781. doi: 10.1038/nn1077

Nakatani, K., & Gibson, E. (2008). Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics*, *46*(1), 63–87. doi: 10.1515/LING.2008.003

Nelson, M. J., Karoui, I. E., Giber, K., Yang, X., Cohen, L., Koopman, H., ... Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669-E3678. doi: 10.1073/pnas.1701590114

Newmeyer, F. J. (2005). *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.

Nieuwland, M. S., Coopmans, C. W., & Sommers, R. P. (2019). Distinguishing old from new referents during discourse comprehension: Evidence from

ERPs and oscillations. *Frontiers in Human Neuroscience*, *13*. doi: 10.3389/fnhum.2019.00398

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. doi: 10.1037/0033-295X.88.1.1

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences*, *23*(11), 913–926. doi: 10.1016/j.tics.2019.08.004

O'Donnell, T. J., Hauser, M. D., & Fitch, W. T. (2005). Using mathematical models of language experimentally. *Trends in Cognitive Sciences*, *9*(6), 284–289. doi: 10.1016/j.tics.2005.04.011

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *156869*. doi: 10.1155/2011/156869

Pagin, P., & Westerståhl, D. (2010). Compositionality I: Definitions and variants. *Philosophy Compass*, *5*(3), 250–264. doi: 10.1111/j.1747-9991.2009.00228.x

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. doi: 10.1073/pnas.1018711108

Papitto, G., Friederici, A. D., & Zaccarella, E. (2020). The topographical organization of motor processing: An ALE meta-analysis on six action domains and the relevance of Broca's region. *NeuroImage*, *206*, 116321. doi: 10.1016/j.neuroimage.2019.116321

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, *25*(12), 1649–1653. doi: 10.1016/j.cub.2015.04.049

Partee, B. (1975). Montague grammar and transformational grammar. *Linguistic Inquiry*, *6*(2), 203–300.

Partee, B. H. (1995). Lexical semantics and compositionality. In L. R. Gleitman & M. Liberman (Eds.), *Language: An invitation to cognitive science, Vol. 1, 2nd ed.* (pp. 311–360). Cambridge, MA: The MIT Press.

Partee, B. H. (2007). Compositionality and coercion in semantics: The dynamics of adjective meaning. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 145–161). Royal Netherlands Academcy of Arts and Sciences.

Partee, B. H., ter Meulen, A., & Wall, R. E. (1993). *Mathematical methods in linguistics* (Vol. 30). Dordrecht: Kluwer Academic Publishers.

Payne, J., Pullum, G. K., Scholz, B. C., & Berlage, E. (2013). Anaphoric one and its implications. *Language*, *89*(4), 794–829. doi: 10.1353/lan.2013.0071

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*. doi: 10.3389/fpsyg.2012.00320

Peña, M., & Melloni, L. (2012). Brain oscillations during spoken sentence processing. *Journal of Cognitive Neuroscience*, *24*(5), 1149–1164. doi: 10.1162/jocn_a_00144

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. doi: 10.1016/j.cognition.2010.11.001

Peterson, R. R., Burgess, C., Dell, G. S., & Eberhard, K. M. (2001). Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1223. doi: 10.1037/0278-7393.27.5.1223

Petty, J., & Frank, R. (2021). *Transformers generalize linearly* (No. arXiv:2109.12036). arXiv.

Phillips, C. (2006). The real-time status of island phenomena. *Language*, *82*(4), 795–823.

Phillips, C. (2013). Parser-grammar relations: We don't understand everything twice. In M. Sanz, I. Laka, & M. K. Tanenhaus (Eds.), *Language down the garden path: The cognitive and biological basis for linguistic structures* (pp. 294–315). Oxford: Oxford University Press.

Phillips, C., & Lewis, S. (2013). Derivational order in syntax: Evidence and architectural consequences. *Studies in Linguistics*, *6*, 11–47.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. T. Runner (Ed.), *Experiments at the interfaces* (pp. 147–180). Bringly: Emerald.

Pietroski, P. (2015). Vocabulary matters. In Á. J. Gallego & D. Ott (Eds.), *MIT Working Papers in Linguistics* (pp. 199–210).

Pietroski, P., & Hornstein, N. (2020). Universal Grammar. In A. J. Lerner, S. Cullen, & S.-J. Leslie (Eds.), *Current controversies in philosophy of cognitive science* (pp. 13–28). Routledge.

Pinker, S. (1997). *How the mind works*. New York, NY: W. W. Norton & Company.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York, NY:

Basic Books.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. doi: 10.1016/0010-0277(88)90032-7

Planton, S., van Kerkoerle, T., Abbih, L., Maheu, M., Meyniel, F., Sigman, M., … Dehaene, S. (2021). A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLOS Computational Biology*, *17*(1), e1008598. doi: 10.1371/journal.pcbi.1008598

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*(1-2), 34–55. doi: 10.1080/02643294.2012.710600

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334. doi: 10.1038/s41583-020-0304-4

Post, B., Marslen-Wilson, W. D., Randall, B., & Tyler, L. K. (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition*, *109*(1), 1–17. doi: 10.1016/j.cognition.2008.06.011

Puebla, G., Martin, A. E., & Doumas, L. A. A. (2021). The relational processing limits of classic and contemporary neural network models of language processing. *Language, Cognition and Neuroscience*, *36*(2), 240–254. doi: 10.1080/23273798.2020.1821906

Pulvermüller, F. (2014). The syntax of action. *Trends in Cognitive Sciences*, *18*(5), 219–220. doi: 10.1016/j.tics.2014.01.001

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360. doi: 10.1038/nrn2811

Rawski, J., & Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to Pater. *Language*, *95*(1), e125-e135. doi: 10.1353/lan.2019.0021

Read, C., & Schreiber, P. (1982). Why short subjects are harder to find than long ones. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 78–101). Cambridge University Press.

Reason, J. T. (1979). Actions not as planned. In G. Underwood & R. Stevens (Eds.), *Aspects of consciousness* (pp. 67–90). London: Academic Press.

