# The Role of Auditory Perceptual Gestalts on the Processing of Phrase Structure

Antony Scott Trotter

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Lancaster University

Department of Psychology

February 2019

## Table of Contents

# Declaration

I declare that this thesis is my own work, completed solely by the author under the supervision of Professors Padraic Monaghan, and Gert Westermann, and that it has not been submitted in substantially the same for the award of a higher degree elsewhere.

# List of Tables

**Chapter 2**

**Chapter 3**

**Chapter 4**

**Chapter 5**

**Chapter 6**

## List of Figures

**Chapter 2**

**Chapter 3**

**Chapter 4**

**Chapter 5**

**Chapter 6**

Figure 10. Model estimates of the likelihood of fixating the target over the analysis region.

Figure 11. Mean proportion of looks to target by prosodic condition, for passive structures.

Figure 12. Model estimates of the likelihood of fixating the target over the analysis region.

Figure 13. Model estimates of the likelihood of fixating the target over the analysis region.

# List of Abbreviations

| | |
|---|---|
| AGL | Artificial grammar learning |
| FSG | Finite state grammar |
| G | Grammatical |
| HCE | Hierarchical centre-embedded structures |
| LoE | Levels of embedding |
| PSG | Phrase structure grammar |
| U | Ungrammatical |
| VWP | Visual world paradigm |

# Acknowledgements

I would firstly like to thank my two supervisors, Professors Padraic Monaghan and Gert Westermann for their support, guidance, and many stimulating conversations over the years. I also thank my lab group, who were always willing to chip in to solve any design or analysis problems, and who were patient when I raised the minutiae of some statistics or why violin plots are simply wonderful (much to Padraic’s vexation). Though we might be scattered across a vast swathe of the world now, it will always be wonderful to catch up. A special thank you goes to Rebecca Frost, who’s diligent editing has been immeasurably helpful, and who’s plucky, tireless attitude never fails to inspire. James Brand also deserves an honourable mention. Without his help, many an R script would be far, far more clunky, and his academic curiosity led to some wonderful interrogations of theory. Plus, he once stopped a chap from pick-pocketing me in Brussels.

I also want to thank my friends, particularly David Elliott and Rebecca Iversen. David gets me thinking about complex methods and new skills, while Rebecca keeps me grounded, and reminds me that not everyone understands me when I speak a million miles an hour about my work. I think it’s safe to say I wouldn’t have made it through writing up without my partner, Shal. Tirelessly supportive, and ever patient, she deserves the biggest thanks. So, Shal: Thanks, you’re a diamond!

Finally, I would like to thank my family. Without their help and support I couldn’t have gotten where I am today. Mum was always there to patiently listen to my academic conniptions, even if she didn’t always understand what the issue was. Dad worked hard to support me over the years, and to him I wish a happy retirement! Finally, a special thank you goes to my brother, Bobby, who believes in me, even when I don’t.

# Abstract

Hierarchical centre embeddings (HCEs) in natural language have been taken as evidence that language is not processed as a finite state system (Chomsky, 1957). While phrase structure may be necessary to produce HCEs, finite state, sequential processing may underlie their comprehension (Frank, Bod, & Christiansen, 2012). Under this account, listeners employ surface level cues (e.g. semantic content) to determine the dependencies within an utterance, instead of processing the words in a hierarchy. The acoustic structure of speech reflects the speaker's syntactic representation during production (Cooper, Paccia & Lapointe, 1978). In comprehension, temporal (Snedeker & Trueswell, 2003) and pitch (Watson, Tanenhaus, & Gunlogson, 2008) cues rapidly influence processing. Therefore, temporal and pitch variation in speech could contain cues to dependencies. We examine whether grouping behaviour may be driven by Gestalt principles. Temporal proximity suggests that individuals group sequential words that occur closer together in time. Pitch similarity states that individuals group sequential words that are similar in pitch. In this thesis, I examine whether these Gestalts support dependency detection in speech, providing a mechanism through which hierarchical structure can be processed non-hierarchically.

In Chapter 3, we assessed whether temporal proximity and pitch similarity explicitly relate to the structure of a corpus of spontaneously produced active and passive relative clauses. This was the case for actives; the embedded clause was preceded by a lengthened pause and a large pitch reduction. For passives, a longer pause and pitch reduction occurred after the verb-phrase of the embedded clause, counter to prediction. The results for actives suggest that temporal proximity and pitch similarity cues could be used to group the phrases of the embedded clause, obviating the need to process hierarchically structured speech hierarchically.

Two artificial grammar learning studies assessed whether pitch similarity and temporal proximity cues support the acquisition of phrase structure grammar. Chapter 4 emphasised temporal proximity cues, while chapter 5 emphasised pitch similarity cues. In Chapter 5, pitch similarity cues improved classification performance for structures with two levels of embedding. In both, participants did not benefit from temporal proximity cues. However, the results of a cross-species meta-analysis of artificial grammar learning studies (Chapter 2)  raised the possibility that reflection-based measures (e.g. grammaticality judgements) are not well suited for assessing processing-based learning, such as online speech processing (Christiansen, 2018). To properly assess the role of Gestalt cues in speech processing therefore requires processing-based measures.

To assess the influence of auditory Gestalts on online speech processing, in Chapter 6 we analysed participants' gaze behaviour in response to pitch similarity and temporal proximity cues using the visual world paradigm. Participants heard speech-synthesised active-object and passive relative clauses, whilst viewing four potential targets. Each sentence had a prosodic structure consistent with either syntactic form (Chapter 3), or two control prosodic structures. Pitch similarity results indicated that these cues facilitated processing. Temporal proximity cues consistent with syntactic structure did not facilitate processing, instead results suggested a general benefit of increased processing time.

Overall, these studies suggest that participants can use the pitch similarity Gestalt to group together syntactically dependent phrases in hierarchical speech, offering a mechanism through which individuals could process hierarchical structures non-hierarchically. The results of Chapters 4, 5, and 6 suggest temporal proximity cues did not facilitate performance to the same extent. Thus, we suggest that unfilled pauses in isolation may be insufficient to facilitate groupings on the basis of temporal proximity.

# 1. Introduction

## 1.1 The theoretical importance of hierarchical structure

Determining the mechanisms that underpin human language processing remains a key focus of psycholinguistic research. Chomsky (1957; 1959) proposed a generative hierarchy of rule systems capable of generating an infinite number of sequences by defining increasing constraints on possible structures. At the lowest level of this hierarchy lie finite state grammars. Finite state grammar sequences can be fully specified by transitional probabilities between a finite number of states (Hauser & Fitch, 2004). The resulting sequences require only a large enough memory stack to hold sequential states, and the transitions between them, in order to concatenate them into longer sequences. An example of a finite state grammar sequences would be, "[$_{C-1}$ Rebecca cooks,] [$_{C-2}$ Dave cleans]". Phrase structure grammars are the next level of the hierarchy. Much like finite state grammars they can concatenate items. Crucially, it has recently been proposed that through the recursive application of the *merge* operation (Chomsky, 1995), phrase structure grammars can embed strings within other strings, resulting in phrase structures and long-distance dependencies, e.g. "[$_{C-1}$ The man [$_{C-2}$ the doctor treated] was in pain]". The mechanisms required to generate and process these complex phrase structures are more sophisticated, requiring an open-ended memory system, in addition to the perceptual mechanisms necessary to recognise them (Hauser & Fitch, 2004). Crucially, phrase structure grammars (Chomsky, 1959), and more recently, *merge* (Berwick & Chomsky, 2016), have been suggested as the defining characteristic of human language, allowing for an infinite number of meaning to be expressed with a finite number of word (Hauser, Chomsky, & Fitch, 2002). Probing the processing differences between sequences generated by these two grammars provides insights into the mechanisms of human language processing.

## 1.2 Hierarchical centre-embedded structures in comprehension

The existence of hierarchical centre-embeddings (HCE) in natural language has been taken as evidence that language is not a finite state system. However, whilst individuals can process hierarchical-centre embeddings, their ability to do so is limited. HCEs with more than three levels of embedding are challenging to process, even for proficient, native speakers (e.g. e.g. Bach, Brown, & Marslen-Wilson, 1986; Hudson, 1996; Newmeyer, 1988). This statute from s.1 of the British Road Traffic Act (1972) is a fairly typical example of a three level of embedding construct; "A person [$_{\text{C-1}}$ who, [$_{\text{C-2}}$ when riding a cycle, [$_{\text{C-3}}$ not being a motor vehicle,] on a road or other public place,] is unfit to ride through drink or drugs,] shall be guilty of an offence." As more clauses are embedded, the distance between dependent constituents increases, together with the difficulty of relating them to one another (Lai & Poletiek, 2011). Recent computational work reinforces this idea. In a large sample of syntactic trees taken from the Prague and Stanford corpora, Ferrer-i-Cancho and Gomez-Rodriquez (2016) found a positive correlation between the sum of dependency lengths in a sentence, and the number of crossings a sentence would contain; as dependency lengths increase, the amount of centre-embeddings produced becomes less frequent. Taken together, this suggests that humans' capacity to generate and process hierarchical structure is limited.

Whilst it remains a contentious topic, there are several mechanisms that may explain processing limitations on hierarchical structure. One intuitive constraint is limited working memory (WM) (Karlsson, 2010). As more embeddings are included in a sentence, the more information needs to be stored. WM cost, here, is specifically quantified in terms of the number of syntactic categories that are necessary to complete the current input as a grammatical sentence (Gibson, 1998). Maintaining open syntactic dependencies increases WM cost, therefore, increased distance between the start and end of a dependency results in increased maintenance costs, effectively setting a resource-based cap on the maximum number of embeddings a sentence can contain. WM costs have been used to explain the missing verb effect, whereby centre-embeddings lacking a verb-phrase are viewed as acceptable (e.g. "The patient who the nurse who the clinic hired met jack") (Gibson & Thomas, 1999). However, this effect is not universal; the missing verb effect is not present in speakers of German (Vasishth et al., 2010) and Dutch (Frank et al., 2015), who find the grammatical versions of these sentences easier to process. Verb-final constructions are common in German and Dutch, and require the listener to track dependency relations over long distances, suggesting that experience results in language-specific processing improvements (Christiansen & Chater, 2015). Usage-based accounts (Christiansen & Chater, 2015) could thus account for the difficulties English speakers face with centre-embedded structures.

### 1.3 Hierarchical centre-embedded structures in production

Given that English sentences with hierarchical centre-embeddings tend to be avoided in production, and are challenging for perception, a key question is what leads speakers to produce these kinds of structures when they do occur? Montag and MacDonald (2014) utilised a picture description task to assess the influence of visual competition and animacy on relative clause production. In this study, participants were presented with 20 scenes, and were required to answer probe questions relating to particular objects in each scene. Scenes contained two competing depictions of  events involving the same action, and two competing depictions of the target object, one animate, one inanimate. For example, a scene might depict a nursery room, with a girl hugging a man wearing a green jacket, another girl hugging a white bear, a brown bear on a table, and a man wearing a brown shirt reading in the background. Participants responded to probe questions (e.g. animate-target, "Which man is wearing green?"; inanimate-target "Which bear is white?"). Participants were more likely to produce active-object relative clauses when asked to describe an animate agent interacting with an inanimate object (e.g. "The bear the girl is hugging is brown), and more likely to produce a passive relative clause when an animate agent interacts with an animate target (e.g. "The man being hugged by the girl is wearing green"). This suggests that, first, visual competition biases production towards relative clause structures, and that animacy factors create a bias towards active-object structures.

### 1.4 Sequential processing accounts

Whilst a sentence may possess hierarchical structure, it remains a contentious question as to whether it is processed hierarchically. Recent computational work has suggested that hierarchical structure may not play a role in generating linguistic expectations. Frank and Bod (2011) compared reading-time measurements (generated with the eye-tracking data of 10 participants) from the Dundee corpus (comprised of 2368 sentences) against word-probability estimates generated by three kinds of probabilistic language models containing different psychological mechanisms and representations. The first class were phrase structure grammar models, which employed hierarchical structure induced from syntactic trees. The second (Markov models) and third (echo state networks) class of models had access only to sequential structure. The phrase structure grammar models failed to estimate variance in reading time data over and above all sequential structure models, suggesting that hierarchical sentence structure did not effectively predict response times associated with the generation of expectations about upcoming words.

Recently, Frank, Bod, and Christiansen (2012) have proposed that the comprehension of hierarchical structure is potentially achieved through sequential processing. Under this account, to comprehend a hierarchical structure in a rapidly unfolding temporal context (i.e. incoming speech), the listener would need to rely on superficial surface level cues to determine the dependencies within an utterance, instead of processing the incoming sequence of words hierarchically, in accord with the structure of the sentence. Returning to our example from the British Road Traffic Act (1972), world knowledge can be used to determine the dependencies. Bicycles can neither be guilty of a criminal offence, nor can they be inebriated, but they are ridden on roads and in public spaces. On the other hand, humans ride bikes, imbibe drink and drugs, feel their effects, and can be guilty of offences. Through semantic knowledge, then, you can generate the following units, "person riding bike" "(if) bike is on road", "(if) person drunk", "(then) person is guilty".

## 1.5 The relationship between syntax and prosody

The sequential processing account (Frank et al..2012) suggests that sentence processing can proceed by using salient, low-level cues to group phrasal units. That is, in addition to semantic cues, the acoustic structure of speech may provide a mechanism through which participants can form an initial parse of hierarchical structure. Prosody – the rhythmic and melodic features of speech – has been shown to relate to syntactic structure. For instance, in English, clauses are often cued with phrase-initial pitch-resetting, phrase-final declining pitch contour (Pierrehumber, 1979), increased duration on phrase final words (Langus, Marchetto, Hoffman, Bion, & Nespor, 2012), as well as increased duration syllable-finally within words (Cooper, Paccia, & Lapointe, 1978). Based on studies of language production, it has been suggested that the tonal and temporal structure of utterances is generated by the speaker's syntactic representation.

### 1.5.1 Syntax and prosody in production

Cooper and Sorensen (1977) demonstrated that the pitch dynamics of a sentence reflect its hierarchical structure. In this study, speakers were presented with sentences to read aloud, and were first asked to read them silently, in isolation, and consider their semantic interpretation. Speakers then practiced the sentence aloud, receiving corrective feedback about their stress pattern, before reading the sentence aloud again for recoding. In experiment 1, the materials were sentences that either included two main clauses ("[$_{C\text{-}1}$ Marie was listening to the song] [$_{C\text{-}2}$ and Del was playing]"), or a main clause and an embedded clause ("[$_{C\text{-}1}$ Marie was listening to the song [$_{C\text{-}2}$ Adelle was playing]]"), matched for total number of syllables and approximate stress contour. In the former, the internal syntactic boundary is between the end of the first, main clause, and the onset of the second. In the latter, it is only the onset of the embedded clause. The authors measured the peak $F_0$ value in *song* ($P_1$), the lowest value in the same syllable (V), and the peak value in the stressed syllable (*Del/Delle*) following the boundary ($P_2$). Critically, there was a pitch difference between the sentences; sentences with two conjoined clauses had a larger $F_0$ reduction between $P_1$ and V, and a larger subsequent increase between V and $P_2$, relative to sentences with an embedding. As the sentences were matched on stress pattern and phonological environment, it suggested that the differences in pitch variation are syntactically driven.

This theory extends to durational cues at syntactic boundaries. Cooper, Paccia, and Lapointe (1978) conducted a series of experiments to study the influence of several syntactic ambiguities on durational cues produced by speakers. The authors utilized several ambiguities that arise due to hierarchical structure, for example, in experiment 5, the sentence "Pam asked the cop who Jake confronted" can have two interpretations; (a) "who did Jake confront?", and (b) "which cop? The cop that Jake confronted?". When cop occurs in the indirect question interpretation, it is produced at the third level of the syntactic hierarchy; it is a noun-phrase, nested within a prepositional phrase, which in turn is nested within the complex verb-phrase. When cop occurs as part of a relative clause interpretation, it is only at the second level of the hierarchy; a noun, occurring within a complex noun-phrase. The results – across six ambiguities – demonstrated that when the critical syllable (e.g. */ka/* in "cop") occurred at a deeper level of the syntactic hierarchy, participants lengthened the syllable more, and paused for a longer duration following the boundary final word.

Whilst these observations suggest that prosody is mapped onto syntax during production, it is not particularly clear as to how this is achieved. In the sequential processing account proposed by Frank, Bod, and Christiansen (2012), in production, utterances are generated from constructions (Construction Grammar, Goldberg, 2006), which are linguistic forms paired with meanings. At the simplest level, these constructions are individual word-meaning pairs, e.g. a noun (brush), combined with its mental representation. Constructions can also be comprised of multiple words (e.g. dustpan and brush), where a frequently occurring word sequence can become merged into a novel construction. Constructions can contain abstract elements that openly correspond to noun phrases, e.g. “pick X up”. Building a sentence thus corresponds to creating a sequence out of these constructions. Frank, Bod, and Christiansen (2012) suggest this occurs by switching between multiple sequential streams that run in parallel, where one stream may contain put x down, a second knife and fork – corresponding to x – and a third, including your, the combination of which results in put your knife and fork down. Constructing utterances would involve generating phrases through this process and combining them.

Given a sequential processing account, prosodic cues would be inserted over chunks. Pitch-declination would occur naturally over chunks, and resetting at the start of new chunks, giving rise to fall-rise patterns that appear useful for inferring syntactic constituency (Gleitman & Wanner, 1982; Morgan, 1986; Peters, 1983). Similarly, pauses would be produced at the end of these chunks, potentially reflecting the patterns outline by Cooper, Paccia, and Lapointe (1978) and Cooper and Sorensen (1977). Crucially, however, prosody may additionally provide a useful signal for speech-error-detection. Under production-based speech error detection accounts (e.g. Nozari, Dell, & Schwartz, 2011), an important source of information is the perceptual-loop, i.e. detecting errors using the sensory processing of your own speech. In section 1.6.4, we outline evidence suggesting that the human auditory cortex is tonotopically organized (Pantev, Hoke, Lehnertz, Lutkenhoner, Anogianakis, & Wittkowski, 1988; Elberling, Bak, Kofoed, Lebech, & Saermark, 1982; Tiitinen, Alho, Huotilainen, Ilmoniemi, Simola, & Naatanen, 1993; Yamamoto, Uemaura, Llinas, 1992; Yamamoto, Williamsen, Kaufman, Nicholson, Llinas, 1988; Bertrand, Perrin, Pernier, 1991), and that hemispheric dominance drives the temporal and spectral aspects of speech (Flinker, Doyle, Mehta, Devinsky, & Poeppel, 2019). Given that this is the case, prosodic cues may help to drive the detection of syntactic errors in speech; if individuals do not produce grouping cues, such as pitch-resetting, unfilled pauses, or final-lengthening, it is plausible that this will provide a useful error detection cue. While this may be the case, at present, this assumption has not been empirically tested, and should only be assumed with caution.

If prosodic cues are generated during speech production, then it follows that they may have some kind of functional purpose in language processing. Indeed, the literature has demonstrated their utility in language comprehension, and the speed at which acoustic structure becomes available to listeners.