Reinhart, T. (1983). *Anaphora and semantic interpretation*. London: Croom Helm.

Resnik, P. (1992). Left-corner parsing and psychological plausibility. In *Proceed-*

ings of the 14th International Conference on Computational Linguistics (pp. 191–197).

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review*, *77*, 481–495. doi: 10.1037/h0029964

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences*, *22*(10), 870–882. doi: 10.1016/j.tics.2018.08.003

Rimmele, J. M., Poeppel, D., & Ghitza, O. (2021). Acoustically driven cortical delta oscillations underpin prosodic chunking. *eNeuro*, *8*(4). doi: 10.1523/ENEURO.0562-20.2021

Rizzi, L. (2009). The discovery of language invariance and variation, and its relevance for the cognitive sciences. *Behavioral and Brain Sciences*, *32*(5), 467–468. doi: 10.1017/S0140525X09990574

Rizzi, L. (2013). Introduction: Core computational principles in natural language syntax. *Lingua*, *130*, 1–13. doi: 10.1016/j.lingua.2012.12.001

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, *25*(5), 762–776. doi: 10.1162/jocn_a_00337

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science*, *26*(4), 525–554. doi: 10.1016/j.humov.2007.04.001

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*(4), 493–515. doi: 10.1006/jmla.2000.2759

Saito, M., & Fukui, N. (1998). Order in phrase structure and movement. *Linguistic Inquiry*, *29*(3), 439–474. doi: 10.1162/002438998553815

Scheepers, C., & Sturt, P. (2014). Bidirectional syntactic priming across cognitive domains: From arithmetic to language and back. *The Quarterly Journal of Experimental Psychology*, *67*(8), 1643–1654. doi: 10.1080/17470218.2013.873815

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex*, *96*, 105–120. doi: 10.1016/j.cortex.2017.09.002

Schneider, E., Maruyama, M., Dehaene, S., & Sigman, M. (2012). Eye gaze reveals a fast, parallel extraction of the syntax of arithmetic formulas. *Cognition*, *125*(3), 475–490. doi: 10.1016/j.cognition.2012.06.015

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. doi: 10.1073/pnas.2105646118

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18. doi: 10.1016/j.tins.2008.09.012

Schütze, C. T., & Gibson, E. (1999). Argumenthood and English prepositional phrase attachment. *Journal of Memory and Language*, *40*(3), 409–431. doi: 10.1006/jmla.1998.2619

Schwartz, M. F. (2006). The cognitive neuropsychology of everyday action and planning. *Cognitive Neuropsychology*, *23*(1), 202–221. doi: 10.1080/02643290500202623

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. doi: 10.1016/j.neuropsychologia.2019.107307

Sheather, S. J. (2009). Diagnostics and transformations for multiple linear regression. In S. Sheather (Ed.), *A modern approach to regression with R* (pp. 151–225). New York, NY: Springer. doi: 10.1007/978-0-387-09608-7_6

Shen, Y., Tan, S., Sordoni, A., & Courville, A. (2019). Ordered neurons: Integrating tree structures into recurrent neural networks. In *Proceedings of the 7th International Conference on Learning Representations*.

Sheng, J., Zheng, L., Lyu, B., Cen, Z., Qin, L., Tan, L. H., ... Gao, J.-H. (2019). The cortical maps of hierarchical linguistic structures during speech perception. *Cerebral Cortex*, *29*(8), 3232–3240. doi: 10.1093/cercor/bhy191

Shi, R., Legrand, C., & Brandenberger, A. (2020). Toddlers track hierarchical structure dependence. *Language Acquisition*, *27*(4), 397–409. doi: 10.1080/10489223.2020.1776010

Shlens, J. (2014). *A tutorial on principal component analysis* (No. arXiv:1404.1100). arXiv. doi: 10.48550/arXiv.1404.1100

Simon, H. A. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review*, *79*(5), 369–382. doi: 10.1037/h0033118

Smith, N., & Tsimpli, I.-M. (1995). *The mind of a savant*. Blackwell Publishers.

Smith, N., & Tsimpli, I.-M. (1997). Reply to Bates. *International Journal of Bilingualism*, *1*(2), 180–186.

Smith, N. V., Tsimpli, I.-M., & Ouhalla, J. (1993). Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, *91*(4), 279–347. doi: 10.1016/0024-3841(93)90002-E

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90. doi: 10.1038/416087a

Smolka, E., Rabanus, S., & Rösler, F. (2007). Processing verbs in German idioms: Evidence against the Configuration Hypothesis. *Metaphor and Symbol*, *22*(3), 213–231. doi: 10.1080/10926480701357638

Snijders, T. M., Vosse, T., Kempen, G., Van Berkum, J. J., Petersson, K. M., & Hagoort, P. (2009). Retrieval and unification of syntactic structure in sentence comprehension: An fMRI study using word-category ambiguity. *Cerebral Cortex*, *19*(7), 1493–1503. doi: 10.1093/cercor/bhn187

Spenader, J., & Blutner, R. (2007). Compositionality and systematicity. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 163–174). Royal Netherlands Academcy of Arts and Sciences.

Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*(2), 161–184. doi: 10.1016/j.jml.2005.11.001

Sprouse, J., & Hornstein, N. (2016). Syntax and the cognitive neuroscience of syntactic structure building. In G. Hickok & S. L. Small (Eds.), *Neurobiology of language* (pp. 165–174). Elsevier.

Stanojević, M., Bhattasali, S., Dunagan, D., Campanelli, L., Steedman, M., Brennan, J., & Hale, J. (2021). Modeling incremental language comprehension in the brain with Combinatory Categorial Grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 23–38). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.cmcl-1.3

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 425. doi: 10.1037/0278-7393.32.2.425

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: MIT press.