### 1.5.2 Syntax and prosody in comprehension

Watson, Tanenhaus, and Gunlogson (2008) demonstrated that pitch cues rapidly affect listeners' linguistic expectations. The question of interest was whether pitch accents on critical vowels in phonological competitors (e.g. *camel*/*candle*) would bias fixations towards a new item in the discourse, or a previously mentioned item. Participants were given a series of instructions (e.g. "Click on the camel and the dog. Move the dog to the right of the square. Now, move the *camel*/*candle* below the triangle.") to perform on a visual display comprised of eight objects; four shapes, and four objects, two of which were in the same phonological cohort (e.g. *camel*/*candle*). In the final command, the underlined vowel in the critical word would either rise to the speaker's maximum pitch (H*), or initially drop, followed by a subsequent increase in pitch (L + H*). The authors aimed to assess whether the H* accent was used by speakers when introducing a new item (discourse new, e.g. *candle*), and the L+H* accent when the speaker intends to contrast a previously mentioned item with a salient alternative (contrast, e.g. *camel*). The results indicated that participants were rapidly able to use the pitch accent to direct their eye movements; the L + H* accent increased fixations to contrast members (*camel*). H* accented vowels increased fixations to all potential referents with names consistent with the input, regardless of whether they were previously mentioned (*camel*), or new to the discourse (*candle*). This provides evidence that listeners can use pitch cues present in the input to form expectations about upcoming material.

Snedeker and Trueswell (2003) showed that listeners could rapidly use temporal cues to disambiguate temporarily ambiguous syntactic structures. In their study, speakers had to instruct a listener to perform an action on an array of objects in front of them (e.g. "Tap the frog with the flower"). The array contained an instrument-holding animal (e.g. a frog holding a flower), separate instances of the animal (e.g. a frog) and instrument (e.g. a large flower), and distractor items. Speakers were given a modifier (the experimenter picks up the instrument, and touches the animal) or instrument (the experimenter touches the instrument-holding animal) demonstration of the action. This affected the way speakers produced the instruction. In the modifier condition, speakers paused for a longer duration following "Tap". In the instrument condition, they paused for a shorter duration following "Tap", lengthened "frog", and paused for a longer duration between "frog" and the by-phrase. Listeners' gaze behaviour was analysed in two critical regions: 200 – 500ms after the onset of the direct object noun, and 200 – 800ms after the onset of the prepositional object. During the direct object noun, instrument prosody produced equal looks to both frogs. Modifier prosody biased gaze towards the frog holding the flower. In the prepositional object region, instrument prosody resulted in more looks to the flower, while modifier prosody elicited more looks to the frog with the flower. These results suggest that participants can use pause (and other durational) cues to eliminate competitors during an unfolding utterance. Taken together, these studies suggest that while tonal and temporal cues may result from production processes, they are salient, and rapidly affect comprehension processes.

Given the sequential processing account, speech comprehension can be conceptualised as rapidly computing dependencies between sequential units utilising low-level cues. Prosody provides a plausible mechanism with which to do so. The human auditory cortex is tonotopically organised (Pantev, Hoke, Lehnertz, Lutkenhoner, Anogianakis, & Wittkowski, 1988; Elberling, Bak, Kofoed, Lebech, & Saermark, 1982; Tiitinen, Alho, Huotilainen, Ilmoniemi, Simola, & Naatanen, 1993; Yamamoto, Uemaura, Llinas, 1992; Yamamoto, Williamsen, Kaufman, Nicholson, Llinas, 1988; Bertrand, Perrin, Pernier, 1991), and processing of the spectral and temporal aspects is underpinned by specialised hemispheric processing (Flinker, Doyle, Mehta, Devinsky, & Poeppel, 2019): Low-level grouping cues in speech are easily detectable in primary auditory areas of the cortex. Changes in pitch, such as those noted in Cooper and Sorensen (1977) will be processed in adjacent, fine-tuned areas of the auditory cortex, which in turn will project to the language processing network. The right superior temporal gyrus (Flinker, Doyle, Mehta, Devinsky, & Poeppel, 2019) is uniquely sensitive to the temporal aspects of speech, providing a potential mechanism to detect temporal cues to clausal boundaries. Provided prosodic cues co-occur with syntactic boundaries, or sequential chunks, bottom-up information from the speech signal will provide useful information about their onset and closure.

### 1.5.3 Prosody and the acquisition of syntax

The relationship between prosodic and syntactic structure has led to the proposal of the prosodic bootstrapping hypothesis (Gleitman & Wanner, 1982; Morgan, 1986; Peters, 1983): Infants draw on the prosodic information contained in speech to help identify word-, phrase-, and clause-boundaries, and to help infer constituency and hierarchical syntactic structure. This hypothesis has received much support. Nazzi, Kemler Nelson, Jusczyk, and Jusczyk (2000) tested 6-month-old infants' ability to utilise prosodic cues present at clausal boundaries using a Head-turn Preference Procedure. The familiarization sentences were extracted from passages that were read aloud. Sentences were either well-formed, entire utterances ("Leafy vegetables taste so good") or ill-formed and made up of two distinct utterances ("…leafy vegetables. Taste so good"), and thus contained a syntactic and prosodic boundary. At test, infants heard both passages containing the familiarization sentences. Infants looked longer to the passage containing the well-formed sentence. The infants also listened significantly longer to novel well-formed test stimuli than novel ill-formed sequences taken from new passages, demonstrating that 6-month-olds infants can recognize prosodic cues consistent with syntactic boundaries.

Similarly, Juczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, and Piwoz (1992) tested whether 9-month-old infants prefer to listen to speech that contains pauses consistent with phrase boundaries (e.g. "*What happened? Did you / spill your cereal*") over passages containing pauses that occur elsewhere in the sentence (e.g. "*What happened? Did you spill / your cereal?*"). The results demonstrated infants preferentially attended to versions consistent with syntactic boundaries, providing evidence that 9-month-olds are sensitive to temporal makers of clauses. In line with the prosodic bootstrapping hypothesis, these findings suggest that infants are sensitive to prosodic boundary cues, resulting in a preference for syntactically well-formed groupings. Adult learners can also benefit from prosodic cues that reinforce syntactic structure.

Prosodic cues taken from learners' native and non-native language have been shown to facilitate the acquisition of hierarchical structure. Langus, Marchetto, Hoffman, Bion, and Nespor (2012) conducted an artificial grammar learning study in which grammatical sequences comprised two clauses. Clauses were cued using final-syllable lengthening, and sentences were cued with a descending pitch contour, and transitional probabilities favoured adjacent dependencies. Critically, participants were either trained with prosodic cues modelled on their native (Italian) or a non-native language (Japanese). Learning was assessed using a two-alternative forced choice task. Regardless of whether participants received native or non-native prosodic cues, they chose novel rule-phrases and sentences over part-phrases, suggesting that participants relied on prosodic over distributional cues. Notably, the results suggest that prosodic cues are salient to participants, whether they reflect their prior language experience or not.

Overall, the finding that prosody is useful in acquisition may reflect the underlying neural architecture of the auditory system. As mentioned above, the human auditory cortex is organised tonotopically, allowing pitch changes to be detected in the primary auditory cortex, hence the rise-fall pattern of pitch prosody over syntactic structures will be naturally salient; small changes will make use of adjacent areas of the auditory cortex, large changes will make use of spatially distant areas of the auditory cortex. Thus, to the language acquiring brain, these differences will be easily detected, and salient. This should occur regardless of language-specific variation, potentially explaining the findings of Langus et a. (2012). As a result, it seems reasonable to assume that these bottom-up processing biases will enable infants to decompose the complex acoustic structure of speech into syntax-like units. Whilst not perfectly reliable, as prosodic structure may rely on factors other than syntax (Kraljic & Brennan, 2005), it will permit a rudimentary chunking, and allow infants to recognize proper linguistic units, explaining the findings of Nazzi et al. (2000) and Juczyk et al. (1992).

### 1.5.4. Conclusion

In sum, research indicates that prosodic cues rapidly affect processing, are highly salient across development, and result from the speaker's representation of the utterance's syntactic structure. Under sequential processing accounts, individuals use salient, superficial cues to determine the dependencies of incoming speech, instead of processing incoming words in a hierarchy. This suggests that prosodic cues provide a plausible mechanism through which sequential processing could be achieved. Further, if speakers can employ both native and non-native prosody to support processing, this raises the question of whether domain-general auditory processing behaviours may support the processing of prosodic structure.

In the preceding sections we have also reviewed evidence suggesting that prosodic processing may be underpinned by the neural architecture of the auditory system, explaining the ability of infants to process basic prosodic groupings. During production planning, if groupings are computed sequentially, prosodic boundaries may inserted at the end of syntactic chunks, and if so, may provide useful information for speech error detection; if participants fail to signal the end of a syntactic grouping, low-level perceptual information will allow individuals to rapidly detect, and subsequently correct their speech. In adult comprehension, sensitivity to acoustic grouping cues will allow participants to rapidly detect prosodic (and therefore syntactic) boundaries utilising bottom-up information. A question remains, however, as to how these three aspects of prosody interact.

Christiansen and Chater (2016) note that the cultural transmission of language enforce an iterative relationship between the three. If a given prosodic cue is useful in language acquisition, it is therefore useful for comprehension. If a cue is useful in comprehension, it will be present in production. In this case, it will be used by a subsequent generation when teaching the subsequent generation. Thus, while this is likely to produce linguistic variation, it makes it therefore more likely that linguistic features driven by domain-general processes (such as underlying auditory processing biases) will be more salient, and robust to change (Christiansen & Chater, 2016). Thus, if prosodic structure is driven by low-level computations rooted in the architecture of the auditory system, they are likely to be salient, robust, and useful for language acquisition.

## 1.6 Domain-general structural processing

It is important to consider the processing of musical structure in comparison to language, due to several similar features. Musical and linguistic structure are similarly organised; speech and music are both auditory stimuli that adhere to hierarchical structural rules (Zhang, Jiang, Zhou, & Yang, 2016), however music does not have a formal semantics. Speech and music also have similar structural acoustic cues; both are grouped into phrases marked by pauses, differences in tone height, and the durations of beats or syllables (Patel, 2003). Due to theses similarities, several authors have proposed that shared perceptual or cognitive mechanisms are recruited in the acquisition (McMullen & Saffran, 2004) or processing (Patel & Iversen, 2007) of musical and linguistic structure. The question of interest therefore becomes; how do listeners group speech and musical stimuli into coherent sub-sequences?

### 1.6.1 The Gestalt principle of pitch similarity

Two domain general strategies for grouping auditory-perceptual information (Gestalts, or grouping rules) are particularly relevant for the present thesis; pitch similarity and temporal proximity. The pitch similarity Gestalt states that individuals form sequential links between tones that are close in pitch, and to distinguish between those that are further apart (Deutsch, 2013). Miller and Heise (1950) provide an excellent demonstration of this effect. Participants were presented with two tone frequencies (A and B), delivered at a rate of 10 tones per second in an ABAB pattern. When the frequency difference between A and B was small, participants perceived the sequence as a trill; a single percept. When there was a large tone difference, participants perceived two interrupted and unrelated tones; i.e. two unique acoustic structures. This provides a demonstration that pitch similarity can be a powerful factor in the context of grouping acoustic structures.

### 1.6.2 The Gestalt principle of temporal proximity

The temporal proximity Gestalt states that individuals form sequential links between tones that occur closer in time, and distinguish between those that are further apart (Deutsch, 2013). Lerdahl and Jackendoff (1983) proposed that musical grouping boundaries are placed at longer intervals between note onsets, and at changes in values of attributes including the pitch range. Deliège (1987) demonstrated this by presenting participants with Western classical music, and asking them to mark boundaries between musical groupings. The boundaries participants chose corresponded with Lerdahl and Jackendoff’s (1983) grouping cues; the strongest effects were present following long notes (i.e. iambic groupings), with changes in timbre and dynamics also exerting influence. This, in turn raises the question of how tonal and temporal cues interact during auditory perception.

### 1.6.3 The interaction of pitch similarity and temporal proximity

Hamaoui and Deutsch (2010) assessed the interplay of temporal proximity and pitch similarity in a grouping preference study, in which the two Gestalt mechanisms were in conflict. Participants were presented with twelve-tone sequences, where pitch similarity suggested four groups of three tones, and temporal proximity suggested three groups of four tones. Tonal groups were separated by two, five or eleven semitones, and temporal groups were separated by pauses 15, 30, 45, or 60ms. With increased tonal differences, participants grouped the sequences based on pitch similarity. However, participants were more likely to rely on temporal cues with pause durations over 30ms, even with large pitch differences. In a second experiment in this study, Hamaoui and Deutsch (2010) presented new sequences, where the tones could be either hierarchically structured or unstructured but were otherwise matched in pitch. Here, participants grouped sub-sequences on the basis of the hierarchical pitch structure, and these groupings were more robust to conflicting temporal cues than unstructured sequences. This suggests top-down preferences based on experience influence grouping behaviour.

### 1.6.4 The neural bases of auditory Gestalt perception

The prior sections have provided evidence that Gestalt processes affect auditory processing. However, these sections are agnostic towards the underlying mechanisms producing these behaviours. First, it is notable that several studies assessing auditory-evoked potentials with electroencephalography (EEG) and magnetoencephalography (MEG) have suggested that the human auditory cortex is tonotopically organised (Pantev, Hoke, Lehnertz, Lutkenhoner, Anogianakis, & Wittkowski, 1988; Elberling, Bak, Kofoed, Lebech, & Saermark, 1982; Tiitinen, Alho, Huotilainen, Ilmoniemi, Simola, & Naatanen, 1993; Yamamoto, Uemaura, Llinas, 1992; Yamamoto, Williamsen, Kaufman, Nicholson, Llinas, 1988; Bertrand, Perrin, Pernier, 1991), with a posteromedial to anterolateral representation of increasing sound frequencies. A highly detailed, single-unit level examination of tonotopic mappings has been conducted in epileptic patients using implanted microelectrodes (Howard, Volkov, Abbas, Damasio, Ollendieck, & Granner, 1996). Each unit corresponds to an electrode site, and the authors discovered that the units responding to sounds exhibited a frequency-dependent response pattern. Around three quarters of the units generated finely tuned, frequency-related excitatory responses, while the remaining quarter exhibited large receptive fields, and excitatory responses to nearly the whole range of frequencies. Whilst it typically quite difficult to conduct detailed analyses of the neural correlates of auditory processing with functional magnetic resonance imaging (fMRI) due to its inherently noisy scanning environment, several studies have supported the tonotopic organisation of the auditory cortex (Bilcecen, Scheffler, Schmid, Tschopp, & Seelig, 1998; Wessinger, Buonocore, Kussmaul, & Mangun, 1997; Lantos, Liu, Shafer, Knuth, & Vaughan, 1997; Strainer, Ulmer, Yetkin, Haughton, Daniels, & Millen, 1997; Talavage, Ledden, Sereno, Rosen, & Dale, 1997; Talavage, Benson, Galaburda, & Rosen, 1996; Yang, Engelien, Engelien, Xu, Stern, & Silbersweig, 2000). In addition to this tonal sensitivity, recent MEG evidence has suggested a left-hemisphere dominant sensitivity to temporal modulations in the superior temporal gyrus (Flinker, Doyle, Mehta, Devinsky, & Poeppel, 2019). While these studies suggest the neural architecture that may permit allow Gestalt processing of auditory stimuli, they do not provide direct evidence for a mechanism to do so.

Recently, using EEG, Costa-Faidella, Sussman and Escera (2017) found evidence suggesting that auditory grouping behaviour may be driven by attentional processes. In their study participants were required to attend to a melody embedded within a longer sequence of tones, and judge their duration, and count the number of tones comprising the sequence. In other words, in ambiguous tone sequences, participants were required to attend to a target melody, and ignore others. The results indicated that brain oscillations concurrently entrained to the rate of all competing sound patterns. Critically, however, entrainment to the ignored sequence was restricted to auditory regions, whilst entrainment to the attended sequence was spread across the auditory-motor network. In other words, the entrainment was gated based on participants attention. This may suggest that auditory grouping behaviour is reliant upon task or environmentally specific demands. This observation is backed up by research in animals (Kuchibhotla & Batherllier, 2018), wherein cognitively demanding tasks can both suppress, and facilitate auditory cortical responses (Kuchibhotla, Gill, Lindsay, Papdoyannis, Field, Sten, Miller, & Froemke, 2017; Carcea, Insanally, & Froemke, 2017; Rodgers & DeWeese, 2014; Runyan, Piasini, Panzeri, & Harvey, 2017): Anticipatory top-down inputs from the prefrontal cortex prepares the auditory cortex to receive incoming sensory information based on behavioural conditions.

## 1.7 Research objectives of the Thesis

If we assume domain-general perceptual and cognitive mechanisms are recruited in the processing of musical and linguistic structures (e.g. Patel & Iverson, 2007), then it raises the question of the extent to which speech production includes temporal proximity and pitch similarity cues that align with syntactic boundaries, and further, whether listeners can recruit this information during processing to facilitate the processing of syntactic structure. If there is evidence for both points, then it suggests low-level, domain-general groupings mechanisms could be used to comprehend hierarchical structure non-hierarchically. In the present thesis, we aim to test these claims, using several methodologies.

Trotter, Monaghan, Beckers, and Christiansen (Chapter 2) conducted a meta-analysis of key artificial grammar learning experiments in humans and non-human animals. The artificial grammar learning paradigm is a frequently employed, versatile technique for investigating grammatical processing in humans and non-human animals. However, studies have employed a wide range of participants (human adults, infants, several species of primates and birds), methods (e.g. two alternate forced choice, eye-tracking paradigms, serial reaction time, head-turn preference procedures) and grammatical structures ($A_nB_n$, $AB_n$, AxC), making it difficult to determine which aspects of these studies influence learning. Thus, the primary question addressed by this study was whether differences in performance can be attributed to learners acquiring the artificial grammar, or to sensitivity to surface level features of the language, training regime, or testing methodology.

Chapter 2 reviewed the literature on cues that support learning of different structural dependencies. In Chapter 3, I therefore tested the extent to which prosodic cues were available in natural speech to support hierarchical structure. Trotter, Frost, and Monaghan (Chapter 3) therefore conducted a speech corpus analysis, assessing whether pitch and temporal cues provide syntactic boundary information consistent with Gestalt processing. The corpus comprised spontaneously produced active-object ("[The bear [the girl hugs] is white]") and passive relative clauses ("[The man [being hugged by the girl] is wearing green]") elicited using a picture description task. The analysis revealed that for active-object relative clauses, pitch similarity and temporal proximity cues were consistent with the syntactic structure; a large difference in pitch occurred between the noun phrases of the main and the embedded clause ("The bear [pause/pitch reduction] the girl hugs…”). Thus, temporal proximity and pitch similarity reinforce syntactic structure by supporting grouping of the phrases of the embedded clause, distinguishing them from the phrases of the main clause. However, this was not the case for passives, where a large pitch reduction and pause occurred between the embedded verb- and noun-phrase (The man being hugged [pause/pitch reduction] by the girl…”). Thus, for passive structures, pitch similarity and temporal proximity cues were inconsistent with syntactic structure.

The literature review of Chapter 2 demonstrated that multiple cues can support learning complex hierarchical structures in an artificial language paradigm. Chapter 3 showed that prosodic cues reflect phrase structure in spontaneous speech. However, the literature is unclear about the extent to which prosodic cues can interact with grammatical structure in acquisition, rather than processing, of that structure. Trotter, Frost, and Monaghan (Chapter 4) therefore employed the artificial grammar learning paradigm to assess whether additional pitch similarity, temporal proximity, and semantic similarity (marked with phonological cues) cues to dependencies would support the acquisition of a phrase structure grammar. Consistent with sequential processing accounts (Frank, Bod, & Christiansen, 2012), we hypothesised that these low-level grouping cues would facilitate learning by supporting dependency detection, relative to a baseline condition where only distributional cues were available to participants. The pitch cues in this study were modelled on a German corpus (Fery and Schubö, 2010), and the pause cues were taken from an artificial language learning task assessing clausal membership (Hawthorne & Gerken, 2014). Participants were assigned to one of five cue conditions; baseline, pitch (syntactically dependent syllables occurred in a similar pitch), pause (lengthened pauses were added between syntactically unrelated syllables), phonological similarity (dependent syllables always started with the same phoneme), or combined (pitch, pause and phonological similarity). Participants were first trained on grammatical structures, after which their learning was assessed with a grammaticality judgement task performed on novel grammatical and ungrammatical sequences.

Trotter, Monaghan and Frost (Chapter 5) conducted a follow-up study to assess whether tonal and temporal grouping cues consistent with an English-speaking corpus (Trotter, Frost, and Monaghan, Chapter 3) would improve learning of the same artificial phrase structure grammar. In comparison to Trotter, Frost, and Monaghan (Chapter 4), pitch variation was higher, increasing the salience of pitch cues, whilst the duration of pauses was reduced, resulting in lower salience.

Chapters 4 and 5 investigated how prosodic cues can assist the acquisition of hierarchical structure. However, to determine the mechanisms of language processing, it is helpful to consider how natural language is processed. Artificial grammar learning studies typically - and in Chapter 4 and 5 – assess participant learning using grammaticality judgement, or two-alternate forced choice tasks. Critically, both of these tasks rely on having participants make explicit decisions after stimulus delivery; they are offline, reflection-based tasks. Speech processing, however, is an online, processing-based task. Recently, the literature (e.g. Christiansen, 2018) has suggested that reflection-based measures are ill-suited for assessing processing-based tasks. This questions the extent to which the results of Chapters 4 and 5 can be explained by use of a reflection-based task, and whether a processing-based task would be more sensitive to any processing benefit of prosodic cues.

To address this issue, Trotter, Monaghan, and Frost (Chapter 6) assessed the role of temporal proximity and pitch similarity information using a processing-based task; the visual world paradigm. In this study, participants heard speech-synthesised active-object ("The boy the girl kicks walks") and passive relative clauses ("The boy kicked by the girl walks"), whilst viewing four potential targets; the target scene, a scene containing an agent-verb violation (e.g. the girl ignores the boy), a scene containing a patient-verb violation (e.g. the boy squats instead of running), and one in which their roles are reversed (the boy kicks the walking girl). Based on the results of Trotter, Frost, and Monaghan (Chapter 3), sentences could contain a number of prosodic boundaries; immediately preceding the first phrase of the embedded clause (active-congruent), immediately following the first phrase of the embedded clause (passive-congruent), in both locations (a high variance control), or no prosodic boundaries (a low variance control). A pitch boundary was defined as a 15Hz reduction in $F_0$, while a temporal boundary was comprised of a 111ms pause.

Taken together, these studies tested the extent to which participants are able to use the pitch similarity and temporal proximity Gestalts to group syntactically dependent phrases in hierarchical structures. Together, the studies of this thesis test whether low-level processing mechanisms can operate through which individuals could detect long-distance, complex dependencies in speech, obviating the need to process words hierarchically.

# Chapter 2: Exploring variation between artificial grammar learning experiments: Outlining a meta-analysis approach

Antony S. Trotter[1], Padraic Monaghan[1, 2, 3], Gabriël Beckers[4], & Morten H. Christiansen[5]

1. Lancaster University
2. Max Planck Institute for Psycholinguistics
3. University of Amsterdam
4. Utrecht University
5. Cornell University

Text as it appears in Trotter, Monaghan, Beckers, & Christiansen (2019). *Exploring variation between artificial grammar learning experiments: Outlining a meta-analysis approach.*. Manuscript accepted.

This study was conducted a literature review of what cues human and non-human animals draw on during the acquisition of syntax, and to determine what measures and task are best suited to measuring linguistic processing.

## Statement of Author Contribution

In the Chapter entitled, "What do humans and animals learn in artificial grammar learning experiments? A focused meta-analysis", the authors agree to the following contributions:

Antony S. Trotter — 65% (Writing, data collection, and analysis)

Signed: ______________________ Date: __________

Professor Padraic Monaghan — 25% (Experimental design, writing, and review)

Signed: ______________________ Date: 21/2/19__________

Professor Gabriel Beckers — 5% (Review)

Signed: ______________________ Date: 22/2 // 13

Professor Morten H. Christiansen — 5% (Review)

Signed: ______________________ Date: 21/2/19__________

# Abstract

Artificial grammar learning (AGL) has become an important tool used to understand aspects of human language learning and whether the abilities underlying learning may be unique to humans or found in other species. Successful learning is typically assumed when human or animal participants are able to distinguish stimuli generated by the grammar from those that are not at a level better than chance. However, the question remains as to what subjects actually learn in these experiments. Previous studies of AGL have frequently introduced multiple potential contributors to performance in the training and testing stimuli, but meta-analysis techniques now enable us to consider these multiple information sources for their contribution to learning – enabling intended and unintended structures to be assessed simultaneously. We present a blueprint for meta-analysis approaches to appraise the effect of learning in human and other animal studies for a series of artificial grammar learning experiments, focusing on studies that examine auditory and visual modalities. We identify a series of variables that differ across these studies, focusing on both structural and surface properties of the grammar, and characteristics of training and test regimes, and provide a first step in assessing the relative contribution of these design features of artificial grammars as well as species specific effects for learning.

## Introduction

Artificial grammar learning (AGL) studies present learners with sequences of stimuli that inhere particular structural properties (Miller, 1958) of differing complexity (e.g., Reber, 1967), and then test learners on their ability to respond to sequences that incorporate aspects of this structure. Such an approach has been a very powerful method enabling investigations within a species into the possibilities and constraints on structural learning, such as distinctions between phrase-structure grammars or finite state grammars (e.g., Bahlmann, Schubotz, & Friederici, 2008), or the extent to which adjacent or non-adjacent dependencies in sequences are available to the learner (e.g., Conway et al., 2010; Gomez & Gerken, 1999; Jamieson & Mewhort, 2005; Lai & Poletiek, 2011; Vuong, Meier & Christiansen, 2016). The paradigm is also of great potential use across species, and has been extensively used to address questions about what structures are learnable by which species, and under what conditions (e.g., Abe & Watanabe, 2011; Chen et al., 2015; Fitch & Hauser, 2004; Saffran et al., 2008).

There has already been substantial progress made in addressing these questions, resulting in an intensive array of studies of learning in birds (e.g., Abe & Watanabe, 2011; Chen & ten Cate, 2015; Gentner et al., 2006; Spierings et al., 2015, 2017), non-human primates (e.g., Endress et al., 2010; Heimbauer et al., 2018; Wilson, Smith, & Petkov, 2015), as well as human children and adults (e.g., Frost & Monaghan, 2017; Gomez & Gerken, 1999; Saffran et al., 2008), addressing acquisition of multiple grammatical structures across these species. The other papers in this special issue provides a host of further examples of the paradigm in use.

However, testing different structures and different species raises substantial methodological problems when it comes to direct comparisons between grammars and between species. Potential confounds both within and across studies have caused substantial concern in the past in terms of the validity of conclusions being drawn from studies (e.g., Beckers et al., 2012, 2017; de Vries et al., 2008; Perruchet & Pacteau, 1990; Perruchet et al., 2004), such as determining exactly what aspect of the structure is being responded to – whether that be the actual structures themselves, or some other feature of the stimuli (see, e.g., Knowlton & Squires, 1996). However, by using current meta-analysis techniques, the presence of these potential confounds can actually provide valuable opportunities for teasing apart some of the multiple factors that may contribute to learning. Thus, the pattern of such confounds across studies provides a backdrop against which the contribution of specific experimental design decisions can be assessed in terms of their effect on participant learning. Critically, meta-analysis permits researchers to quantify the effects of different kinds of stimuli within a species, but also differences across species in how they may respond to different grammatical structures. In the present study, we present an analysis of a subset of AGL studies, providing a framework that more comprehensive analyses can follow.

In cross-species comparisons, a key topic of interest is to determine which grammatical structures are potentially learnable by distinct species (Fitch & Friederici, 2018; Ghirlanda et al., 2017). The prospect of such discoveries has broad repercussions for the evolution of communicative systems, and the human specificity of language structure. The stakes are thus high. As one influential example, Fitch and Hauser (2004) conducted a study that required human adults and cotton-top tamarins to distinguish between strings generated by a phrase-structure and a finite-state grammar. Only the humans were able to make this distinction when trained on strings from the phrase-structure grammar. Subsequent research, however, has revealed several confounds in this study, suggesting that the humans may have relied on other sources of information to make their responses instead of the intended structural information (e.g. de Vries et al., 2008; Perrruchet & Rey, 2005).

An ideal, perfectly-controlled methodological study would isolate a particular grammatical structure and test learning of that particular structure without influence from other properties of the stimulus. However, the complexity of language structure and the practical challenges of training and testing different species on language-like structures introduces variation into the actual tasks being conducted. Ensuring that only one particular aspect of language structure is tested, and tested in the same way across studies involving different species, remains a substantial, potentially insoluble, challenge.

In a recent small-scale review of cross-species studies of artificial grammar learning, Beckers et al. (2017) identified several characteristics that could have biased learning toward accepting the grammatical structure being tested without necessarily indicating learning of the structure. These included the extent to which the test sequence had previously occurred in the same form during exposure to the training sequences (either wholly or in part), whether the test sequence shared the same onset as the training sequences, and whether the test and training sequences were cross-correlated even if they did not contain exactly the same sequences or subsequences. Thus, in a study containing one or more of these specific properties, it would be impossible to conclusively demonstrate that the grammatical rule was acquired by the learner. Such questions have been raised for almost as long as artificial grammar learning studies have been conducted – the extent to which learning is of particular grammatical structures or instead responding to lower-level fragments in the sequences (cf. Knowlton & Squire, 1996; Perruchet & Pacteau, 1990—see Frost, Armstrong, Siegelman & Christiansen, 2015, for a review).

Artificial grammars also differ on fundamental structural properties. Some AGL studies contain dependencies between adjacent stimuli, whereas others contain dependencies between non-adjacent elements in the stimuli. Furthermore, artificial grammars may differ in terms of the number of distinct stimulus elements that sequences contain, and the number of different categories to which these stimulus elements belong. An artificial grammar with a larger versus a smaller vocabulary, or a larger versus smaller set of grammatical categories, may affect learning distinctly. Learning studies can also vary in terms of the modality of the stimuli – whether they are auditory or visual (Heimbauer et al., 2018). For example, whilst cotton-top tamarins are often trained on auditory (e.g. human non-words, monkey calls; Neiworth et al., 2017) and visual materials (e.g. structured visuospatial sequences; Locurto, Fox, & Mazzella, 2015), zebra finches only receive auditory materials consisting of manipulations of species-specific birdsong (e.g. Chen and ten Cate, 2015; van Heijningen et al., 2009). Modality is known to have distinctive effects on learning sequence structure (for reviews, see Frost et al., 2015; Milne, Wilson & Christiansen, 2018), and for these reasons modality is taken as a focus of the literature that we will analyse.

Artificial grammar learning studies also differ in terms of how training and testing is conducted. Studies of complex sequences with non-human primates and birds may require substantial training time – several thousand trials over several weeks – whereas studies with human adults are typically constrained to short training sessions with a constrained set of training trials. Testing also varies in terms of how the effects of learning are measured. For instance, in testing human adults and children there is frequently a distinction between explicit, reflection-based tasks for adult responses, such as alternative forced choice, or go/no-go responses, and implicit, processing-based tasks such as head-turn preferences or looking times. These tasks may tap into different mechanisms, with processing-based tasks more effective for assessing processing-based learning, such as acquisition of grammatical structures (Christiansen, in press; Frizelle, O’Neill, & Bishop, 2017; Isbilen et al., 2018).

As we have summarised, studies of artificial grammar learning may vary along several of these dimensions simultaneously. In this paper, we present a blueprint for how a meta-analysis approach could proceed to quantify how various design features of AGL studies might influence performance. We analyse a subset of AGL studies that have focused on presenting stimuli in either auditory or visual modalities, as reflected in the key words used within these articles. As we focus only on a subset of AGL studies, the conclusions drawn within the analysis may not generalise to the wider literature. The primary aim of our study is thus to provide a meta-analytic framework that a more comprehensive study may adopt. We show how meta-analytical methods enable us to measure the relative contributions of multiple potential confounds – reconsidered here as moderators – in influencing the size of the observed effects. This means that what was once considered a confound can actually be reinterpreted as providing a valuable and interesting source of data towards determining the limits and constraints on learning within and across species.

## Method

### *Literature Search*

We conducted the literature search and meta-analysis in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009), pre-registering the encoding and analysis to be conducted (https://aspredicted.org/wf2uk.pdf). The literature search was conducted on the SCOPUS database (Scopus, 2019) on articles published up to March 2019. In order to focus our literature review, we searched for studies that considered explicitly the modality of presentation in artificial grammar learning. We therefore conducted two searches of keywords appearing in titles, keywords, and abstracts of articles. In the first, we searched the keywords “artificial grammar learning” and “vision” OR “visual”. In the second, we used the keywords “artificial grammar learning” and “auditory” or “audio” or “audiovisual”. The results were then merged into a master list, and submitted to study selection criteria.

The search we performed avoided bias in selecting publications for analysis, in accordance with PRISMA guidelines, but it is important to note that the results of the search were not comprehensive in including all papers that conducted AGL studies with auditory or visual stimuli. The literature search for instance failed to include several influential artificial grammar learning studies (e.g., Gentner et al., 2006; Hauser & Fitch, 2004; Reber, 1967; Saffran et al., 2001, 2008). Our approach therefore outlines a blueprint for conducting meta-analyses of potential design differences in AGL research, rather than to provide a final, comprehensive answer as to the size of effects of learning in AGL studies.

### *Study selection*

The literature search resulted in 91 records. Of these, 11 were duplicates. Of the 80 articles remaining, 8 were review articles, 3 presented computational modelling and no behavioural data, 1 study reported neuroimaging data of primates with no behavioural data, and 2 reported a case study on an aphasic population with no control group. These articles were removed, and the remaining 66 articles contained 78 studies involving 3559 subjects (this includes subjects tested more than once in the same article – see Results section for how the analysis took into account multiple studies within articles). Figure 1 shows the PRISMA literature search flowchart. The list of studies included are reported in the Supplementary Materials.

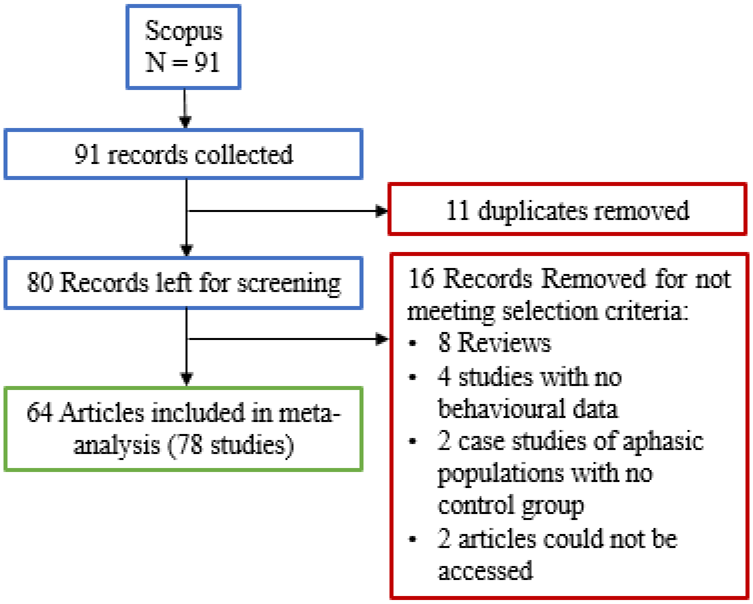

Scopus
N = 91
91 records collected
11 duplicates removed
80 Records left for screening
16 Records Removed for not meeting selection criteria:
• 8 Reviews
• 4 studies with no behavioural data
• 2 case studies of aphasic populations with no control group
• 2 articles could not be accessed
64 Articles included in meta-analysis (78 studies)

Figure 1. Flowchart of the PRISMA literature search criteria used in the current meta-analysis.

*Data extraction and effect size calculation*

The effect size for each study was initially computed as Cohen's d, and subsequently corrected to Hedge's g, with the variance of g computed in accordance with Borenstein et al. (2009). Formula (1) provides correction factor *J*, which is multiplied with Cohen's d to provide Hedge's g (2). The variance of Hedge's g, $V_g$, was provided by (3), where the variance of Cohen's *d* is computed, and corrected by *J*.

$$(1)\ J = \left(1 - \frac{3}{4df - 1}\right)$$

$$(2)\ g = J \times d$$

$$(3)\ V_g = \left(\frac{1}{n} + \frac{d^2}{2 \times n}\right) \times J^2$$

Cohen’s d was derived for each type of dependent variable, the dependent variable for each study is shown in the Supplementary Materials. For studies reporting the number correct, numbers endorsed or responded to, or go/no-go responses as dependent variable, the effect size was computed from the difference to chance responding in a one sample test (see Equation 4):

$$(4)\ d = \frac{Mean - Chance}{SD_{Within}}$$

In cases where tests and language structures were similar over different test sessions or conditions (e.g. Cope et al., 2017; Goranskaya et al., 2016; Mueller et al., 2010), we combined the means and SDs from each of the multiple test sessions, and computed the one sample difference from chance. The pooled mean was simply computed as the arithmetic mean across the sessions, weighted by number of participants in the session. For pooled SD, we took the average SD using equation (5),  where $n_1$ is the number of items in test session 1, $n_2$ is the number of items in test session 2, etc., and $SD_1$ is the observed standard deviation of the test session 1 response accuracy, etc. (see van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017):

$$(5)\ SD_{Average} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + (n_3 - 1)SD_3^2 + (n_4 - 1)SD_4^2}{n_1 + n_2 + n_3 + n_4 - 4}}$$

Subsequently, we computed *d* using equation (4), with the pooled mean, 50% as chance, divided by the SD $_{\text{Average}}$. In serial reaction time studies, the effect was measured as the standardised mean difference in RT between presentations of a trained vs. an untrained structure, with SD$_{\text{Average}}$ computed as in (5), which assumes conservatively that there is a correlation of 1 between the trained and untrained structure responses across participants (a lower correlation would result in a lower SD, so this formula provides a conservative upper limit for the effect size). For instance, for Kemeny and Nemeth's (2017) data represented in Figure 3, presenting the mean response time (RT) and SEM per testing block. In this case, we pooled the mean RT for the grammatical blocks 4 and 6 weighted by the number of participants in the session, and computed *d* as the difference to the mean RT for the ungrammatical block 5, with SD computed as the SD $_{\text{Average}}$ across blocks 4, 5, and 6, using (5).

For sequence reproduction tasks, the effect size was computed as difference in mean accuracy for grammatical sequences and ungrammatical sequences, with *SD* as the SD $_{\text{Average}}$ computed using (5).

In head-turn preference paradigms (e.g. Gomez & Gerken, 1999), effect size was the proportion of trials where the participant turned towards the grammatical violation sequences over the grammatical sequences, indicating observation of the violation. These values were compared to chance and *d* computed in the same way as for response accuracy measures.

For looking time paradigms (e.g. Milne et al., 2018), the effect size was computed as the difference in fixation duration between grammatical and ungrammatical sequences, computed using the same approach as that for sequence reproduction paradigms. Positive effects were generally computed as longer looking to ungrammatical than grammatical sequences (a novelty effect). However, in cases where the interpretation of the authors suggested that longer looking times to grammatical stimuli (or preferences in head-turn to grammatical sequences) reflected greater learning (i.e., a familiarity effect), we re-signed these effects.