Steedman, M. (2002). Plans, affordances, and Combinatory Grammar. *Linguistics and Philosophy*, *25*(5), 723–753. doi: 10.1023/A:1020820000972

Stolk, A., Todorovic, A., Schoffelen, J.-M., & Oostenveld, R. (2013). Online and

offline tools for head movement compensation in MEG. *NeuroImage*, *68*, 39–48. doi: 10.1016/j.neuroimage.2012.11.047

Stout, D., & Chaminade, T. (2009). Making tools and making sense: Complex, intentional behaviour in human evolution. *Cambridge Archaeological Journal*, *19*(1), 85–96. doi: 10.1017/S0959774309000055

Sun, Z., & Firestone, C. (2021). Seeing and speaking: How verbal "description length" encodes visual complexity. *Journal of Experimental Psychology: General*, *151*(1), 82. doi: 10.1037/xge0001076

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks* (No. arXiv:1312.6199). arXiv.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370. doi: 10.1016/j.jml.2004.01.001

Takahashi, E., & Lidz, J. (2007). Beyond statistical learning in syntax. In A. Gavarró & M. J. Freitas (Eds.), *Language acquisition and development: Proceedings of GALA* (pp. 446–456). Cambridge Scholars Publishing.

ten Cate, C., & Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: Natural vocalizations and artificial grammar learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1984–1994. doi: 10.1098/rstb.2012.0055

ten Oever, S., & Martin, A. E. (2021). An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *eLife*, *10*, e68066. doi: 10.7554/eLife.68066

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. doi: 10.1126/science.1192788

Tettamanti, M., Alkadhi, H., Moro, A., Perani, D., Kollias, S., & Weniger, D. (2002). Neural correlates for the acquisition of natural language syntax. *NeuroImage*, *17*(2), 700–709. doi: 10.1006/nimg.2002.1201

Tettamanti, M., & Moro, A. (2012). Can syntax appear in a mirror (system)? *Cortex*, *48*(7), 923–935. doi: 10.1016/j.cortex.2011.05.020

Tettamanti, M., & Weniger, D. (2006). Broca's area: A supramodal hierarchical processor? *Cortex*, *42*(4), 491–494. doi: 10.1016/S0010-9452(08)70384

-8

Tezcan, F., Weissbart, H., & Martin, A. E. (2022). *A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension.* bioRxiv.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. doi: 10.1080/15475440709336999

Titone, D. A., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, *31*(12), 1655–1674. doi: 10.1016/S0378-2166(99)00008-9

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: MIT Press.

Tran, K., Bisazza, A., & Monz, C. (2018). The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4731–4736).

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454–475. doi: 10.1006/jmla.1996.0025

Tyler, L. K., Stamatakis, E. A., Post, B., Randall, B., & Marslen-Wilson, W. (2005). Temporal and frontal systems in speech comprehension: An fMRI study of past tense processing. *Neuropsychologia*, *43*(13), 1963–1974. doi: 10 .1016/j.neuropsychologia.2005.03.008

Uddén, J., & Bahlmann, J. (2012). A rostro-caudal gradient of structured sequence processing in the left inferior frontal gyrus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2023–2032. doi: 10.1098/rstb.2012.0009

Uithol, S., van Rooij, I., Bekkering, H., & Haselager, P. (2012). Hierarchies in action and motor control. *Journal of Cognitive Neuroscience*, *24*(5), 1077–1086. doi: 10.1162/jocn_a_00204

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories. *Social Psychology*, *51*(5), 285–298. doi: 10.1027/1864-9335/a000428

van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1597–1605). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/N15-1183

Van Wagenen, S., Brennan, J., & Stabler, E. P. (2014). Quantifying parsing complexity as a function of grammar. *UCLA Working Papers in Linguistics*,

18, 31–47.

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*(4), 533–567. doi: 10.1080/01690960903310587

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, *22*(8), 1682–1700. doi: 10.1162/jocn.2009.21293

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. doi: 10.1016/j.jml.2009.04.002

Wang, L., Amalric, M., Fang, W., Jiang, X., Pallier, C., Figueira, S., ... Dehaene, S. (2019). Representation of spatial sequences using nested rules in human prefrontal cortex. *NeuroImage*, *186*, 245–255. doi: 10.1016/j.neuroimage.2018.10.061

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112. doi: 10.1016/S0010-0277(02)00087-2

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. doi: 10.1162/tacl_a_00321

Waters, G. S., & Caplan, D. (2004). Verbal working memory and on-line syntactic processing: Evidence from self-paced listening. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *57*(1), 129–163. doi: 10.1080/02724980343000170

Watumull, J., Hauser, M. D., Roberts, I. G., & Hornstein, N. (2014). On recursion. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.01017

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. doi: 10.1162/jocn_a_01467

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, *57*, 59–70. doi: 10.1016/j.neuropsychologia.2014.03.001

Wexler, K. (1990). Innateness and maturation in linguistic development. *Developmental Psychobiology*, *23*(7), 645–660. doi: 10.1002/dev.420230708

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019). Structural

supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3302–3312).

Xiang, B., Yang, C., Li, Y., Warstadt, A., & Kann, K. (2021). *CLiMP: A benchmark for Chinese language model evaluation* (No. arXiv:2101.11131). arXiv.

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, *110*(16), 6324–6327. doi: 10.1073/pnas.1216803110

Yang, C. (2015). Negative knowledge from positive evidence. *Language*, *91*(4), 938–953. doi: 10.1353/lan.2015.0054

Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456. doi: 10.1016/j.tics.2004.08.006

Yang, C. D., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, *81*, 103–119. doi: 10.1016/j.neubiorev.2016.12.023

Yoshida, M., Dickey, M. W., & Sturt, P. (2013). Predictive processing of syntactic structure: Sluicing and ellipsis in real-time sentence processing. *Language and Cognitive Processes*, *28*(3), 272–302. doi: 10.1080/01690965.2011.622905

Zaccarella, E., Meyer, L., Makuuchi, M., & Friederici, A. D. (2017). Building by syntax: The neural basis of minimal linguistic structures. *Cerebral Cortex*, *27*(1), 411–421. doi: 10.1093/cercor/bhv234

Zaccarella, E., Papitto, G., & Friederici, A. D. (2021). Language and action in Broca's area: Computational differentiation and cortical segregation. *Brain and Cognition*, *147*, 105651. doi: 10.1016/j.bandc.2020.105651

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3. doi: 10.1037/0033-2909.127.1.3

Zhang, H., Zhen, Y., Yu, S., Long, T., Zhang, B., Jiang, X., ... Wang, L. (2022). Working memory for spatial sequences: Developmental and evolutionary factors in encoding ordinal and relational structures. *The Journal of Neuroscience*, *42*(5), 850–864. doi: 10.1523/JNEUROSCI.0603-21.2021

Zhang, L., & Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study. *NeuroImage*,

111, 228–240. doi: 10.1016/j.neuroimage.2015.02.028

Zoefel, B., & VanRullen, R. (2015). The role of high-level processes for oscillatory phase entrainment to speech sound. *Frontiers in Human Neuroscience*, *9*. doi: 10.3389/fnhum.2015.00651

Zoefel, B., & VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage*, *124*, 16–23. doi: 10.1016/j.neuroimage.2015.08.054

Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., ... Ding, N. (2019). Auditory and language contributions to neural encoding of speech features in noisy environments. *NeuroImage*, *192*, 66–75. doi: 10.1016/j.neuroimage.2019.02.047

Nederlandse samenvatting

Om een kralenketting te maken, moet je één voor één de kralen aan het touw rijgen. Het produceren van taal lijkt op het eerste gezicht een vergelijkbaar proces. Bij het van maken zinnen, bijvoorbeeld, zetten we woorden één voor één achter elkaar, alsof die woorden kralen aan een zinsketting zijn. Als deze vergelijking klopt, dan betekent dit dat de structuur en betekenis van zinnen bepaald worden door de volgorde waarin woorden achter elkaar worden gezet. In zekere zin is dat ook wel zo – "man bijt hond" is immers niet hetzelfde als "hond bijt man" – maar in een belangrijk opzicht is die gedachte te simpel. Van buitenaf gezien zijn zinnen inderdaad *lineair*, maar onder die lineaire sequentie gaat een bepaald type structuur schuil. Die structuur is *hiërarchisch*, wat inhoudt dat woorden worden gecombineerd in constituenten, die vervolgens weer worden gecombineerd met andere woorden in grotere constituenten, en zo door. Die hiërarchische structuur wordt vaak weergegeven in zogenaamde boomstructuren, waarvan er een aantal in dit proefschrift te vinden zijn (bijvoorbeeld op bladzijde 14 in hoofdstuk 1).

Het ingewikkelde aan dit concept is dat hiërarchische structuur zowel onzichtbaar als onhoorbaar is; het enige dat we zien of horen is die lineaire stroom aan woorden. Maar dat de structuur er wel degelijk is, is te zien aan het feit dat zinnen structureel ambigu kunnen zijn. Neem bijvoorbeeld de zin "de vrouw bekijkt de man met de verrekijker". Deze zin heeft twee betekenissen, terwijl geen van de woorden ambigu is. De ambiguïteit volgt uit het feit dat de constituent "met de verrekijker" gelinkt kan worden aan zowel "de man" als "zag de man". In het eerste geval is er een man met een verrekijker die bekeken wordt door een vrouw. De daarbij behorende hiërarchische structuur is [zag [de man met de verrekijker]] (de woordgroepen die hier tussen vierkante haken staan, zijn constituenten). En in het tweede geval gebruikt de vrouw de verrekijker om de man te bekijken. De structuur die daarbij hoort is [[zag [de man]] met de verrekijker]. Dit voorbeeld laat zien dat zinnen meer zijn dan kettingen van woorden. Om taal te begrijpen, moeten we aan de hand van lineaire input een hiërarchisch georganiseerde structuur opbouwen. Dit proefschrift gaat over de vraag hoe we dat doen.

In hoofdstuk 2 bekijk ik of hiërarchische structuur ook wordt opgebouwd in taalgebruik. In recente studies is namelijk beweerd dat hiërarchie vaak genegeerd wordt bij het gebruiken van taal, en dat taalverwerking voornamelijk een lineair en sequentieel proces is. Als het inderdaad zo is dat taalgebruik fundamenteel lineair is, dan zou je verwachten dat mensen een voorkeur, of *bias*, hebben voor de lineaire interpretatie van zinnen en constituenten. Dat wil zeggen, als er constituenten zijn waarvoor de hiërarchische en de lineaire interpretatie niet hetzelfde zijn, en mensen hebben een lineaire bias, dan zouden ze over het algemeen voor de lineaire interpretatie moeten kiezen. Dit heb ik getest in een gedragsexperiment waarin proefpersonen plaatjes te zien kregen met een rij groene en blauwe ballen. Vervolgens werden ze gevraagd om de *tweede blauwe bal* aan te wijzen. Een constituent zoals *tweede blauwe bal* is op twee manieren te interpreteren. Als je het lineair interpreteert, dan zeggen zowel *tweede* als *blauwe* iets over *bal*; het geheel verwijst dan naar iets dat een blauwe bal en ook een tweede bal is, oftewel, de tweede bal, die blauw is. Bij de hiërarchische interpretatie, daarentegen, worden *blauwe* en *bal* eerst samengevoegd. Omdat *blauwe* en *bal* nu een eenheid vormen, zegt *tweede* niet iets over *bal* maar over *blauwe bal*. Het geheel verwijst dan naar de tweede van de set blauwe ballen. De cruciale manipulatie in dit experiment was dat de plaatjes zó waren opgebouwd dat de lineaire en de hiërarchische interpretatie *niet* naar dezelfde bal verwezen. Om een voorbeeld te geven: stel je een rijtje ballen voor, waarbij de eerste bal groen is (bal 1), de tweede blauw (bal 2), en de derde ook blauw (bal 3). Als mensen een lineaire bias hebben, zouden ze bal 2 moeten kiezen, omdat deze bal blauw is en de tweede in de rij. Dit is echter niet de tweede van de set blauwe ballen, omdat de eerste bal groen is. Dus, hebben mensen een bias voor hiërarchie, dan kiezen ze voor bal 3. De resultaten van dit experiment kunnen heel kort worden samengevat: alle proefpersonen interpreteerden *tweede blauwe bal* hiërarchisch. Dat laat zien dat hiërarchie meer is dan alleen een manier om taalstructuur te beschrijven. Hiërarchie heeft psychologische relevantie; het is de manier waarop onze geest structuur geeft aan de taal die we om ons heen zien en horen.