In studies where means and variance were reported only in figures, we contacted authors for data, and utilized the Digitizeit digitizer software (available from: http://www.digitizeit.de/) when such data was not available, to extract the means and SDs. In cases where graphs displayed the mean and 95% confidence intervals (Hall et al., 2018), confidence intervals were converted into SDs according to (6), which assumes that the authors had computed the confidence intervals using the t-distribution (which is more conservative than assuming confidence intervals based on the Z-distribution), where tcrit is the critical value of the *t*-distribution for n-1 degrees of freedom at *p* = .05:

$$(6)\ SD = \sqrt{n} \times \frac{upperlimit - lowerlimit}{2 \times tcrit[n-1]}$$

Each study was encoded for several features in order to test their influence on learning performance. We encoded the animal class and species that was tested, and in the case of human studies, distinguished whether the study was on children (<18 years) or adults.

For properties of the AGL structure, we encoded whether the study contained at least some repetitions of the stimuli experienced during training in the testing, whether the artificial grammar contained adjacent dependencies or did not contain adjacent dependencies, and whether the artificial grammar contained non-adjacent dependencies or did not contain non-adjacent dependencies.

For characteristics of training and testing, we encoded the type of test response that was being collected – whether this was a Yes versus No judgment, a go or no-go task, a scale judgment, a forced choice test between two or more alternatives, serial reaction time, head-turn preference, looking time, sequence production, or frequency estimation task. We subsequently grouped these variables into whether they required reflection on the grammatical structure (reflection-based; forced choice tests, yes versus no judgement, go/no-go, scale judgement), or more directly tapped into the underlying processing of the grammatical structure (processing-based; looking time, head-turn preference, serial reaction time, sequence production) (Christiansen, in press). We encoded the amount of exposure to the artificial grammar that participants experienced in terms of the total number of stimulus tokens from the grammar during exposure (training length).

Importantly, we also encoded a number of surface features of the AGL, including whether the stimuli were visual, auditory, or a combination of both visual and auditory, in order to determine whether learning varied according to the modality of the task. Further, we also encoded the size of the artificial grammar in terms of the size of the vocabulary in the grammar (or the number of distinct items), as well as the number of different categories in the grammar (e.g., for a phrase-structure grammar with four nouns, two verbs, two adjectives, and two determiners, the number of categories is 4 (noun/verb/adjective/determiner) and the size of the vocabulary is 14.

## Results

*Evidence of acquisition of structure from AGL studies*

The overall effect size across the studies, and the extent to which each of the encoded study variables predicted differences in effect sizes across the studies, was determined by conducting a random effects meta-analysis of effect sizes, using the R package metafor (Viechtbauer, 2010). This approach takes into account inconsistencies between the studies analysed, provides an estimate of sampling error, and also permits a measurement of the effects of each of the variables in moderating the size of the overall behavioural effect (Borenstein, Hedges, Higgins, & Rothstein, 2010; Borenstein, Higgins, & Rothstein, 2009). We encoded each experiment in an article and each test in an experiment as a separate study, and as these cannot be assumed to result in effect sizes independent from one another, we encoded article as a nested multilevel variable in the analysis (Konstantopoulos, 2011).

The model was run using the rma.mv function with the restricted maximum likelihood (REML) method. We utilised the *t* method to generate test statistics and confidence intervals. The model was run using the rma.mv function with restricted likelihood (REML) method, and the t-adjustment to calculate the model estimates of standard errors, p values and confidence intervals. Effect sizes for individual studies and the overall average weighted effect sizes are presented in Figure 2. A positive effect size indicates greater preference for stimuli conforming to the AGL structure, while a negative effect size indicates preference for non-conforming stimuli (except in the case of the looking studies, where a positive effect indicates longer looking to violating stimuli – as this was the predicted effect of such studies in reflecting AGL acquisition, e.g., Gomez & Gerken, 1999).

The meta-analysis resulted in the average weighted effect size = 1.069 (SE = .130, 95% CI [.813, 1.326], p < .0001), indicating that overall there was strong evidence of learning in AGL studies.

### *3.2 Publication bias*

To determine whether there was publication bias in the sample, we conducted a Peters' test (Peters et al., 2006) on the random multilevel meta-regression model. The Peters' test revealed a significant asymmetrical distribution, *t*(154) = -2.290, p = .023, indicating the presence of publication bias in our sample. The funnel plot (Figure 2) displays the standard error (a measure of study precision) against the effect sizes of the individual studies. In the absence of publication bias, studies should be symmetrically distributed around the average weighted effect size in a funnel shape, with high precision studies being closer to the average weighted effect size, and lower precision studies symmetrically distributed around the average weighted effect size. The distribution indicates that there are more large positive effect sizes for smaller sample sizes than would be expected from a standard distribution of studies, suggesting a potential publication bias. The size of the effect of AGL acquisition, and the sources of heterogeneity of the effects, should thus be considered in light of possible bias in the studies published.

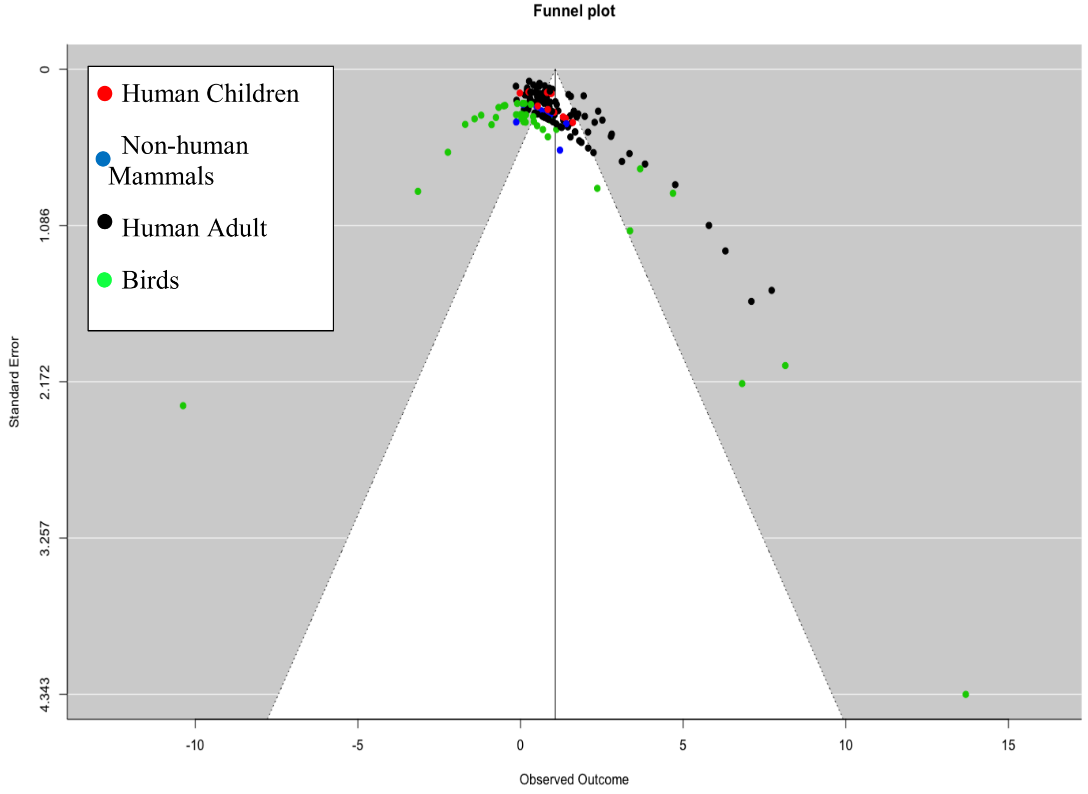


Figure 2. Funnel plot showing the relationship between the standard error and the effect size of the individual studies. Points are colour-coded according to animal class. Black points illustrate Human Adult Studies, blue illustrate Non-human mammals studies, red are Human Child studies, and green are Bird studies.

### *3.3 Heterogeneity in effect size variance associated with study variables*

Cohran's Q-test for heterogeneity was significant ($Q(155) = 1185.657$, $p < .0001$), indicating that variance in the data cannot be explained by random measurement error, but that different aspects of studies are contributing to the effect size. We thus analysed the effects of each of the set of variables we encoded from each of the studies as moderators, shown in Table 1.

For the effect of animal class (but also distinguishing human adults and human children from non-human mammals), there were significant differences on the size of effect of learning between different species. For human adults, the overall effect size was 1.252 (SE = .148, 95% CI [0.958, 1.545], *p* < .0001). For human children, the overall effect size was 0.615 (SE = .231, 95% CI [.101, 1.129], *p* = .0237). For non-human mammals, the overall effect size was 0.626 (SE = .172, 95% CI [.221, .1.032], *p* = .008). For birds, the overall effect size was 0.428 (SE = 0.533, 95% CI [-0.653, 1.509], *p* = .427).

Properties of training and testing of AGL studies were found to produce significant differences in effect sizes. Log-transformed number of training trials related negatively to effect size, -0.188 (SE = 0.054, 95% CI [-0.295, -0.0815], *p* = .0006). Further, repetition of trained items at test resulted in larger effects 1.051 (SE = 0.279, 95% CI [0.499, 1.602], *p* = .0002).

Surface level features of the language did not significantly moderate the variance of effect sizes (see Table 1), and this included also the modality of stimulus delivery. The number of categories, the vocabulary size, and critically, whether the stimuli were visual or auditory were not found to affect the overall effect size.

For the structural properties of the language, there were moderating effects. The presence of repetition of items from training to test positively influenced effect sizes, with an overall effect of 1.051 (SE = 0.279, 95% CI [0.499, 1.602], *p* = .0002).

As there were different sized effects of learning for each animal class, and possible confounds between study design characteristics and animal class tested, we conducted further analyses of moderator variables for human adult, human child, birds, and non-human mammals separately.

Table 1. Contributions of each moderating variable to account for variance in effect sizes across studies.

| Moderator | | *F* | *Df1, Df2* | *p* |
|---|---|---|---|---|
| Population | | | | |
| | Animal Species | 2.613 | (10, 145) | < .0001*** |
| | Animal Class | 5.811 | (3, 152) | .0009*** |
| | Human vs. Non-human | 7.555 | (2, 153) | .0007*** |
| Training and testing | | | | |
| | Log Training Length | 12.149 | (1, 154) | < .0001*** |
| | Stimulus Modality | 0.095 | (2, 153) | .909 |
| | Test Response | 1.624 | (10, 145) | .105 |
| | Test Type | 3.698 | (1, 154) | .056 |
| Surface level properties | | | | |
| | Categories in Language | 0.0001 | (1, 154) | .992 |
| | Number of unique vocabulary items | 3.021 | (1, 154) | .084 |
| Structural Properties | | | | |
| | Repetition of items | 14.162 | (1, 154) | .0002** |
| | Adjacent dependencies | 0.238 | (1, 154) | .627 |
| | Non-adjacent dependencies | 0.118 | (1, 154) | .608 |

Note. *F* is the statistic for testing whether the moderator accounts for some heterogeneity between studies; *p* is the significance for the *F*-test *** p < .001, ** p < .01, **p* < .05. Note that Animal Class distinguishes birds, non-human mammals, human adult, and human child. Animal species also distinguishes human adult and human child.

### *3.4 Moderator Analysis of Human Adults*

There was significant heterogeneity of variance in the effect size in studies testing human adults (*Q*(99) = 707.273, *p* < .001), so we analysed the effect of each moderator (see Table 2 for the significance of each moderator). There was a significant effect of the presence of non-adjacent dependencies (effect = 0.582, SE = 0.259, 95% CI [0.068, 1.096], *p* = .027), suggesting that adult human participants are overall successful in learning non-adjacencies in artificial grammars.

Table 2. Contributions of each moderating variable to account for variance in effect sizes in Human Adult studies.

| Moderator | | *F* | *Df1*, *Df2* | *p* |
|---|---|---|---|---|
| Training and testing | | | | |
| | Log Training Length | 0.415 | (1, 98) | .521 |
| | Stimulus Modality | 0.306 | (2, 97) | .737 |
| | Test Response | 0.671 | (8, 91) | .716 |
| | Test Type | 1.884 | (1, 98) | .173 |
| Surface level properties | | | | |
| | Categories in Language | 0.319 | (1, 98) | .574 |
| | Number of unique vocabulary items | 1.023 | (1, 98) | .305 |
| Structural properties | | | | |
| | Repetition of items | 0.036 | (1, 98) | .851 |
| | Adjacent dependencies | 1.745 | (1, 98) | .190 |
| | Non-adjacent dependencies | 5.050 | (1, 98) | .027* |

### *3.5 Moderator Analysis of Human Children*

There was significant heterogeneity (*Q*(10) = 49.953, *p* < .0001), so we further analysed the effect of each moderator (see Table 3). In this analysis, the only significant moderator was the test response participants made. This analysis indicated that head-turn preference paradigms produced an overall effect of 1.301 (SE = 0.1663, 95% CI [0.772, 1.831], *p* = .004). Sequence production paradigms, by comparison, produced an effect that failed to statistically differ from 0 (effect size = 0.150, SE = 0.144, 95% CI [-0.433, 0.721], *p* = .395). Finally, binary yes-no judgement tasks produced an overall effect of 0.822 (SE = 0.099, 95% CI [0.506, 1.137], p = .004).

Table 3. Contributions of each moderating variable to account for variance in effect sizes in human child studies.

| Moderator | | *F* | *Df1, Df2* | *p* |
|---|---|---|---|---|
| Training and Testing | | | | |
| | Log Training Length | 0.214 | (1, 9) | .654 |
| | Stimulus Modality | 3.427 | (1, 9) | .097 |
| | Test Response | 15.978 | (2, 8) | .002* |
| | Test Type | 0.271 | (1, 9) | .615 |
| Surface level properties | | | | |
| | Categories in Language | 0.059 | (1, 9) | .813 |
| | Number of unique vocabulary items | 0.862 | (1, 9) | .377 |
| Structural properties | | | | |
| | Repetition of items | 2.503 | (1, 9) | .148 |
| | Adjacent dependencies | 0.023 | (1, 9) | .884 |
| | Non-adjacent dependencies | 0.012 | (1, 9) | .917 |

### *Moderator Analysis of Non-human Mammals*

There was significant heterogeneity ($Q(7) = 15.928$, $p < .026$), therefore we analysed the effect of each moderator (see Table 4). Non-human mammals only took part in studies delivered in the auditory modality, and all of which were processing based, included adjacent dependencies, and did not include repetitions at test, and hence we did not include a moderator analysis of testing modality, repetition of items, adjacency, and testing type. No moderator accounted for a significant proportion of variance in this dataset.

Table 4. Contributions of each moderating variable to account for variance in effect sizes in non-human mammal studies.

| Moderator | | *F* | *Df1, Df2* | *p* |
|---|---|---|---|---|
| Training and testing | | | | |
| | Log Training Length | 1.121 | (1, 6) | .331 |
| | Test Response | 1.262 | (1, 6) | .304 |
| Surface level properties | | | | |
| | Categories in Language | 0.760 | (1, 6) | .418 |
| | Number of unique vocabulary items | 0.365 | (1, 6) | .567 |
| Structural properties | | | | |
| | Non-adjacent dependencies | 0.111 | (1, 6) | .750 |

*Moderator Analysis of Birds Studies*

There was again significant heterogeneity ($Q(36) = 259.498$, $p < .0001$), therefore we analysed the effect of each moderator (see Table 5). Birds, however only took part in classification-based tasks, and thus, we did not analyse the effect of test type. Log training length accounted for a significant portion of the variance, increased training resulted in a lower effect size -0.739 (SE =

.268, 95% CI [-1.283, -0.195], *p* = .009). Increased vocabulary sizes tended to increase effect sizes (effect size = 0.099, SE = 0.038, 95% CI [0.022, 0.177], *p* = .014). Stimulus modality explained a significant portion of variance, with visual stimuli producing larger effects (effect size = 1.993, SE = 0.788, 95% CI [0.395, 3.592], *p* = .016) than auditory stimuli. The response task used also accounted for a significant portion of variance of effect sizes, however, the meta-analytic estimate for both 2AFC tasks (effect size = 2.288, SE = .135, 95% CI [-0.488, 5.065], p = .090) and go/no-go tasks (effect size = -0.042, SE = 0.294, 95% CI [-0.642, 0.559], p = .889) failed to significantly differ from 0. This reflects the fact that variance of effect sizes in birds was large; to properly account for the moderating effect of task type on the variance in effect size for bird studies, a larger set of studies for inclusion would be helpful. Finally, the repetition of items accounted for a significant portion of the variance of effect sizes, whereby repeating items at test resulting in an effect size of 5.013 (SE = 0.740, 95% CI [3.511, 6.515], *p* < .0001). This effect is explained by the only study including repetitions of whole strings at test (Spierings & ten Cate, 2016) produced large effect sizes.

Table 5. Contributions of each moderating variable to account for variance in effect sizes in birds studies.

| Moderator | | *F* | *Df1, Df2* | *p* |
|---|---|---|---|---|
| Training and testing | | | | |
| | Log Training Length | 7.609 | (1, 35) | .009** |
| | Stimulus Modality | 6.407 | (1, 35) | .016* |
| | Test Response | 6.407 | (1, 35) | .016* |
| Surface level properties | | | | |

|  |  |  |  |  |
|---|---|---|---|---|
| Structural properties | Categories in Language | 0.053 | (1, 35) | .819 |
|  | Number of unique vocabulary items | 6.712 | (1, 35) | .014* |
|  | Repetition of items | 45.926 | (1, 35) | < .0001*** |
|  | Adjacent dependencies | 2.462 | (1, 35) | .126 |
|  | Non-adjacent dependencies | 1.661 | (1, 35) | .206 |

## Discussion

We presented a focused literature search analysing AGL studies that address the modality of stimulus presentation, taking into account the varieties of designs, as well as species, that are tested across these studies. This approach provides a blueprint for how meta-analysis in AGL studies can assess the influence of multiple moderators on learning, providing insight into the conditions under which learning of regularities in artificial grammars can be observed. Confounds and differences between studies – both intended and unintended (and previously viewed as adding opacity to the field of research) – can be considered sources of information for disentangling multiple contributors to learning of artificial grammar stimuli, rather than serve only as an impediment to comparison between studies. Heterogeneity of design can actually be analysed through an estimate of heterogeneity of variance which can then be associated with the presence or absence of differences across studies.

The current analysis was conducted to provide a framework for how future, more comprehensive meta-analyses might robustly identify patterns in the artificial grammar learning literature. However, our literature search was constrained by a restricted set of keywords that selected only papers where AGL and modality of presentation were explicitly tagged as features of the study. We know that influential studies in the literature were omitted by our approach. Whereas our focus here was to avoid bias in selecting the papers for inclusion in our analysis by conducting an objective keyword search, this absence of key studies highlights that there are relevant papers that are not included in the current analysis, and so the comprehensiveness of our search cannot be assumed. Consequently, the precise results of the meta-analysis and the moderator analysis should not be taken as the final word on this topic. Instead, we have shown how a future analysis, on an even more comprehensive set of studies, may help move the field forward. Such a study will be a considerable undertaking; a Scopus search with the keywords "artificial grammar learning" or "statistical learning", for instance, resulted in 6,511 records and still failed to include the landmark studies by Fitch and Hauser (2004), Gentner et al. (2006), and Reber (1967), mentioned in the Introduction, though the search did succeed in including the key studies by Saffran (2001) and Saffran et al. (2008). Finding principled ways to limit the literature search, without omitting key articles, presents an additional interesting challenge in this field of research.

This shortcoming raises concerns about terminological specificity in the field of artificial grammar learning. If we take Fitch and Hauser's (2004) study, this paper explicitly implements an AGL method, however, it instead describes it as a "familiarization/discrimination paradigm" in its abstract. Gentner and colleagues (2006) do not describe their method in the abstract, and in text specify it as a go/no-go operant conditioning procedure of $AB^n$ and $A^nB^n$ grammars. Similarly, Saffran's (2001) and Saffran et al.'s (2008) methods are variously described as statistical learning, grammatical pattern learning, or familiarization-discrimination.