De tweede vraag uit hoofdstuk 2 is in hoeverre computationele taalmodellen in staat zijn om dit gedrag te simuleren. Vandaag de dag zijn er enorm veel handige computersystemen die taal gebruiken, van Google translate op je computer tot autocomplete op je smartphone. Hoewel die systemen enorm succesvol zijn, is het niet duidelijk of ze taal leren en verwerken op de manier waarop mensen dat doen. In het tweede gedeelte van dit hoofdstuk bekijk ik of een artificieel

neuraal netwerk, een veelgebruikt computationeel taalmodel, *tweede blauwe bal* hiërarchisch of lineair interpreteert. De resultaten laten zien dat het model beide interpretaties kan leren, zolang het maar op de juiste manier getraind wordt. Als het alleen maar lineaire antwoorden ziet tijdens de trainingsfase, dan geeft het lineaire antwoorden tijdens de testfase. Ziet het gedurende de trainingsfase alleen hiërarchische antwoorden, dan leert het *tweede blauwe bal* hiërarchisch te interpreteren. Maar als de experimentele trials gedurende de trainingsfase ambigu zijn (voor *tweede blauw bal* is dit het geval wanneer de eerste twee ballen blauw zijn), dan leert het neurale netwerk alleen de lineaire interpretatie. Het lijkt er dus op dat mensen een voorkeur hebben voor de hiërarchische interpretatie, maar dat neurale netwerken taal bij voorkeur lineair interpreteren. Ze kunnen wel leren om hiërarchische antwoorden te geven, maar daarvoor moeten ze expliciet getraind worden met volledig niet-ambigue hiërarchische data.

Hoofdstuk 3 bouwt voort op het idee dat veel huidige artificiële neurale netwerken geen goede modellen van het menselijk taalvermogen zijn. Om dit standpunt te ondersteunen geef ik twee primaire redenen. Aan de ene kant zijn de netwerken vaak niet sterk genoeg, omdat er taalkundige afhankelijkheden zijn die de modellen niet gemakkelijk kunnen leren. Een voorbeeld hiervan zijn coreferentiële relaties, die gaan over zinsstructuren waarin persoonsnamen en persoonlijk voornaamwoorden wél en niet naar elkaar kunnen verwijzen (bijvoorbeeld “Evy denkt dat zij ...” vs. “Zij denkt dat Evy ...”). Aan de andere kant zijn ze juist té sterk, omdat ze in staat zijn onmogelijke talen te leren. Dit zijn artificiële talen met regels die niet voorkomen in natuurlijke talen en door mensen dus nooit geleerd worden. Een voorbeeld hiervan zijn artificiële talen met lineaire regels, die verwijzen naar de ordinale positie van woorden in de zin, in plaats van hiërarchische regels. In dit hoofdstuk bespreek ik twee aanpassingen aan computationele taalmodellen die deze problemen zouden kunnen verhelpen. Aan de ene kant is het van belang dat er meer nadruk wordt gelegd op de betekenis van taal, en niet slechts op de vorm. Veel computationele taalmodellen worden namelijk getraind aan de hand van een corpus van grammaticale zinnen, waardoor ze dus alleen maar weten wat de vorm van taal is. Het is daardoor erg ingewikkeld, zo niet onmogelijk, om betekenis te leren. Aan de andere kant moeten de modellen beperkt worden, zodat ze onmogelijke talen niet makkelijk kunnen leren. Dit kan bijvoorbeeld gedaan worden door een hiërarchische *bias* of *constraint* in het model in te bouwen, zodat het een voorkeur krijgt voor hiërarchische in plaats van lineaire regels. Aan de hand van deze aanpassingen

wordt het mogelijk om computationele modellen te gebruiken om het menselijk taalvermogen beter te begrijpen.

Nadat ik heb laten zien dat hiërarchische structuur een belangrijke rol speelt in taalgebruik, richt ik me in hoofdstukken 4 en 5 op de verwerking van hiërarchische structuur in het brein. Hoofdstuk 4 bouwt voort op voorgaand onderzoek dat heeft laten zien dat, wanneer we naar gesproken zinnen luisteren, onze elektrische hersenactiviteit de structuur van die zinnen volgt. Dit proces wordt *cortical tracking* van syntactische structuur genoemd. In dit hoofdstuk heb ik onderzocht welke factoren van invloed zijn op dit trackingproces. Proefpersonen luisterden naar gesproken stimuli die varieerden in structuur en betekenis. Tegelijkertijd werd hun elektrische hersenactiviteit gemeten aan de hand van elektro-encefalografie (EEG). De resultaten laten zien dat de structuur van gesproken zinnen sterker gevolgd wordt wanneer die zinnen bestaan uit echte woorden in plaats van pseudowoorden. Het trackingproces wordt echter niet beïnvloed door de betekenis van zinnen: de structuur van zinnen werd even sterk gevolgd in reguliere zinnen als in zinnen met een afwijkende betekenis. Dit betekent dat, wanneer we tijdens taalverwerking hiërarchische structuur opbouwen, onze hersenen gevoelig zijn voor de elementen waaruit deze structuur is opgebouwd. Tegelijkertijd laten de resultaten zien dat onze hersenen tijdens dit proces niet beïnvloed worden door de uiteindelijke betekenis van de structuur. Of we nou een reguliere zin verwerken of een afwijkende zin, zoals “een vierkant is een driehoek”, het hersenproces dat structuur opbouwt wordt op dezelfde manier uitgevoerd.

In hoofdstuk 5 maak ik gebruik van magneto-encefalografie (MEG) om verder te onderzoeken hoe die hiërarchische structuur in de hersenen wordt opgebouwd tijdens taalverwerking. Taalverwerking is een incrementeel proces, wat betekent dat mensen continu proberen elk nieuw woord te integreren in de hiërarchische structuur van de zin. Ze wachten dus niet tot het laatste woord voordat ze de zin interpreteren. Maar zelfs als we uitgaan van incrementele verwerking is het nog steeds mogelijk dat mensen (waarschijnlijk onbewust) verschillende verwerkingsstrategieën hanteren. Ze kunnen bijvoorbeeld proberen de zinsstructuur te *voorspellen*. Een andere mogelijkheid is dat ze juist een afwachtende houding aannemen, waarbij ze een woord pas *integreren* in de zinsstructuur wanneer ze zeker weten wat de juiste analyse is. Om dit te onderzoeken, liet ik proefpersonen naar een luisterboek luisteren terwijl hun hersenactiviteit werd gemeten met MEG. Vervolgens berekende ik voor elk woord in het luisterboek twee waarden: de ene waarde was gekoppeld aan de *voorspellende* verwerkingsstrategie, de an-

dere aan de *integrerende* strategie. De voorspellende waarde voor een woord geeft aan hoeveel structuur al voorspeld kan worden op het moment dat dat woord gepresenteerd wordt. De integrerende waarde voor hetzelfde woord geeft aan hoeveel structuur geïntegreerd kan worden op dat moment. Door te kijken naar de correlaties tussen deze waarden en de hersenactiviteit die door die woorden wordt opgewekt, is het mogelijk om te bepalen of mensen een voorspellende of een integrerende verwerkingsstrategie hanteren. De resultaten laten zien dat de voorspellende waarden sterker correleren met hersenactiviteit (voornamelijk in de frontale en temporale kwab in de linkerhersenhelft), wat laat zien dat mensen continu proberen te voorspellen wat de zinsstructuur van de huidige zin gaat zijn. Een belangrijke bevinding van deze studie is dat de resultaten niet geheel overeenkomen met de resultaten van vergelijkbare studies die gedaan zijn in het Engels, wat suggereert dat het type verwerkingsstrategie dat mensen hanteren afhangt van hun moedertaal en/of de taal die ze op dat moment verwerken. De resultaten benadrukken daarom het belang van cross-linguïstisch onderzoek waarbij gekeken wordt naar de verwerking van talen anders dan het Engels. Op die manier is het mogelijk om te bepalen of taalverwerking een universeel proces is, of dat er juist taalspecifieke processen een rol spelen.

Tot slot vergelijk ik in hoofdstuk 6 de structuur van taal met die van acties. In de literatuur wordt vaak beweerd dat taal en acties op elkaar lijken omdat ook acties hiërarchisch georganiseerd zijn. In een sequentie van opeenvolgende handelingen (bijvoorbeeld alle handelingen die gedaan worden bij het zetten van een kopje koffie) kunnen alle bij elkaar horende handelingen gegroepeerd worden in zogenaamde *actieconstituenten* (vergelijkbaar met constituenten die uit woorden bestaan), die vervolgens weer gecombineerd worden met andere handelingen om nog grotere actieconstituenten te maken. Zo ontstaat er een hiërarchische structuur die lijkt op het type boomstructuur dat we kennen uit de taalwetenschap. In dit hoofdstuk laat ik zien dat de structuur van acties andere eigenschappen heeft dan de structuur van zinnen. Het belangrijkste verschil is dat taalstructuur *sterk compositioneel* is, terwijl de structuur van acties *zwak compositioneel* lijkt te zijn. Dat wil zeggen dat het doel van een reeks handelingen slechts afhankelijk is van de volgorde waarin de handelingen worden uitgevoerd (e.g., de koffie wordt in de machine gedaan *voordat* de machine wordt aangezet). De hiërarchische structuur van acties heeft er geen invloed op of het doel van de actie bereikt wordt. Met andere woorden, het maakt niet uit hoe opeenvolgende handelingen hiërarchisch gegroepeerd worden; zolang de volgorde van de handelingen juist is, wordt het doel van de actie bereikt. Dit is duidelijk anders dan

in taal, waarin de hiërarchische structuur van zinnen juist wel van invloed is op de betekenis. Denk bijvoorbeeld aan de structurele ambiguïteit van “de vrouw bekijkt de man met de verrekijker”. Dat de structuur van taal en acties fundamenteel verschillend is, suggereert dat ze op een andere manier in de hersenen worden gerepresenteerd.

Samengevat, in dit proefschrift heb ik de rol van hiërarchische structuur in taalgebruik onderzocht. Ik heb laten zien dat mensen een voorkeur hebben om taal hiërarchisch te interpreteren, en dat zo'n hiërarchische bias waarschijnlijk niet op een natuurlijke manier zal ontstaan in huidige artificiële neurale netwerken. In twee neurowetenschappelijke hoofdstukken laat ik vervolgens zien hoe die hiërarchische structuur in de hersenen wordt opgebouwd. Tot slot laat ik zien dat die structuur taalspecifiek lijkt te zijn. Andere cognitieve systemen die hiërarchisch georganiseerd zijn hebben fundamenteel andere eigenschappen. De hoofdstukken in dit proefschrift laten dus zien dat taalverwerking meer is dan het lineair verwerken van de stroom aan woorden die binnenkomt. Met die woorden wordt een hiërarchische structuur opgebouwd die taalspecifieke eigenschappen heeft en die op een speciale manier in het brein wordt gerepresenteerd.

Research data management

Data availability

Three chapters in this thesis contain experimental data. The behavioral data of Chapter 2 and the EEG data of Chapter 4 were acquired at the Max Planck Institute for Psycholinguistics. Both chapters have been published and their data have been archived at the MPI for Psycholinguistics Archive. Below, I provide the link to the corresponding collections in the MPI Archive. The MEG data of Chapter 5 were acquired at the Donders Centre for Cognitive Neuroimaging. These data are part of a larger project, for which data collection is not yet complete. The data will be archived at the Donders Repository when all data have been acquired. Below, I provide the Data Acquisition Collection (DAC) identifier for this project.