Cumming (2014) provided a compelling argument for favouring magnitude estimation over null hypothesis significance testing in assessing experimental effects. A tenet of this approach is to employ meta-analytic thinking throughout the research process, including writing, reporting, and publication. The diversity of terms utilised to describe related methods makes it difficult to devise a singular, constrained set of search terms that would gather them together in a given search. Moving forward, we would suggest that using informative, umbrella keywords will ameliorate this issue, facilitating meta-analyses, and in Cumming's (2014) view, support research integrity.

In terms of the results of our focused meta-analysis in terms of what can be learned across animal classes, the analyses showed that the size of learning effects varies according to the species tested, though the evidence of publication bias and the potential lack of comprehensiveness in the search mean that interpretations based on size of effects must be treated with caution. The overall largest effect was observed for studies involving adult humans, but there were also overall significant effects of learning associated with child humans, non-human mammals, though not for birds. However, there are many differences between studies designed to appraise learning in different species, and heterogeneity of the variance within studies addressing each species points to ways in which these design differences may have profound effects on learning. The analyses of moderator effects within each animal class demonstrated that multiple variables were affecting learning, highlighting potential distinctions across species.

The size of the observed effects for human children was affected by the test response required, with similar effect sizes for head-turn preference and Yes/No judgement tasks. Whilst sequence production tasks did not significantly differ from 0, this likely reflects the small number of child studies included in the present analysis. For birds, the presence of training items at test produced large effects, perhaps unsurprising given the large amount of training they receive. Intriguingly, a greater number of training trials related negatively to effect size. This is likely correlated with the specific species of bird tested, and thus represents an important variable to focus on in a comprehensive meta-analysis. For adult humans, larger effects were produced by grammars containing non-adjacent dependencies than sequences without those dependencies, which have traditionally been difficult to observe in individual studies (e.g., Frost & Monaghan, 2016; Lai & Poletiek, 2011; Perruchet et al., 2004), see Wilson et al. (in press) in this issue for further discussion. The absence of a significant effect of adjacent dependencies was unexpected, but highlights the variation that can occur in the effect sizes across studies testing these structures.

Further meta-analytical techniques can help determine the additional sources of information that might support such learning, such as use of reflection- versus processing-based test measures (Vuong et al., 2016). In order to measure the effect of learning on processing, rather than explicit decision-making based on the structures experienced by the learner, a task that probes processing is proposed to be more effective (Christiansen, in press; Frizelle et al., 2017; Isbilen et al., 2018), however, in the present analysis there was no statistically reliable difference between the two. This may be a consequence of the comparatively large number of reflection-based effects (135) relative to processing-based effects (21) included in this analysis, or of the range of grammars that tend to be tested in AGL studies, a large number of studies use Reber-style (1967) grammars, where explicit testing may produce a similar magnitude of effects. Moreover, the effect of reflection-based measures may also have been inflated by including the non-human animal data as they are unlikely to engage in the kind of conscious reflections often observed in human studies. Finally, the presence of a potential publication bias combined with the much longer use of reflection-based assessments in AGL studies going more than half a century may further explain this pattern.

A key issue that emerged during our analysis was that individual stimuli within a test may contain alternative structures or vary in the presence of surface features. The analyses in this paper report effect sizes and features of the stimuli across sets of stimuli, which can obscure the individual influence of these features. Making raw data sets publicly available would enable this by-items analysis to reveal the precise contribution of multiple variables to learning behaviour (e.g., Beckers et al., 2017).

The studies included here were selected from an objective literature search on SCOPUS, intending to avoid bias in our selection of tests, focusing on studies of AGL that describe the modality of the stimuli. Interestingly, except in the case of birds, modality was not found to affect the results, but this may also have been affected by observed publication bias. Expanding further to a literature search of an even broader literature would help to determine more clearly which moderators are affecting performance, and which are orthogonal to artificial grammatical learning. There are, for instance, other structures that are of key interest to both language acquisition research, and cross-species investigations of the limits of grammar learning – such as distinctions between phrase structure and finite-state grammars (Fitch & Friederici, 2012; Fitch & Hauser, 2004), or focused on hierarchical centre-embedded structures (Lai & Poletiek, 2011). Debates on the learnability of these structures (e.g., de Vries et al., 2008) will be facilitated by a wider survey of the published literature. In our blueprint for a meta-analysis approach in this field, we have made an illustrative first step toward providing a perspective on what is learned and what is learnable within and across species.

# Acknowledgements

AT, PM, and MHC were supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (UK) [ES/L008955/1]. GJLB is part of the Consortium on Individual Development (CID), which is funded through the Gravitation Program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO; grant number 024.001.003).

# Chapter 3: Chained Melody: Low-level acoustic cues as a guide to phrase structure in comprehension

Antony Scott Trotter[1], Rebecca L.A. Frost[2], & Padraic Monaghan[1,2,3]

1. Lancaster University
2. Max Planck Institute for Psycholinguistics
3. University of Amsterdam

Chapter 3 was carried out to assess whether particular acoustic features of human speech may facilitate the acquisition of hierarchical structure. We focused on acoustic cues, as above, the meta-analysis illustrated little difference of effect sizes between studies using auditory and visual presentation. Further, speech is the most frequent form of language use, and critically, the modality from which language is acquired. This paper is currently in a draft ready for submission.

## Statement of Author Contribution

In the Chapter entitled, “Chained Melody: Low-level acoustic cues as a guide to phrase structure in comprehension”, the authors agree to the following contributions:

Antony S. Trotter – 70% (Writing, experimental design, and data analysis)

Signed: ______________________________ Date: ____________

Dr, Rebecca L. A. Frost – 20% (Review, and experimental design)

Signed: ______________________________ Date: 22.2.2019

Professor Padraic Monaghan – 10% (Review, and experimental design)

Signed: ______________________________ Date: 22.2.2019

## Abstract

To accurately process and respond to speech requires rapidly determining the structural dependencies between words in order to comprehend meaning. While phrase structure may be necessary for producing syntactically complex sentences, it has been argued that sequential processing along may be sufficient for comprehension (Frank, Bod, & Christiansen, 2012), with low-level statistical correspondences supporting dependency detection. In the present study, we investigated the extent to which prosody may support low-level processing of long-distance dependencies in complex syntactic structures. We hypothesised that syntactically dependent phrases would be similar in pitch, enabling grouping according to the Gestalt principle of similarity. Further, we hypothesised that pause duration could reflect the Gestalt principle of proximity; pauses occurring between clauses will render them distinct if they are longer than elsewhere in speech.

To explore this possibility, we analysed a corpus of speech data from Montag and MacDonald (2014), in which American English speakers (n = 64) spontaneously produced active ("[The bear] [the girl] [hugs] [is green]") or passive relative clauses ("[the bear] [being held] [by the girl] [is green]"). The results for actives supported our hypotheses; the embedded clause was preceded by a long pause, and phrases within it were similar in pitch. Passives differed, with a large reduction in pitch and a long pause following the verb phrase of the embedded clause. The results for actives suggest that Gestalt principles could be used to group the phrases of the embedded clause, obviating the need to process hierarchically structured speech hierarchically.

# 1. Introduction

Learning to process the hierarchical structure of language is of critical importance for language comprehension; in just a single sentence, listeners encounter multiple phrases, each comprising multiple words, which are in turn composed of multiple morphemes. Comprehending linguistic input requires understanding how each of these levels of language structure inter-relate. Classical descriptions of the comprehender's model of language refer to a system of generative rules at each level of this hierarchy, that permit the production and comprehension of an infinite number of phrases and sentences from a finite set of morphemes and words (Langus et al., 2012).

But how do learners arrive at the ability to make sense of complex language structure? There are two principal views of this process. The first states that language experience triggers innately-specified linguistic structure (Chomsky, 2005; Pinker, 1991) because the environment itself is insufficient to constrain the generation of linguistic structure. The alternative is that linguistic structure is learned through experience (e.g. Saffran & Aslin, 1996). Recently, it has been suggested that the application of domain-general learning mechanisms may drive this process (e.g., Christiansen & Chater, 2008). Under the latter view, multiple information sources can be brought to bear on constraining language structure – from beyond those constraints that apply between words or the grammatical categories to which they belong.

For instance, Farmer, Christiansen, and Monaghan (2006) demonstrated that phonological cues have an early influence on comprehenders' interpretation of sentences: when a word contains phonological properties consistent with nouns, it promotes sentence parsing when the word occurs in a noun position. However, it would impede sentence processing when the word occupied the position of a verb. Thus, the autonomy of syntax appears to be violable (Newmeyer, 2017);

statistically typical phonological cues influence syntactic processing. If this is the case, this raises the question as to how other sources of information in the environment – such as prosodic cues – influence syntactic processing.

One source of information that has been shown to relate to syntactic structure is prosody – the rhythmic and melodic features of speech. Critically, clauses are often cued with phrase-initial pitch-resetting, phrase-final declining pitch contour (Pierrehumbert, 1979), in addition to increased duration on phrase final words (Langus, Marchetto, Hoffman, Bion, & Nespor, 2012), as well as syllable-finally within words (Frost, Monaghan, & Tatsumi, 2017). The prosodic bootstrapping hypothesis (Gleitman & Wanner, 1982; Morgan, 1986; Peters, 1983) states that prosody may assist infants as they learn to process linguistic input. In support of this proposal, sensitivity to prosodic information has been documented in the earliest stages of infancy; research has demonstrated that new born infants can discriminate amongst languages on the basis of rhythm (Nazzi, Bertocini, & Mehler, 1998; see e.g., Toro, Trobalon, & Sebastian-Galles, 2003; Ramus, Hauser, Miller, Morris, & Mehler, 2000 for evidence of this in other mammalian species such as rats and macaques), and can detect changes in pitch at 1-2 months old (Kuhl & Miller, 1982). At 4.5 months, infants have been found to prefer to listen to passages with pauses inserted at clausal boundaries rather than other places in the sentence (Jusczyk, Hohne, & Mandel, 1995). Thus, it follows that learners may draw on the prosodic information contained in speech to help them during language acquisition.

Specifically, the prosodic bootstrapping hypothesis proposes that infants can draw on the prosodic information contained in speech to help identify word-, phrase-, and clause-boundaries, and to help infer constituency and hierarchical syntactic structure.  Nazzi, Kemler Nelson, Jusczyk and Jusczyk (2000) tested 6-month-old infants' ability to utilize prosodic cues present at clausal boundaries, using a Head-turn Preference Procedure. The familiarization stimuli were sentences extracted from passages that were read aloud, differing only by whether they were an entire utterance (e.g. *"Leafy vegetables taste so good."*), or made up of parts of two distinct utterances (e.g. *"…leafy vegetables. Taste so good…*"). In other words, the ill-formed utterances contained an erroneous prosodic boundary. In experiment one, at test, infants heard the entire passages containing the familiarization sentences. One contained the well-formed sentence, and the other contained its ill-formed counterpart. Infants looked significantly longer to passages containing the well-formed sentence compared to the ill-formed sentence. The infants also listened significantly longer to test stimuli that contained novel well- and ill-formed sequences taken from new passages. This demonstrates that 6-month-olds infants can recognize prosodic cues consistent with syntactic boundaries, even when the speech occurred within a longer passage. Extending these findings, in experiments 2 and 3, the familiarization sentences were extracted from new spoken passages, resulting in new intonational contours; they differed acoustically from test sentences, meaning that infant preference could not be based upon an acoustic match; it had to rely upon the prosodic parsing by the infants. Taken together, these findings suggest that the advantage of the well-formed sequences results from infants' use of prosody to parse continuous speech

Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, and Piwoz, (1992) demonstrated that at nine months old, infants prefer to listen to speech that contains pauses which are consistent with phrase boundaries over speech containing pauses that occur elsewhere in the sentence. In their study, infants heard passages comprised of seven to nine clauses, taken from a corpus of a mother interacting with her child. The passages were modified to have pauses that were either consistent or inconsistent with phrasal boundaries. Natural pauses that were over four seconds were removed from the passages, and one second pauses were added at phrasal boundaries (specifically, between the subject-noun and the verb, e.g. consistent – "*What happened? Did you*

*/ spill your cereal?"* vs. inconsistent – *"What happened? Did you spill / your cereal"*, where "/" denotes a pause). The results indicated that the infants preferentially attended to the consistent versions, providing evidence that 9-month-olds are sensitive to acoustic markers of clausal units, in particular pauses. However, in a follow up study, it was found that 6-month-olds were insensitive to the experimental manipulation, indicating a progressive ability to employ different prosodic markers across the developmental trajectory.

Taken together, the above studies support the prosodic-bootstrapping hypothesis. In both Nazzi et al. (2000), and Jusczyk et al. (1992), infants were able to recognize natural prosodic clauses, and preferentially attended to them, even in cases where acoustic matches were not possible. Critically, in the former, prosodic boundaries aligned with syntactic boundaries, supporting the notion that acoustic cues present in speech may facilitate the processing of syntactic structure. Here, the 6-month old infants had both pitch and temporal cues present, and thus could perform the task. However, in Jusczyk et al. (1992), pause cues were not useful cues for 6-month-olds. Thus, it may be the case that for the infant learner of English, that pitch cues are a more useful cue in earlier in development, with pause cues becoming accessible later. As descending pitch contours, final lengthening, and pauses are typically present at clausal boundaries, pitch cues may be more salient for English acquiring infants at an early developmental stage, with the usefulness of pauses being built upon pitch. Cross-linguistic comparisons can be informative in this regard: In a behavioural study by Seidl (2007), English acquiring infants were found to be sensitive to prosodic boundaries whether a pause cue was present or absent, however, in Männel and Friederici's (2009) ERP study using 5-month-old German acquiring infants, a pause at the end of a prosodic phrase was necessary to evoke the closure positive shift – a purely prosodic ERP that is elicited by the closure of a prosodic phrase. German has a larger number of inflections and a flexible word in contrast to English; thus, the functional demands on prosody may be greater for English speakers in highlighting phrasal structure (Männel & Friederici, 2009). Thus, pitch prosodic cues may be acquired earlier in English than German based on their functional importance within a language, with German acquiring infants acquiring pause prosodic cues at an earlier developmental stage. However, the relative importance of pitch and temporal cues may shift across development, or in cases where you are not processing your native language.

In a series of production studies, it has been demonstrated that the hierarchical structure of speech determines the types – and respective strength – of prosodic cues produced by speakers. Cooper, Paccia, and Lapointe (1978) employed a novel sentence reading method to study the influence of several different types of ambiguity on durational cues produced by speakers. In this paradigm, the speaker is provided with ambiguous sentences, and a corresponding semantic interpretation (e.g. Experiment 5: "Pam asked the cop who Jake confronted", (a) "Who did Jake confront?", (b) "Which cop? The cop that Jake confronted", see Figure 1 for the corresponding tree representations). Speakers first rehearse, then read the sentence aloud twice for recording. For the example provided, acoustic measures were taken of the duration of the key segment (/ka/) in "cop" (syllable lengthening), and the subsequent pause (pause duration). In this study, experiments one through six assessed whether speakers would employ stronger durational cues depending on the level of the syntactic hierarchy the critical word occurs in. Figure 1 illustrates that when "cop" occurs in the indirect question interpretation, it is at the third level of the syntactic hierarchy; it is a noun-phrase, nested within a prepositional phrase, which is in turn nested within the complex verb-phrase. When part of a relative clause interpretation, it is only at the second level of the hierarchy; a noun, occurring within a complex noun-phrase. The results, across all studies, demonstrated that when the measured syllable occurred at the closure of a deeper syntactic level, participants lengthened it more, and subsequently paused for a longer duration. This suggests that during production, speakers compute a representation of the syntactic structure of the utterance, that influences the temporal structure of the subsequent speech. Experiment seven assessed whether this was the case, or whether the findings were driven instead by audience design. Speakers now produced short narratives that provided a disambiguating discourse context, prior to the ambiguous sentence. The results remained the same; speakers produced longer syllables and pauses at deeper levels of the syntactic hierarchy, indicating that durational cues are driven by production processes, not audience design.

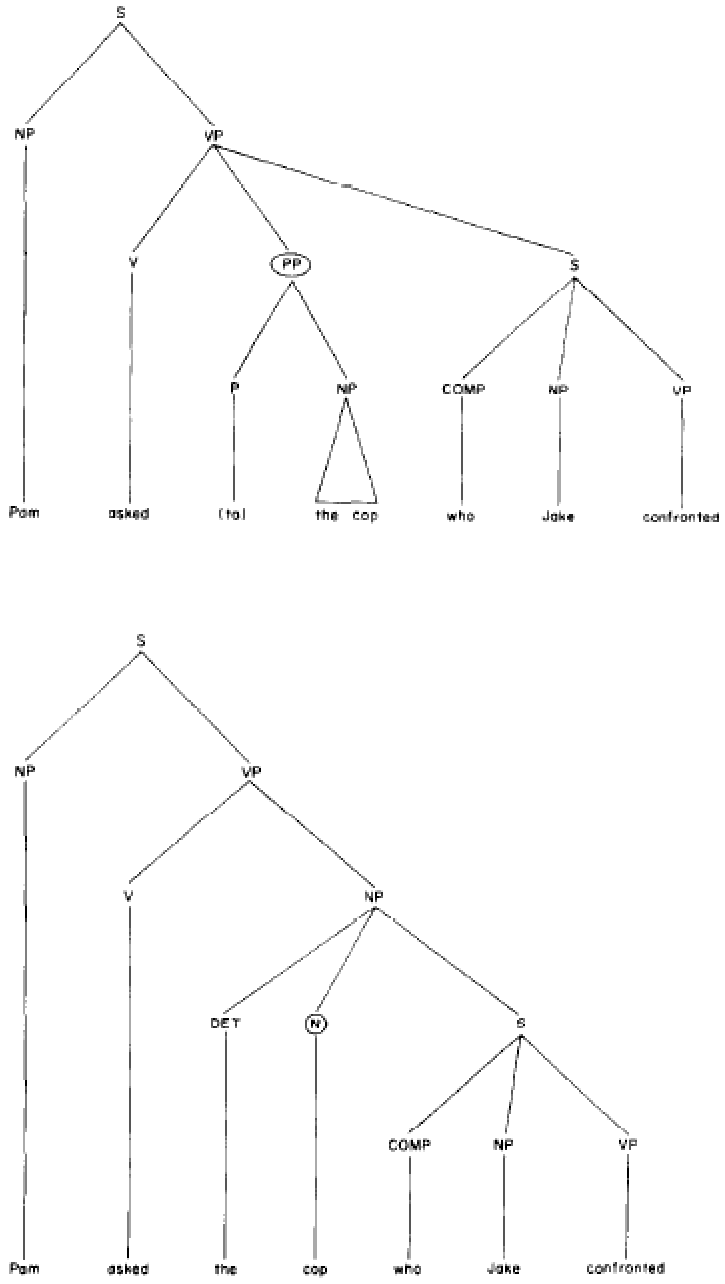


**Fig. 1**. Adapted from Cooper, Paccia, and Lapointe (1978), tree representations for an indirect question interpretation (top), and a relative clause interpretation (bottom) of the test sentence. In the top panel, "who Jake confronted" modifies the verb-phrase, in the bottom, it modifies the noun-phrase. In the top example, the prepositional phrase "(to) the cop", is embedded within the complex verb-phrase.