Chapter 2

Code and data of the behavioral experiments are available at: `https://hdl.handle.net/1839/044f49c1-ef24-4e20-a1f7-0a5592ed9b19`
Code and data of the computational simulations are available at: `https://github.com/CasCoopmans/second_blue_ball`

Chapter 4

Code and data of the EEG experiment are available at: `https://hdl.handle.net/1839/aea66e37-587d-426f-859c-59e15d9ca525`

Chapter 5

DAC: `di.dccn.DAC_3027007.01_451`

Acknowledgements

I often hear other PhDs say that doing doctoral research feels like being on a solo journey. I tend to agree with this, but that does mean that I did not get help or support from many important people around me.

Andrea, our meetings rarely go as I think they will – they usually start with some brief, practical updates and before I know it we've strayed into this scientific utopia full of new ideas, exciting research directions, and what-if scenarios. That's great, one might say, but doing a PhD requires translating thoughts into actions, and actions into writing. This, I think, is what I should thank you most for: you encouraged yet challenged my ideas, and you urged me to formulate my thoughts clearly (audience design is key). While doing so, you never lost sight of that dot on the horizon, always reminding me to do those things that will help me achieve my future goals. I feel very fortunate to have worked under your supervision, and am glad that we can continue our successful collaboration. I'm curious to see what's next.

Peter and Helen, thank you for giving me the freedom to pursue my own interests in the past four years. Even though you are both very busy people, your doors were always open when I had questions. **Peter**, it has been a immense pleasure to work in the amazing research environment you have built, with top-notch facilities and brilliant people (including you). I continue to be amazed by your ability to ask exactly the right question after *any* presentation on *any* topic. And I am glad that I can say the same about your response to my ideas, results and writing; while often minimal in quantity, your comments always get right to the heart of the matter. **Helen**, when I sent you a draft of the last chapter of this thesis, your response was as positive as your reply to the PhD proposal I sent you four and a half years earlier. This is something I appreciate in you – your feedback is serious, but it is always formulated in such a way that I feel inspired and motivated to improve my work. Thank you, too, for letting me teach two courses in the Linguistics program. Doing this has made me a better academic.

I would like to thank the members of my manuscript committee, Prof. Herbert Schriefers, Prof. David Poeppel, and Dr. Nina Kazanina, for taking the time to read and evaluate my thesis.

I am immensely grateful to my paranymphs for being good colleagues and even better friends. **Laura**, for the past six and a half years, many of our academic adventures were perfectly aligned. This started with the first course in the CNS master and ended with the submission of our PhD theses. On a daily basis, we talked, gossiped, and complained about science and its people – these interactions were quite often the highlight of my working day. You are an amazing scientist and a wonderful person, and I still think that we should write a paper together (let's be quick, we won't be office mates forever). **Gert-Jan**, I am happy we did our PhDs at the same time. As a psychologist interested in language, it was great that I got work so closely with a linguist doing psychology (please accept this characterization of you). From writing papers to teaching Syntax, it always seems to work out great if we team up (and we have a lot of fun in the process, too!). I hope that we can continue pushing the syntactic agenda of (psycho)linguistic research in Nijmegen.

It is an understatement to say that I learned a lot in the four years of my PhD. However, I felt quite prepared before I even started, and to a large extent that's due my teachers and mentors from a pre-PhD era. I want to single out three people who played an important role. **Mante Nieuwland**, you taught me many things about academia, from analyzing data to revising a manuscript, and from presenting (myself) to dealing with reviewers. As I would be a very different scientist if I hadn't been supervised by Andrea and you, I consider the two of you my academic parents. **Aoju Chen**, thank you for taking me on as a student when I was still doing bachelor in Psychology. You helped me set up my first experiment, encouraged me to present the results at an international conference, and, most importantly, you did not (seem to?) lose hope that the paper would eventually be published. It took us about eight years, three journals, and multiple rounds of revisions to get it out, but I am very happy that we made it work. **Jungmin Kang**, you might not know it, but you played a very important role in me getting to this stage. I thoroughly enjoyed the syntax course you taught during my 2015 exchange stay at WashU in St. Louis. Thanks for being endlessly enthusiastic about complicated syntactic analyses and for assuring me that, yes, syntacticians are indeed crazy. I also want to thank you for being the first person to suggest that I should pursue a PhD in linguistics. While I am not entirely sure that the work in this thesis counts as linguistics, I think it's fair to say that it would contain much less linguistics if you hadn't gotten me excited about syntax.

The MPI for Psycholinguistics is a great place for young scientists. I had the joy of working with some of them very closely. **Karthikeya Kaushik**, your positive spirit is contagious. I really enjoyed the projects we worked on, not in the least because I learned a ton about things I had never even thought about before starting my PhD (there is quite a bit of modeling in this thesis!). Because of the pandemic most of our meetings were via Zoom, which is a pity because sitting down with you over a coffee and thinking about a complex problem is a real treat for the mind. Enjoy sunny Berkeley, I hope to be able to visit you sometime. **Junyuan Zhao**, in our very first meeting we talked about ways to study the neurobiological basis of VP-internal subjects, which is a pretty crazy idea given the state of our experimental methods. In the months following that meeting, you came up with an experiment that was more feasible and eventually led to a wonderful master thesis. If all students are like you, being a supervisor is a piece of cake, so it's only logical that you found a place to continue your academic career. I look forward to seeing the ideas you will come up with in Ann Arbor. **Rowan Sommers**, thank you for disagreeing (at least initially) with basically everything I ever said about language. I am pretty sure that we left many of our disagreements unresolved (shall we agree to disagree, or is that boring?), but your Cartesian doubt has helped me understand what I really think I think. And to be fair, the reason chapter 6 exists at all is largely due to you not being convinced by my claim that language and action are quite different.

Doing psycholinguistics in Nijmegen means regularly interacting with many smart people at the MPI, Donders Institute, and Centre for Language Studies. I want to thank the PhDs and postdocs in the NBL department and the LaCNS and G&C groups for regular meetings full of interesting science. Thanks to **Gert-Jan Schoenmakers**, **Stefan Frank** and **Roel Willems** for the courses we taught together. **Ashley Lewis** and **Yingying Tan**, thank you for a great one-day trip to Tallinn, and for being one of the few NBL members who enjoy talking about football. And thanks to **Kevin Lam** for running the IMPRS so smoothly.