Kraljic and Brennan (2005) additionally found that disambiguating temporal cues are produced by speakers irrespective of the presence of an audience, or the presence of syntactic ambiguity. In their first experiment, they had participants perform a referential communication task in which speakers (hitherto directors) instructed listeners to manipulate a set of objects. Directors were provided with a picture that they viewed prior to issuing each instruction that pictures indicated which objects which objects were to move and where, and other objects that needed to be mentioned as part of the instruction. These instructions elicited syntactically ambiguous utterances in which the prepositional phrase (PP) could be interpreted as a modifier (in the utterance, *put the dog in the basket on the star*, “in the basket” could be used to specify a particular dog) or a goal (to first put the dog into a basket, and then place that on the star). The array could be ambiguous (contain a dog, a basket, and a dog sitting in a basket) or disambiguate the utterance (contain only a dog in a basket). Directors provided disambiguating prosodic cues; the first prosodic boundary (a relative measure, taken as the total duration of the noun phase and the following pause) was longer for the goal interpretation, and shorter for the modifier interpretation, regardless of whether the scene was ambiguous, supporting Cooper, Paccia, & Lapointe’s (1978) assertion that prosodic cues are generated during production, and not driven by audience design. Experiment 2 illustrated that participants were poor at judging the ambiguity of scenes. In Experiment 3, directors addressed a matcher, or took part alone. Again, the degree of prosodic boundary marking did not change, reinforcing the idea that durational cues do not reflect audience design, but are instead a feature of production planning.

Production based processing also influences the pitch contour of utterances. Cooper and Sorensen (1977) applied a similar research paradigm to investigate the differences in pitch contour at major phrase and syntactic boundaries. In experiment 1, the materials included sentence pairs (matched on phonetic environment and stress pattern) that either contained two conjoined main phrases, or a main clause and an embedded clause, e.g. (1a) "[$_{\text{C-1}}$ Marie was listening to the song] [$_{\text{C-2}}$ and Del was playing]", (1b) "[$_{\text{C-1}}$ Marie was listening to the song [$_{\text{C-2}}$ Adelle was playing]]". In (a) versions of the sentences, the internal syntactic boundary is between the end of the first main clause and the onset of the second. The internal syntactic boundary in the (b) versions is only the onset of the embedded clause. Three measurements were taken to assess fall-rise patterns in the $F_0$ contour: The peak $F_0$ value in *song* (the syllable prior to the major boundary), $P_1$, the lowest value in the same syllable, *V*, and the peak value in the stressed syllable following the boundary, ("delle" in *Adelle*), $P_2$. Across all test sentences, the reduction in $F_0$ between $\mathrm{P}_1$ and V was significant, as was the rise in $F_0$ between V to $\mathrm{P}_2$. Critically, however, there was a larger reduction between $\mathrm{P}_1$ and V, and a larger subsequent increase between V and $\mathrm{P}_2$, in conjoined sentences. Given that the sentences were matched on stress pattern and phonetic environment, it suggests that the effects are syntactically driven during speech production.

Whilst tonal and temporal cues may be indicative of speech production processes, these information sources may be critically important for comprehension and acquisition. Snedeker and Trueswell (2003) demonstrated that individuals can use temporal cues for disambiguation rapidly during comprehension. In their study, speakers were required to provide an instrument or modifier interpretation of a sentence ("Tap the frog with the flower", instrument; touch an empty-handed frog with a flower, modifier; touch a frog holding a flower). When producing an instrument instruction, speakers lengthened "frog", and paused for a longer duration following between "frog" and "with". When producing a modifier instruction, speakers paused for a longer duration following "tap". Listeners were able to rapidly use the appropriate durational cues identify the target interpretation in either the with-phrase (instrument prosody), or the onset of the direct object noun (modifier prosody).

Watson, Tanenhaus, and Gunlogson (2008) have also provided evidence that participants can rapidly use H* and L + H* pitch accenting on vowels to rapidly determine if a speaker intends to refer to a given or contrast item. Participants heard a series of commands (e.g. “Click on the camel and the dog. Move the dog to the right of the square. Now, move the *camel*/*candle* below the triangle”, where the underlined vowel is accented) to perform on a scene containing four shapes (e.g. *triangle*, *square*) and four objects, two of which were phonological competitors (e.g. *camel*, *candle*). The research question was whether pitch accents on the vowel would bias fixations towards the discourse new (newly mentioned, *candle*) or to contrast a previously mentioned item with a salient alternative (*camel*) element. Pitch accents rapidly affected processing; when exposed to an L + H* accented vowel, fixations increased to contrast items (“candle”), H* accents increased fixations to all potential referents with names consistent with the input, regardless of whether they were contrast or discourse new (e,g, “candy”). This suggests that individuals can rapidly use pitch information during comprehension. Thus, whilst it may be the case that pitch prosodic and durational cues may be driven by speech production processes, they can be critical for comprehension, and - as work with prosody in infancy shows – acquisition.

Learners’ sensitivity to the possible alignment of prosody with syntactic structure has been demonstrated in artificial grammar learning studies for both infant (Hawthorne & Gerken, 2014) and adult learners (Langus, Marchetto, Hoffman Bion, & Nespor, 2012), who were both found to draw on prosodic cues to assist processing of hierarchical grammatical structure. In Langus and colleagues’ study, each sentence consisted of two-clauses. Clauses were cued using final lengthening (the final pseudo-word of a clause was given a longer duration), and sentence-level prosody was cued with a descending pitch contour (the pitch of the initial syllable was the highest, and the pitch of the last-syllable the lowest). Learning was assessed using a two-alternate forced choice task, where participants were presented with either a rule-conforming novel phrase/sentence, vs. a familiar-part phrase/sentence. Transitional probabilities favoured adjacent dependencies (familiar part-phrases). Participants preferred novel rule-phrases and sentences over familiar part-phrases, indicating that participants were relying on prosodic over statistical cues. Critically, half of the participants were trained using the prosody simulating their native language, Italian, whilst the other half were trained with prosodic cues mimicking Japanese. Both groups performed above chance, demonstrating that experience with the prosodic cues was not required to employ it for acquisition of the artificial grammar. Similarly, Hawthorne, Mazuka and Gerken (2015) demonstrated that both Japanese and English acquiring infants could successfully use non-native prosody to acquire the experimental syntax. Taken together, it appears that prosody is salient and useful for acquisition and processing syntactic structure for both adults and children.

Many prosodic cues can be seen to relate to broader properties of auditory processing that are not specific to linguistic stimuli. For instance, it is well documented that learners tend to group auditory elements alternating in duration iambically (with the longest element last), whereas elements that alternate in intensity (strong to weak, or high to low pitch) are grouped trochaically (with the stressed element first, Hay & Diehl, 2007). Such grouping principles have been shown to extend beyond processing language structure – playing a similar role in the processing of musical structure - (Hay & Diehl, 2007; Frost et al., 2017), and are unlikely to be a consequence of transfer from language processing (e.g., Frost et al., 2017). Thus, an important literature to consider in comparison to prosodic processing is that of music perception. There are several reasons for this, primarily that the rhythmic properties of music and language share several features; both are grouped into phrases marked by pauses, as well as by differences in tone height and the durations of beats and syllables (Patel, 2003). Pitch-resetting at intonational boundaries can be seen as consistent with domain-general processing constraints, such as similarity – the likelihood that similar pitches are likely to be grouped together (Palmer & Krumhansl, 1987). Thus, a large change in pitch can be processed as closure on a group for structuring.

Indeed, several authors have proposed that shared perceptual or cognitive mechanisms are recruited in the acquisition (McMullen & Saffran, 2004) or processing (Patel & Iversen, 2007) of music and language. For the current piece, the key question is essentially; how do listeners group musical pieces into coherent sub-sequences? Western tonal music is often represented as tonal-temporal hierarchies; individual tones combine to form phrases, which then combine to form phrase groups, continuing to the level of the entire piece, all of which function according to a grammatical system (a musical style, or idiom) (Deutsch & Feroe, 1981; Farbood et al., 2015; Lerdahl & Jackendoff, 1983; Zhang, Jiang, Zhou, & Yang, 2016). In relation to the current study, we will primarily be concerned with two strategies for grouping perceptual information (i.e. Gestalts, or grouping rules), that of similarity, and proximity. According to the similarity Gestalt, in an array of five items, if three of these are orange, and the others blue, then you automatically perceive the array as two groups, one of blue items and one of orange items. According to the proximity Gestalt, if this array is instead made up of five identical objects, but two are close to one another, but more distant from the other three, then again you will perceive them as two groups, one of three, and one of two items that are close to one another, but distant as a group.

First, we will consider how the similarity Gestalt can applies to pitch processing. Individuals tend to form sequential links between tones that are close in pitch, and to distinguish between those that are further apart (Deutsch, 2013). Miller and Heise (1950), provide an excellent example of this. Participants were presented with two pure tones at different frequencies (A and B), delivered at a rate of 10 tones per second in an ABAB pattern. When the frequency difference between A and B tones was small, participants perceive the sequence as a trill; a single percept. However, when there was a large frequency difference between the tones, participants perceived two interrupted and unrelated tones; the perceived two distinct auditory percepts. This effect is quite robust, and has been demonstrated with more complex musical stimuli. Dowling, Lung, and Herbold (1987) investigated the role of pitch similarity on melody perception. Here, participants were presented with a novel target melody, followed by a probe melody that was interleaved with a distractor sequence. Participants made same/different judgements, and performance increased with larger pitch separations between the probe melody and distractor tones. Here we see evidence that pitch similarity can be a powerful factor in the context of grouping of distinct sequences. However, the timing of tones – their temporal proximity – plays an important perceptual role.

Grouping by temporal proximity has been shown to be the most powerful cue for the perception of musical phrase boundaries. Lerdahl and Jackendoff (1983), for example, proposed that grouping boundaries are placed at longer intervals between note onsets, and at changes in values of attributes including pitch range. Indeed, Deliège (1987) presented subjects with excerpts of Western classical music, and tasked them with marking boundaries between groupings. The boundaries participants chose corresponded to a high degree with Lerdahl and Jackendoff's grouping cues; the strongest effects were present following long notes (i.e. iambic groupings), with changes in timbre and dynamics also exerting influence. Given Zhang et al.'s (2016) observation that tonal systems operate according to a grammatical system, or idioms, we can pose the question of how these factors interact in cases where auditory input is hierarchically structured into melodies.

Dowling (1973) presented participants with patterns constructed with five-tone sequences separated by pauses. At test, participants made recognition judgements on whether test sequences were embedded in these patterns. Participants were more accurate at recognising a sequence that occurred in a single temporal period, and less so when a pause intervened; when temporal cues suggested the initiation of a new grouping, judgements on the basis of pitch suffered. Similarly, Hamaoui and Deutsch (2010) conducted a grouping preference study using stimuli where these cue types disagreed. However, it should be noted that this study only utilised four participants, and as a result, these conclusions should be interpreted with caution. Participants were presented with twelve-tone sequences, where pitch similarity suggested four groups of three tones, whilst the presence of pauses suggested three groups of four tones. Tonal groups were created by between group semitone distances of two, five or eleven semitones, and temporal groups were created with pauses of 15 to 60ms. As the distance between tone groups increased, participants were more likely to group the sequence on the basis of pitch; more dissimilar groups are more likely to be discriminated. However, participants had an increasing tendency to rely on temporal cues with pause durations over 30ms, even with large pitch distances. In a subsequent experiment, Hamaoui and Deutsch (2010) presented participants with sequences in which tones were either hierarchically structured or unstructured, but otherwise matched in pitch. Participants formed groupings based on hierarchical pitch structure, and these groupings were more robust to temporal cues than the unstructured sequences. Intriguingly, sequences that conformed to hierarchical structure, not simply pitch proximity, produced stronger groupings, implying that top-down preferences based on experience of musical systems drive grouping preferences. If we assume the similarity and proximity Gestalts reflect general properties of acoustic processing, then we can question whether they play a similar role in speech perception.

If such low-level, domain-general auditory processing constraints are found to be consistent with syntactic structure, then this opens up the possibility that syntax acquisition can be supported, or driven, by auditory Gestalts. If this is the case, then processing syntactic structures may be vastly simplified. For instance, Frank, Bod, and Christiansen (2012) suggested that sentence comprehension may be underwritten by sequential, rather than hierarchical processing. Under this account, to comprehend a hierarchical structure, the listener would need to rely on surface level cues (such as semantics) to determine the dependencies within the utterance, instead of processing the incoming works in a hierarchy. If sequential processing can be supported by low-level auditory cues, then this provides further support for the possibility of listeners processing syntactic structures without requiring complex hierarchical structure.

Centre-embeddings (e.g. “The rat the cat chases runs away”) have been extensively studied as a key example of hierarchical syntactic processing (de Vries, Monaghan, Knecht, & Zwitserlood, 2008; Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006; Lai & Poletiek, 2011) – where words are grouped in phrases (or ‘constituents’), which combine into higher–level phrases, up to the level of sentences - because they require a long-distance dependency (“The rat… runs away”) to be processed around an intervening centre-embedded phrase (“the cat chases”). In terms of their intonational properties, Fery and Schubö (2010) examined pitch-variance in centre-embedded German structures. In their study, participants read aloud sentences of the form, “[$_{\text{c-0}}$ The pears [$_{\text{c-1}}$ which at the tree [$_{\text{c-2}}$ which green is] hang] are sour]”, where the peak pitch of each underlined word was measured, and these were compared against sentences with no embeddings. Data indicated that the subject noun, and the second part of c-0 (where c-x indicates embedding at each level, so c-0 indicates no embedding, and c-1 indicates an embedded phrase) possessed the highest and lowest-pitch respectively, signaling the start and end of the utterance (lowest pitch usually occurs utterance-finally to indicate the final element, Beckman & Pierrehumbert, 1986). Interestingly, the first half of c-1 saw a significant drop in pitch, which occurred again at the start of c-2, followed by a pitch reset when the second half of c-1 was voiced. Notably, the two-parts of c-1 were produced in a similar pitch range relative to c-0 and c-1, i.e. pitch seems to have reflected the grouping of the two-phrases in the clause. Further, a declining pitch contour provided evidence for a trochaic grouping. Hence, in Fery and Schubö's (2010) analysis, there are several potentially useful prosodic cues that could assist listeners in grouping non-adjacent structures that draw on acoustic processing principles rather than requiring hierarchical phrase structure to determine the dependencies in the sentence.

However, Fery and Schubö (2010) gave participants sentences to read, whereas the potential availability of these cues may be very different in spontaneous speech. Here, we examined the degree to which pitch systematically varies during relative clause production in native-English-speaking adults. Specifically, we compared the influence of syntactic form (active vs. passive), and sentence position on pitch and pause variation. Our study therefore addressed two key questions: Do pitch and temporal cues vary systematically on the basis of structure, facilitating the processing of that structure; and do utterance boundaries correspond with structural boundaries? If so, this may obviate the need to process hierarchical centre-embeddings hierarchically by supporting the application of lower-level acoustic processing to support the identification of dependencies in the sentence, consistent with a sequential processing strategy (Frank et al., 2012). We hypothesized that (1) words spoken in phrasal units containing syntactic dependencies will be more similar in pitch, enabling grouping according to the Gestalt similarity principle. Given that humans are sensitive to a semitone difference of 0.8  (Dowling & Harwood, 1986), we predict a difference of at least one semitone between syntactically unrelated phrases.

Further (2) that pause duration should reflect the Gestalt principle of proximity: pauses occurring between clauses will render those clauses distinct if they are longer in duration than elsewhere in the speech, and (3) that pauses should be more likely to occur between clauses than elsewhere in the speech.

To assess the usefulness of prosody in comprehension from speakers' relative clause productions, we conducted an analysis on speech data from a picture description task, conducted by Montag and Macdonald (2014). In their study, participants were presented with a series of scenes (20 scenes total, see Figure 2), and were required to answer questions relating to particular objects that appeared within them. Each scene contained two competing depictions of events involving the same action, and two competing depictions of the target object (one animate, and one inanimate). Critically, describing one of these instances encouraged participants to use the appropriate verb in the active voice, whereas describing the other influenced participants to use the passive voice. Scenes therefore elicited production of relative clauses with either an active or a passive form. Using these data, we assessed whether the pitch and temporal dynamics of the speech would vary on the basis of the two syntactic forms and their dependency relations.

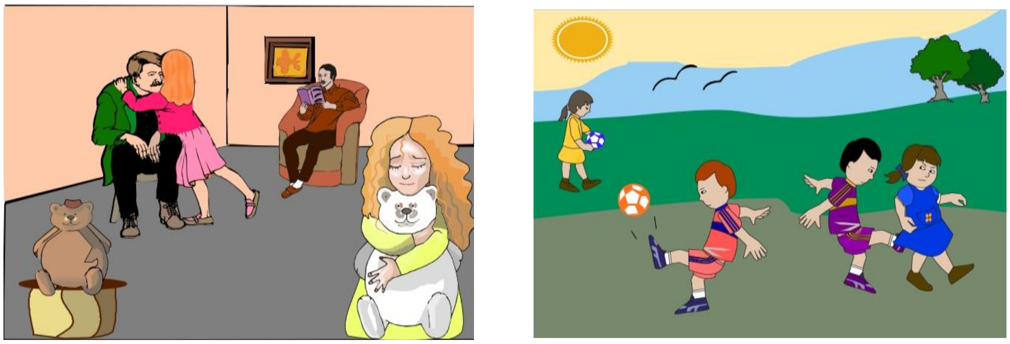


**Fig. 2**. Example stimuli from Montag & MacDonald (2014). In response to the left scene, participants were likely to produce an active, "The bear the girl is hugging is white". In response

to the right scene, participants were likely to produce a passive, “The girl being kicked by the boy is wearing blue”

## 2. Method

### 2.1 Data

The data were taken from Montag and MacDonald’s (2014) study, where English-speaking participants provided descriptions for visual scenes, designed to elicit relative clause completions from participants. The items within the scene varied the animacy of targets and competitors to determine the influence of these visual features on the structural choices made by participants. Each participant described 20 scenes, giving one sentence for each scene (so, each participant provided data for 20 sentences in total).

Participants completed 20 trials. In each, participants were given a probe question, focused on one item within the picture. For example, the probe question for the left scene in Figure 1 would be, “Which bear is white?”. The scene depicts a white bear being hugged by a girl, a man being hugged by a second girl - an action/animate target competitor -, a bear on the left - target distractor -, and an unrelated distractor in the rear of the scene. As a result, participants are implicitly encouraged to foreground information about the bear, and its distinguishing feature (that it is being hugged by the girl), increasing the likelihood of producing a relative clause. In this example, with an animate agent and inanimate patient, participants were more likely to produce an active-object relative clause, e.g. “The bear (that) the girl is hugging (is white)”. In the right scene, we see a girl wearing a blue dress being kicked by kicked by a boy, another boy kicking a ball – an action/inanimate competitor -, and a girl playing in the background. The corresponding probe to this question would be “Which girl is wearing blue?” In scenes like this, depicting an animate agent and patient, participants were more likely to produce passive completions, e.g. “The girl (who is) being kicked by the boy (is wearing blue)”. In each case, the inclusion of the relative pronoun is optional, as it is not required to produce a grammatical utterance. Similarly, participants often omitted “is white/is wearing blue”, due to this information being provided in the probe questions.

In our processing of the data, we distinguished whole and part sentences, and active and passive sentence constructions that each speaker produced. A whole sentence completion (e.g. “The book the girl is reading is green”) was characterized as containing information posed in the trial question (e.g. “Which book is green?”), whereas a part sentence did not (e.g. “The book the girl is reading”).

Data for the 64 participants (hence referred to as speakers) who took part in the original study were provided. Two speakers were removed from analysis due to producing solely highly complex syntactic structures (e.g. “The lady being held by the man in the green hat, green pants and green shoes is wearing red”), or simple noun phrases (e.g. “The lady”). Further, individual trials including recording errors (e.g. participant failed to complete the utterance within the recording period) were eliminated from analysis.

## 2.2 Data analysis

The data were prepared using the acoustic analysis software Praat (Version 6.0.13; Boersma, Paul & Weenink, 2016). Utterances were prepared for analysis using the Prosogram package (Version 2.13; Mertens, 2016). This package was used to automatically segment utterances into phonemes and syllables, and pauses. This procedure utilizes changes in the spectrum (sound timbre) and intensity. The resulting text grid was then used edited to include a word level (informed by syllabic boundaries). Then boundaries were manually inspected and corrected where necessary. For a pause to be defined as such, we used the simple criterion that there was no audible speech in that segment, and that if a pause occurred between two plosives (p, b, g, d, t, k), the boundary for the second word would begin at the conclusion of the first plosive. Utterances were coded such that words were indexed on the basis of which phrase they appeared in (e.g. "[$_1$ The bear] [$_2$ the girl] [$_3$ is hugging] [$_4$ is white]", where the numbered subscripts index phrasal position) for further analysis.

The first phrase for all productions was always a noun phrase. In active productions, the second phrase was the relative clause noun phrase, i.e. the noun phrase of the embedded clause. In actives, phrase three was always the relative clause verb phrase, which contains a dependency relation to the second noun phrase (phrase 2). For actives, phrase four was always the verb phrase of the main clause, which shares a dependency relation with phrase one. Passives differed in their construction (see Figure 3 for the tree diagrams for each syntactic structure). Phrasal position two was always was always the verb phrase of the relative clause, which critically share a dependency relationship with the initial noun phrase. As a result, the first two phrases of a passive constitutes a grammatical utterance, in contrast to actives (e.g. "The bear being hugged" vs. "The bear the girl"). The third phrase in passive constructions was an optional agentive prepositional by-phrase, that attached to the verb-phrase of the relative clause. Similar to active constructions, the final phrase of passives was the verb phrase of the main clause.

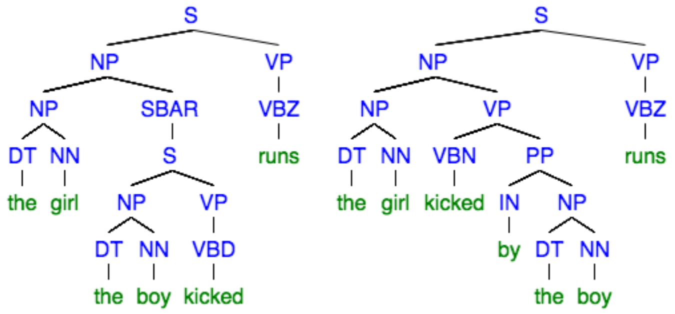


**Fig. 3.** Syntactic trees for reduced active-object (right) and passive (left) relative clauses.

To analyse the degree to which speakers' prosody systematically varied with syntactic structure, several measures were utilised. For each utterance, we calculated the mean pitch (measured in $F_0$hz) per word and its duration (ms), in addition to the duration of any pauses occurring between phrasal positions (ms), that were subsequently coded on the basis of pause location (e.g. a 1 – 2 pause occurred between phrases 1 and 2). To assess whether pauses at clausal boundaries and their duration were more governed by constraints on the vocal system (i.e. a finite amount of air in the respiratory system), for each phrase we calculated the voiced phrase duration (ms) (total phrase duration – phrase internal pause duration). When a pause did not occur in an inter-phrasal position, we additionally coded this as being a pause with no duration – a zero-pause. We reasoned that if a pause at a given location was a useful cue to support syntactic processing, a pause would be more likely to occur. By coding the data to include zero-pauses, we were thus able to assess the probability of pause occurrence in each location. Three analyses were conducted using linear mixed-effects modeling (Baayen et al., 2008) using the lme4 package in R (Bates, Maechler, & Bolker, 2011), which assessed (1) pitch variance in relative pitch (semitone distance from middle C) (2) relative pause duration, and (3) likelihood of pause occurrence on the basis of phrasal position and syntactic form. In each case, either pause type or phrasal position were taken as multi-leveled factors, where comparisons of interest are sequential, i.e. if the baseline predictor was phrasal position 1, the key comparison is with phrasal position 2, if the baseline is position 2, the comparison is with 3, and so on.

## 3. Results

### 3.1 Analysis 1: Semitone Pitch Variance

A common measure of pitch is the semitone which measures relative pitch change to which listeners are sensitive, rather than absolute changes. Whilst hertz, or cycles per second, are the physical correlate of pitch, it can also be represented in terms of musical scales. In this analysis, we focus on pitch in terms of musical scales. If the frequency of any given tone is doubled, it is separated by an octave. These two tones are perceived as similar, an observation that is consistent across cultures (Patel, 2008), with even novice listeners (Dowling & Harwood, 1986) and monkeys (Wright et al., 2000) being sensitive to this relationship, suggesting that the musical system reflects the neurophysiology of the auditory system (e.g. McKinney & Delgutte, 1999). In Western European music, each octave is comprised of 12 equal-sized intervals, with each note being approximately 6% higher in frequency than its predecessor (Patel, 2008). This interval is known as a semitone. For ease of interpretation for the current analysis, we computed the semitone distance for each component word of an utterance from middle C (hitherto $C_4$) on a standard MIDI keyboard, i.e. 261.626hz. To compute the distance, we used the following equation:

$$12 \times log_2(\frac{x}{261.626})$$

Where $x$ is the mean $F_0$Hz value for a given word. Barring this transformation, the analysis remains the same as that for $F_0$hz.

The descriptive statistics (see table 1 for the descriptive statistics) reveal similar observations as the analysis of $F_0$Hz. For actives, the mean semitone distance from $C_4$ is the lowest in the sentence (as the $F_0$Hz value for $C_4$ is higher than all $F_0$Hz values produced by the participants, the semitone units are negative) in phrase 1, similar to passives. For active structures, at phrase 2 there is a 1.31 semitone increased distance from $C_4$ for actives. For passives there is an increase of 0.88 semitones, a comparatively smaller increase. In phrase 3, active structures are 0.87 further from $C_4$. In passives there is an increase of 1.94, over a semitone larger than the comparable increase in active structures. Moving to phrase 4, the semitone distance for actives increases by a further 0.47, and in passives there is an increased distance of 0.71. To summarise; phrases 1 and 4 are the closest and furthest from $C_4$ respectively, an increased semitone distance from $C_4$ at phrase 2 for actives relatives to passives, and a larger decrease from phrases 2 to 3 for passives, relative to actives. In terms of the similarity gestalt, this should result in a greater likelihood of grouping phrases 2 and 3 in active structures (the embedded clause), and a greater likelihood of grouping phrases 1 and 2, and 3 and 4 in passives.

**Table 1**

*Descriptive statistics of semitone distance from $C_4$ by phrase and syntactic form.*

| Syntax | Phrase | Mean Semitone Distance | Std. Deviation |
|---|---|---|---|

|  |  |  |  |
|---|---|---|---|
| Active | 1 | -7.29 | 2.11 |
| Active | 2 | -8.6 | 2.91 |
| Active | 3 | -9.47 | 4.11 |
| Active | 4 | -9.94 | 5.41 |
| Passive | 1 | -7.39 | 2.62 |
| Passive | 2 | -8.27 | 2.79 |
| Passive | 3 | -10.21 | 3.8 |
| Passive | 4 | -10.92 | 5.52 |

To assess these dynamics formally, we utilized a linear mixed effects models assessing the mean semitone distance from $C_4$ per word, predicted by syntactic form (Active = 0.5, Passive = -0.5), phrase (1 – 4, coded as a four-level factor), and their interaction. The model included subjects, items, and voiced phrase duration (as a longer phrase has more time in which for pitch to reduce), with random slopes for syntactic form for each random effect. Models were built iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). All random effects structures were built in a forward manner (first intercepts only, then adding random slopes, until models failed to converge), and in each case, the maximally convergent model is reported. To assess the difference between phrases, models were re-levelled such that each phrase was taken as the baseline predictor, revealing the difference between each level of the factor. Table 2 presents the summary of the maximal model.

The model revealed a significant main effect of phrase; significant differences were found between phrases 1 and 2 (Estimate =-1.1334, SE = -.132), 2 and 3 (Estimate = -1.4861, SE = -0.1302), and 3 and 4 (Estimate =-1.6042, SE = 0.2683), reflecting the global trend to produce words in successive phrases at an increased semitone distance from $C_4$. There was a significant main effect of syntactic form for the comparison of phrases 2 and 3 (Estimate = -0.4978, SE = 0.2223), and phrases 3 and 4 (Estimate = 0.6726, SE = 0.2289). Thus in these phrases, key differences emerged between the two forms. The interaction between phrase and syntactic form was significant only for the contrast between phrasal positions 2 and 3 (Estimate = 1.1704, SE = 0.2605, see figure 5 for greater detail), reflecting the fact that in passives, we see a larger increase in distance from $C_4$ between these phrases than we do for actives.  Taken together, we can conclude that, in active structures, the similarity Gestalt should promote a grouping of phrases 2 and 3, binding them together, and facilitating comprehension of the dependencies in the hierarchical structure. In passives, pitch similarity is highest between phrases 1 and 2, and 3 and 4, which in theory, should support comprehension by combining phrases 1 with 2, and 3 with 4.

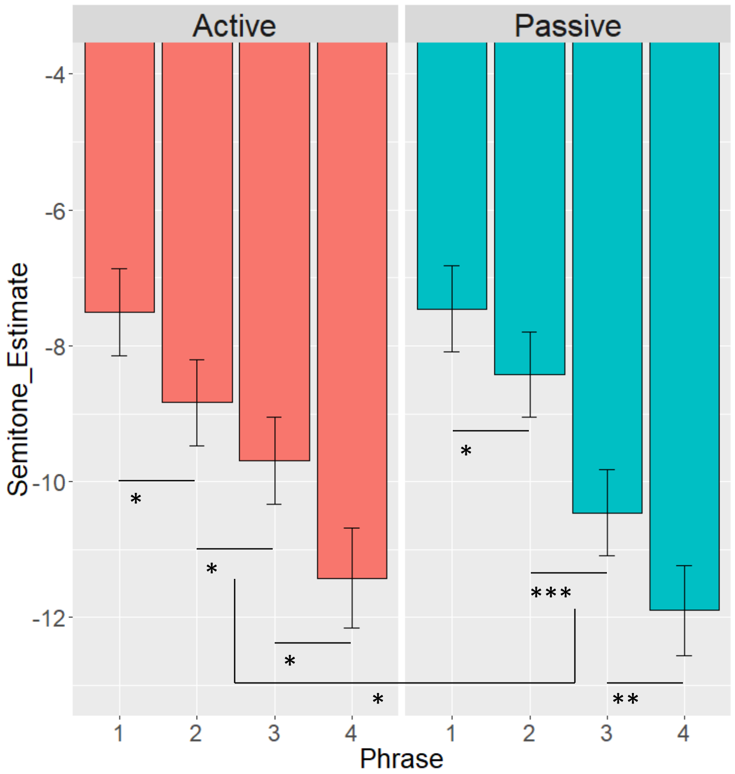


**Fig. 5**. Model estimates of semitone distance from middle C on the basis of syntactic form and phrasal position. The left panel illustrates the data for active constructions, and the right for passive constructions. Black vertical bars display the standard error of the semitone estimate. Black horizontal bars with "*" above them illustrate significant differences (t > 2 & < 5 = *, t > 5 & t < 10 = **, t > 10 = ***).

**Table 2**

*Results of mixed-effects model predicting Semitone distance from middle C by phrasal position and syntactic form. The model contained intercepts and by-subjects and items slopes for syntactic form.*

| Fixed Effect | Baseline | Estimate | Std. Error | *t* |
|---|---|---|---|---|
| Form | Phrase 1 | -0.150 | 0.237 | -0.632 |
| Phrase 2 | Phrase 1 | -1-133 | 0.132 | -8.586** |
| Phrase 3 | Phrase 1 | -2.620 | 0.136 | -19.279*** |
| Phrase 4 | Phrase 1 | -4.224 | 0.269 | -15.703*** |
| Form: Phrase 2 | Phrase 1 | -0.348 | 0.264 | -1.319 |
| Form: Phrase 3 | Phrase 1 | 0.822 | 0.272 | 3.027* |
| Form: Phrase 4 | Phrase 1 | 0.630 | 0.516 | 1.221 |
| Form | Phrase 2 | -0.498 | 0.222 | -2.239* |
| Phrase 1 | Phrase 2 | 1.133 | 0.132 | 8.586** |
| Phrase 3 | Phrase 2 | -1.486 | 0.130 | -11.418*** |
| Phrase 4 | Phrase 2 | -3.090 | 0.266 | -11.616*** |
| Form: Phrase 1 | Phrase 2 | 0.348 | 0.264 | 1.319 |
| Form: Phrase 3 | Phrase 2 | 1.170 | 0.261 | 4.493* |
| Form: Phrase 4 | Phrase 2 | 0.977 | 1.919 | 1.919 |
| Form | Phrase 3 | 0.673 | 0.229 | 2.939* |
| Phrase 1 | Phrase 3 | 2.619 | 0.136 | 19.279*** |
| Phrase 2 | Phrase 3 | 1.486 | 0.130 | 11.418*** |
| Phrase 4 | Phrase 3 | -1.604 | 0.268 | -5.98 |
| Form: Phrase 1 | Phrase 3 | -0.822 | 0.272 | -3.027 |
| Form: Phrase 2 | Phrase 3 | -1.170 | 0.261 | -4.493 |
| Form: Phrase 4 | Phrase 3 | -0.193 | 0.512 | -0.377 |

Model Syntax: F0 ~ (1 + Form:Phrase|Subject) + (1 + Form:Phrase|Item) + (1 + Form:Phrase|Voiced Phrase Duration) + Form + Phrase + Form:Phrase

### 3.2 Analysis 2: Pause Duration

The second analysis assessed whether relative pause duration differed on the basis of phrasal position and syntactic form. To account for the fact that the utterances varied naturally in duration by speaker, and that unfilled pauses are likely to reflect a combination of utterance planning and constraints on the vocal system (Kraljic & Brennan, 2005) such as breathing, and phonetic variation (Ferreira, 2002) we computed each pause as a percentage of the duration of the entire utterance. For active structures, pauses occurring between phrases 1 and 2 (see Table 3 for descriptive statistics) were comparatively long compared to those occurring between phrases 2 and

3 (difference = 1.413%), which were in turn longer than those occurring between phrases 3 and 4 (difference = 2.715%). For passive structures, pauses occurring between phrases 1 and 2 were shorter than those occurring between phrases 2 and 3 (difference = 2.654), which in turn were longer than those between phrases 3 and 4 (difference = 3.579%).

**Table 3**

*Mean pause duration as a proportion of the entire phrase, by pause location and syntactic form*

| Syntax | Phrases paused between | Mean % of utterance duration | Std. Deviation |
|---|---|---|---|
| Active | 1-2 | 4.391 | 10.977 |
| Active | 2-3 | 2.978 | 12.975 |
| Active | 3-4 | 0.263 | 1.470 |
| Passive | 1-2 | 1.394 | 4.598 |
| Passive | 2-3 | 4.048 | 7.814 |
| Passive | 3-4 | 0.489 | 1.628 |

To evaluate these differences, we employed linear mixed effects models assessing pauses as a percentage of the entire utterance on the basis of syntactic form (Dummy coded; passive = -0.5, active = 0.5), and pause location (between 1 and 2, 2 and 3, and 3 and 4, hitherto; 1 – 2, 2 – 3, and 3 – 4, coded as a factor). This model included random intercepts for subjects, items and voiced phrase duration, with random by syntactic form intercepts and slopes for subjects and items. The main effect of pause location was significant (|*t*|s < 2); 1-2 pauses were overall shorter than 2-3 pauses (β = 2.32, SE = 0.72), mainly driven be the long duration of 2-3 pauses in passive structures. 3-4 pauses were longer than 2-3 pauses (β = -3.21, SE = 0.64). The main effect of syntactic form was significant for 1-2 pauses, showing a syntactically driven difference in this location, whereby active structures tended to longer on average. Critically, however, the interaction between pause location and syntactic form was significant when using 1-2 pauses as a baseline predictor, demonstrating that 2-3 pauses for passives tended to be longer on average than those for actives (β = -3.150, SE = 1.154, see figure 6). In regards to the temporal proximity Gestalt, in actives this should deter participants from grouping the first two phrases of the utterance. The lower pause duration between phrases 2 & 3 should facilitate the grouping of the phrases of the internal clause; temporal proximity suggests the phrases of the embedded clause should be grouped. However, in passives, longer pauses occur between phrases 2 and 3 than elsewhere in the sentence, which should bias participants towards grouping the first two phrases of the utterance, agreeing with the grouping suggested by pitch similarity.

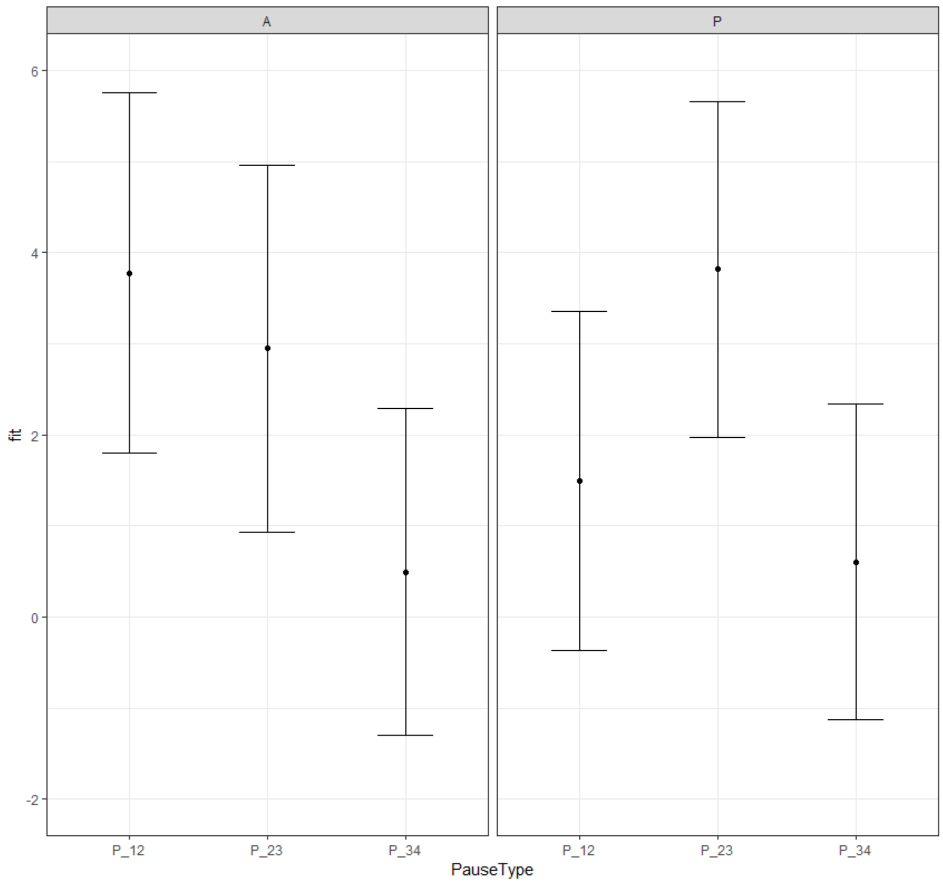


**Fig. 6.** Model estimates of % duration of the entire utterance for pause location by syntactic form. The left panel illustrates the data for active constructions, and the right for passive constructions. Black vertical bars display the standard error of the log duration estimate.