I am grateful to all colleagues who provided help at various stages of the different chapters in this thesis. Thanks to **Stefan Frank** and **Hartmut Fitz** for critically commenting on chapter 2, and to **Charles Yang** and **Bob Berwick** for giving feedback on chapter 3. Bob, this is also the place where I should thank you for letting me use the phrase *triangles in the brain* as the title of my thesis. On chapter 4, thank you, **Greta Kaufeld**, for sharing your knowledge about mutual information analysis, and thanks to **Eva Poort** for recording the stimuli. On chapter 5, **Filiz Tezcan**, thank you for a crash course in TRF analysis at a rather

stressful time when your help was much needed. Thank you, **Sophie Slaats** and **Michelle Suijkerbuijk**, for manually parsing the weirdest sentences into beautiful syntactic structures. **Laura Giglio**, thank you for our endless discussions about syntactic structure building, and for reading the final version of chapter 5. Many thanks to **Giosuè Baggio**, **Cedric Boeckx** and **Bob van Tiel** for commenting on an early version of chapter 6. And thank you, **Fan Bai** and **Shuang Bi**, for making a beautiful cover.

I have many colleagues to thank, but the value of non-academic support should not be underestimated. Thanks to friends from Apeldoorn, Utrecht and elsewhere. You provided the necessary distraction from PhD-related struggles and reminded me, intentionally or not, that there are more important things in life than work. **Mark**, one of the reasons we're such good friends is that we have a lot of things in common. Unfortunately for you, I am just a little bit better at everything (after all, there can only be one number 10). No worries though, finishing this thesis has given me more spare time, which I will devote to teaching you my skills. **Bart**, there are two main things I should thank you for: one is getting me through those 9am MATLAB practicals, the other is making me laugh during 9am Neuroscience classes. For some reason all classes in our bachelor started early in the morning. Thanks, too, for asking me about my work, for reminding me that the academic system is weird, and for making sure that NS trains run on time. I am happy we're still in touch.

Mertuaku, **Yvonne** and **John**, terima kasih banyak telah memberiku rumah kedua and for always showing interest in linguistics and cognitive neuroscience. Being able to explain your work to others is one of the most important aspects of being a scientist. My explanations often leave much to be desired, but I will try to up my game. Thanks to my two **bradders** and **zusje** for support, advice and endless banter. Mom and dad, you always make sure that being in Apeldoorn feels like being ~~in a hotel~~ home. **Mama**, thank you for your unconditional love and support, and for always providing a critical but considerate perspective on the things I do and decide. **Pap**, as my role model in professional and personal life, you taught me many things I know, both about linguistics (e.g., why people can say Chi-*fucking*-cago) and about life in general (e.g., how offside works). I am glad that I can now pay off some of my epistemic debt by teaching you how the human brain works.

And the most important person in my life, **Evy**. Everything in life is easier and more fun when I am with you. You're the best.

Curriculum Vitae

Cas Coopmans was born in Apeldoorn, The Netherlands, in 1994. He obtained his bachelor's degree in Psychology from Utrecht University and his master's degree in Cognitive Neuroscience from Radboud University Nijmegen. In September 2018, Cas started his PhD project at the Max Planck Institute for Psycholinguistics, funded by a four-year PhD Fellowship from the IMPRS for Language Sciences. He is currently a postdoctoral researcher at the Donders Centre for Cognitive Neuroimaging.

Publications

Coopmans, C. W., & Cohn, N. (2022). An electrophysiological investigation of co-referential processes in visual narrative comprehension. *Neuropsychologia, 172*, 108253. doi:10.1016/j.neuropsychologia.2022.108253

Coopmans, C. W., de Hoop, H., Hagoort, P., & Martin, A. E. (2022). Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiology of Language, 3*(3), 386-412. doi:10.1162/nola00070

Coopmans, C. W., de Hoop, H., Kaushik, K., Hagoort, P., & Martin, A. E. (2021). Structure-(in)dependent interpretation of phrases in humans and LSTMs. *The Society for Computation in Linguistics (SCiL) 2021*, pp. 459-463.

Coopmans, C. W., de Hoop, H., Kaushik, K., Hagoort, P., & Martin, A. E. (2022). Hierarchy in language interpretation: evidence from behavioural experiments and computational modelling. *Language, Cognition and Neuroscience, 37*(4), 420-439. doi:10.1080/23273798.2021.1980595

Coopmans, C. W., Kaushik, K., & Martin, A. E. (2023). Hierarchical structure in language and action: A formal comparison. *Psychological Review*.

Coopmans, C. W., & Nieuwland, M. S. (2020). Dissociating activation and integration of discourse referents: Evidence from ERPs and oscillations. *Cortex, 126*, 83–106. doi:10.1016/j.cortex.2019.12.028

Coopmans, C. W. & Schoenmakers, G. T. (2020). Incremental structure building of preverbal PPs in Dutch. *Linguistics in the Netherlands, 37*, 38–52. doi:10.1075/avt.00036.coo

Coopmans, C. W., Struiksma, M., Coopmans, P. H. A., & Chen, A. (2022). Processing of grammatical agreement in the face of variation in lexical stress: A Mismatch Negativity (MMN) study. *Language and Speech*. doi:10.1177/00238309221098116

Nieuwland, M. S., **Coopmans, C. W.**, & Sommers, R. P. (2019). Distinguishing old from new referents during discourse comprehension: Evidence from ERPs and oscillations. *Frontiers in Human Neuroscience, 13*. doi:10.3389/fnhum.2019.00398

MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

VISITING ADDRESS
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

POSTAL ADDRESS
P.O. Box 310
6500 AH Nijmegen
The Netherlands

CONTACT
T +31(0)24 3521 911
F +31(0)24 3521 213
E info@mpi.nl
Twitter @MPI_NL
www.mpi.nl

CLS | Centre for Language Studies
Radboud University