**Table 3**

*Results of final mixed-effects model predicting pause duration from pause location and syntactic form. The model contained random intercepts, and by-subjects, by-items and by-whole vs. part slopes for syntactic form.*

| Fixed Effect | *Baseline* | β Coefficient | Std. Error | *t* |
|---|---|---|---|---|
| 2 - 3 | 1 - 2 | 2.320 | 0.718 | 3.234* |
| 3 - 4 | 1 - 2 | -0.891 | 0.643 | -1.386 |
| Form | 1 – 2 | 2.281 | 0.643 | 2.684* |
| 2 – 3: Form | 1 – 2 | -3.150 | 1.154 | -2.730* |
| 3 – 4: Form | 1 – 2 | -2.390 | 1.005 | -2.379* |
| 1 – 2 | 2 – 3 | -2.320 | 0.718 | -3.234* |
| 3 – 4 | 2 – 3 | -3.211 | 0.644 | -4.986* |
| Form | 2 – 3 | -0.869 | 0.874 | -0.994 |
| 1 – 2: Form | 2 – 3 | 3.150 | 1.154 | 2.730* |
| 3 – 4: Form | 2 – 3 | 0.760 | 1.021 | 0.745 |
| 1 – 2 | 3 – 4 | 0.891 | 0.643 | 1.386 |
| 2 – 3 | 3 – 4 | 3.221 | 0.644 | 4.986* |
| Form | 3 – 4 | -0.109 | 0.669 | -0.163 |
| 1 – 2: Form | 3 – 4 | 2.387 | 1.005 | 2.379* |
| 2 – 3: Form | 3 – 4 | -0.760 | 0.121 | -0.745 |

Model Syntax: Scaled Pause Duration ~ (1 + Form:Pausetype|Subject) + (1 + Form:Pausetype|Item) + Form + PauseType + Form:Pausetype

### 3.3 Analysis 3: Pause Likelihood

The third analysis we conducted assessed the likelihood of a pause having a non-zero duration on the basis of syntactic form and pause location, see figure 7 for pause proportions. The proportion of non-zero pauses between phrases one and two is higher for active relative to passive structures (29.6% vs. 17%). Passives had a higher proportion of non-zero pauses between positions two and three (47.7% vs. 4.3%), and three and four (10% vs. 4.3%).

To formally assess the influence of syntactic form and pause location on the likelihood of pause occurrence, we used a generalized linear mixed effects model (GLMER), predicting the binary variable (non-zero vs. zero pause) of pause occurrence. In this case, we used the binomial distribution using a logit-link function. Again, pause location was coded as a four-level factor (1-2, 2-3, 3-4), and syntax was dummy coded (active = 0.5, passive = -0.5). The models included by-items, by-subjects, and by-voiced phrase duration random intercepts. Only models including random intercepts are reported here, as models including random slopes did not converge, making their results uninterpretable. The results revealed no significant effects (Pr|>z| > 0.05). However, the interaction of syntactic form by pause location did approach significance for pauses occurring between phrases one and two (Estimate = 3.951, SE = 1.791, *p* = 0.073, see figure 7), reflecting the greater proportion of non-zero pauses in active structures, and for pauses occurring between phrases two and three (Estimate = -3.951, SE = 2.21, *p* = 0.073), reflecting the higher proportion of non-zero pauses in this location for passives. Overall, in cases where pauses do occur in actives, they are most likely to occur between positions one and two, notably this matches the location of the largest reduction in pitch, which may allow the proximity Gestalt to reinforce the grouping suggested by pitch similarity. Similarly in passives, pauses are most likely to occur between positions two and three, again matching the largest pitch change. However, these effects are highly marginal, and should thus be interpreted with caution.

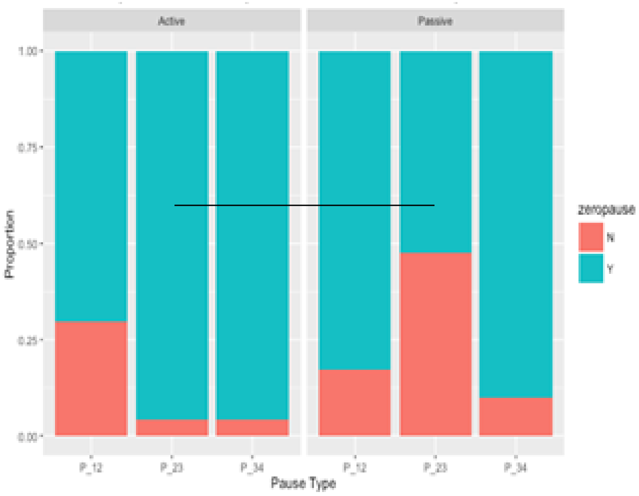


**Fig. 7.** The proportion of non-zero vs. zero-pauses on the basis of syntactic form. The left panel illustrates the proportions for active structures, and the right panel illustrates the data for passive structures. The red area of the bar indicates the pause had a non-zero duration, i.e. it was not a

placeholder, while the blue area illustrates the proportion of placeholder, zero pauses. The horizontal black line illustrates the interaction between pause type and syntactic form that trended towards significance.

**Table 4**

*Results of mixed-effects model predicting pause likelihood from pause location and syntactic form. The model contained random intercepts but no random slopes.*

| Fixed Effect | Baseline Predictor (Pause type) | Estimate | Std. Error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|---|
| 2 - 3 | 1 - 2 | 1.517 | 1.094 | 1.387 | .1655 |
| 3 - 4 | 1 - 2 | -1.280 | 1.775 | -0.721 | .4709 |
| Form | 1 - 2 | 0.775 | 1.282 | 0.605 | .5455 |
| 2 – 3: Form | 1 - 2 | -3.951 | 2.206 | -1.791 | .0733* |
| 2 – 3: Form | 1 - 2 | -2.372 | 3.179 | -0.746 | .4556 |
| 1 – 2 | 2 – 3 | -1.517 | 1.094 | -1.387 | .1655 |
| 3 – 4 | 2 – 3 | -2.797 | 1.769 | -1.581 | .1139 |
| Form | 2 – 3 | -3.176 | 2.045 | -1.553 | .1205 |
| 1 – 2: Form | 2 – 3 | 3.951 | 2.206 | 1.791 | .0733* |
| 3 – 4: Form | 2 – 3 | 1.579 | 3.592 | 0.440 | .6602 |
| 1 – 2 | 3 – 4 | 1.280 | 1.774 | 0.721 | .471 |
| 2 – 3 | 3 – 4 | 2.797 | 1.768 | 1.582 | .114 |
| Form | 3 – 4 | -1.597 | 3.162 | -0.505 | .614 |
| 1 – 2: Form | 3 – 4 | 2.372 | 3.174 | 0.747 | .455 |
| 2 – 3: Form | 3 – 4 | -1.579 | 3.588 | -0.440 | .660 |

Model Syntax: NonZeroPause ~ (1 + Form:PauseType|Subject) + (1 + Form: PauseType) + Form + PauseType + Form: PauseType

## 4.0 Discussion

In the current study, we analyzed speakers' data from Montag and MacDonald's (2014) relative clause elicitation study. In their study, participants described visual scenes (see figure 2), in response to probe questions. Here, we analyzed the temporal and pitch dynamics of active-object and passive relative clauses provided by speakers. We had hypothesized that syntactically dependent phrases would be more similar in terms of pitch, assisting dependency detection. Further, that speakers would pause for a longer duration in sentence positions consistent with clausal boundaries, rendering syntactically dependent phrases temporally distinct. These hypotheses reflected two auditory-perceptual Gestalt principles (or grouping behaviors): Pitch similarity, and temporal proximity. The former states that the more similar in pitch two sounds are, the more likely they are likely to form a grouping, whilst the latter states that the more temporally proximate two sounds are, the more likely they will be grouped together.

The results revealed the auditory cues present in active-object relatives reflect these Gestalts; phrases occurring within the embedded clause were more temporally proximate and more similar in pitch. Thus, the two auditory-perceptual Gestalts should facilitate the grouping of the phrases of the embedded clause, whilst distinguishing it from the first phrase of the external clause. In passive relative clauses, however, the results differed. Words spoken in phrases 1 and 2 - and 3 and 4 - were both more temporally proximate and more similar in pitch. On the basis of the two Gestalts we have hypothesized may help to guide comprehension, this should hinder comprehension. In the introduction, we raised the question of whether pitch and temporal dynamics systematically vary on the basis of structure, and whether this will assist comprehension. Whilst this seems to be clearly present in the active-object constructions, the results for passives are less clear, and seem counter-intuitive. Further, we asked whether utterance and structural boundaries align. Again, this seems to be the case in actives, where the embedded clause (phrases

2 & 3) is clearly delineated through pitch and temporal cues. However, in the passives, where phrases 2 and 3 carry similar information and meaning, this is not the case. In the following discussion, we will address these points.

First, we will consider the predictions and implications of a sequential processing account proposed by Frank, Bod, and Christiansen (2012). In terms of production, utterances are generated from constructions (see Construction Grammar, Goldberg, 2006), which are pieces of linguistic forms paired with meaning. These constructions - in their most basic form - are individual word-meaning pairs, e.g. a noun (*brush*), combined with the corresponding mental representation of a brush. Constructions can also be comprised of multiple words (e.g. *dustpan and brush*), where a frequently occurring word sequence can become merged into its own construction. Further, constructions can contain abstract elements that can openly correspond to noun phrases, e.g. "*pick X up", "I bought X",* etc.

Building a sentence thus corresponds to creating a sequence out of these constructions. Frank, Bod, and Christiansen (2012) suggest this occurs by switching between multiple, sequential streams that run in parallel, where one stream may contain *put x down,* a second *knife and fork* - the noun phrase corresponding to *x* - and a third, including *your*, the combination of which results in *put your knife and fork down*. Turning back to the current study, each utterance is comprised of a number of constructions. Active-object relatives could be construed as "[The A] [the B is C] [is D]", where A and B are nouns, C is a verb that frequently corresponds to agent B, and D is an adjective describing A. In passives, a different sequential structure could be construed with "[W being X] [by the Y] [is Z]" where, departing from the active-object example, W is a noun being acted upon in a frequent manner X, thus, it can be a multi-word construct. "[By the Y]", due to not being required to form a grammatical statement may stand apart from "[W being X]". In actives, however, “[The A]” is unlikely to be part of a construction containing “[the B]”. Why? “[The A the B]” lacks an action of some sort, which is particularly salient when producing sentences about an animate and inanimate noun (a girl and a teddy bear). A more frequent construct containing the relationship between nouns A and B, would be “[A’s B]” (The girl’s bear”), or “[A and B]” (“The bear and the girl…”). Thus, the present set of results suggest that utterance boundaries (defined by large changes in pitch, and longer pauses), are most consistent with the initiation of new constructions in sequential structure. In passives, the results thus conflict with the dependency shared between phrase two (the agentive verb phrase), and the prepositional by-phrase (phrase 3). In actives, this results in agreement between the auditory features and sequential structural boundaries, with the agentive noun- and verb-phrases (phrases 2 and 3) set apart from the patient noun-phrase (phrase 1). To verify whether these results are utilized in comprehension will, however, require further study. Here, we can only say that these cues are present in production of relative clause structures.

Another potential explanation can be drawn from prosodic rules in English (e.g. Fodor & Inoue, 2000). Crucially, this account of prosody relies on the principle of incremental comprehension; due to the temporally transient nature of acoustic signals, they must be processed immediately upon being encountered. Thus, processing prosody is necessarily incremental, and is more in tune with ideas of sequential linguistic processing than a hierarchical account. Dekydtspotter (2008) notes that when a relative clause functions as noun modifier - i.e. adjusting the meaning of the noun phrase – it is integrated into the phonological phrase from the noun. As an example, presented with the sentence, “We adore the secretary of the psychologist who takes a walk”, the prosodic segmentation tends to be, “[$_{\text{Utterance}}$ [$_{\text{Intonational Phrase}}$ [We adore] [$_{\text{Intonational Phrase}}$ the secretary] [$_{\text{Inonational Phrase}}$ [$_{\text{Phonological Phrase}}$ of the psychologist who… $_{\text{Phonological Phrase}}$]]]” In passives, the relative clause, ("The bear being hugged…"), immediately functions to modify the sentence initial noun-phrase, before introducing the second noun phrase. Noun-phrases project their own phonological phrases. In the active-object structures, this mandates a new prosodic grouping starting at the second noun-phrase. As a result, it should be preceded by an optional pause, and paired with a pitch-reset. Recalling the data from Fery & Schubö (2010), in active-object relatives there was a reduction in pitch between the first noun phrase, and the initial part of the second noun-phrase. This would agree with the data presented here; the second noun-phrase sees a reduction in pitch, constituting a new prosodic grouping. In passives, however, as the verb-phrase modifies the initial noun-phrase; it is not obligatorily differentiated from the initial noun-phrase, resulting in shorter pauses, and a smaller pitch difference. Due to the third phrase of passives being a separate noun-phrase, it should suggest that it should form a new prosodic grouping. Thus, the incremental nature of prosody can provide an alternative explanation for our pattern of results.

Studies of speech production have suggested that temporal and tonal boundaries in speech reflect the hierarchical syntax of utterances. Based on a series of experiments, Cooper, Paccia, and Lapointe (1978) suggest that the extent to which speakers lengthen the pre-boundary final syllable, and the duration of the following pause,  increase with greater depth in the syntactic structure; when a boundary occurs at a deeper level of the syntactic structure (e.g. at the end of an embedding), the lengthening effect will be greater. Similarly, Cooper and Sorensen (1977) argue that pitch boundary cues relate to syntactic structure; a syntactic boundary between two conjoined main-phrases produced a larger reduction in pitch in the phrase final syllable, followed by a greater rise in pitch by the first stressed syllable after the boundary. In sentences containing embeddings, there was a similar pitch reduction during the final pre-boundary syllable, followed by a smaller pitch increase by the first stressed syllable following the boundary. Taken together, these studies suggest that the prosodic cues present in speech are automatically mapped onto the speaker's cognitive representation of the utterance's hierarchical syntactic structure. In the present study, we assessed whether lengthened pauses and pitch declination at the boundary of an embedded clause – reflecting hierarchical structure – are reliable indicators of dependency boundaries, and whether these cues may be sufficient to trigger the auditory Gestalt principles of pitch similarity and temporal proximity. For active-object relative clauses, this was the case, suggesting that cues generated during production could be used to process hierarchical structures in comprehension non-hierarchically (Frank, Bod, & Christiansen, 2012), by the supporting grouping of the embedded clause. While we have found evidence suggesting these cues are present, additional comprehension studies will be required to determine whether they are useful for listeners.

The current study forwards the idea that useful prosodic cues are generated during production according to syntax, and are not driven by audience design, given the lack of an interacting partner here, and in other production studies (e.g. Cooper, Paccia, Lapointe, 1978; Cooper & Sorensen, 1977; Kraljic & Brennan, 2005). However, it is necessary to note that syntactic factors are unlikely to be the sole factors affecting prosodic structure. Ferreira (1993) has argued that semantics can mediate the relationship between a syntactic representation and its articulation. Kraljic and Brennan (2005) additionally posit that semantic or pragmatic information could influence prosodic lengthening if that information were available before articulation, but that this information is unlikely to be available if it requires more time and computation than the system ordinarily expends during conversation. This raises a way in which the current study, amongst many, may give a very narrow view of the relationship between syntax and prosody.

To specify a robust account of prosody, it may be necessary to give due consideration to contextual factors that may influence prosodic cues such as lengthening in everyday interaction. Under simulation accounts, such Pickering and Garrod's (2004) Interactive Alignment Theory, interlocuters converge across all levels of linguistic communication, from semantics, to syntax, phonetics, and even gesture in order to predict upcoming speech, reducing the complexity of online speech processing. Restricting the search space to phonetics - in large part due to Ferreira's (2002) assertion that the phonological properties of words influence the degree of lengthening or unfilled pauses -, there is substantial evidence suggesting that interlocuters automatically imitate several aspects of one another's speech, including accent, speech rate, intonation and speech style (Delvaux & Soquet, 2007; Webb, 1969; Goldinger, 1998; Shockley et al., 2004; Pardo et al., 2010). Several of these may also correlate with several aspects of audience design; if your interlocuter does not appear to understand you, lowers their speech rate, speaks more effortfully, and pauses more often, you will likely automatically imitate them, it need not be an explicit production decision. Alignment through imitation of these factors will likely affect many studies on temporal prosodic cues, and presumably, pitch cues as well. Given these observations, it is surprising that work considering prosody, including the study presented here, only consider prosody in tightly constrained situations, the laboratory vacuum, so to speak. If we are to truly generate a fully mechanistic account of prosody and its utility to listeners, it will be necessary to conduct work assessing to what degree these prior findings may be explained by factors such as pragmatics, semantics, and communicative context.

In the introduction we explored the nature of auditory-perceptual Gestalts in relation to musical processing, so here we will briefly discuss the relationship between the two. This discussion was couched mainly in terms of the pitch similarity and temporal proximity Gestalt principles, and the interaction between the two. Quickly, it was demonstrated in Deliège (1987), Dowling (1973), and Hamaoui and Deutsch (2010) that whilst pitch similarity is a strong grouping cue (even more so when it is hierarchically structured), temporal grouping cues are able to over-power them, overall providing a stronger grouping cue at durations greater than 30ms. In the current data, pitch similarity appeared to be a more powerful and reliable cue. Overall, non-zero pauses, i.e. the placeholder pauses used to assess pause likelihood were more frequent. Further, the model assessing pause likelihood did not produce any statistically significant effects. However, when they did occur, they differed on the basis of syntactic form and location within the utterances that agreed with the pitch groupings. What then, does this mean for the relationship between linguistic and musical processing?

The first implication is that similar to music processing, pitch similarity in spoken language can be an effective grouping cue (see Ferreira, 2002; Kraljic & Brennan, 2005). If we consider words as notes, words that are more similar in pitch tend to belong to the same phrase. However, pause cues seemed to have played a weaker role. Pauses may provide information regarding phrasal and clausal membership in English where present, however, they do not always occur. Recall the argument made by Männel and Friederici (2009); pause cues may be necessary to elicit ERPs corresponding to the closure of a prosodic grouping in German acquiring infants compared to their English counterparts. This is due to the fact that English has an inflexible word order, requiring a larger inventory of intonational cues to perform functions word order may perform in German. Thus, the current results may speak more to cross-linguistic variation than to the relationship between auditory language and music processing. Pauses, when they occurred, did so in locations that were congruent with large pitch changes, suggesting they would be help to reinforce the pitch groupings. It may simply be the case that - due to the native English-speaking sample - speakers were more expert with, and thus more likely to use pitch cues to communicate structural relationships with prosody. This argument would, however, require testing using a similar elicitation paradigm with German native speakers. Thus, we suggest that pauses do provide useful cues to grouping but are largely optional in English.

An additional question raised by the lack of reliability of pause cues is whether, in isolation, they may be an insufficient phrasal grouping cue. In both language (e.g. Ferreira, 2002; Snedeker & Trueswell, 2003) and music (Lerdahl and Jackendoff, 1983), phrase-final lengthening is an important grouping cue. Given that we focused solely on unfilled pauses, it may well be the case that they are an infrequent grouping cue in isolation, simply because there are other alternatives. Given Ferreira's (2002) observation that whether structural boundaries are indicated by lengthening a word, or by pausing following it depends on the phonetic characteristics of the word, this seems likely. However, in the present work, unfilled pauses were the best candidate to study, as they have proven potent in music, and they serve as an unambiguous marker of temporal proximity.

To conclude, we sought to answer the question of whether speakers' use of pitch and temporal dynamics would co-vary with structural choices made during sentence production, specifically in active-object and passive relative clauses. To pursue this question, we assessed whether there were cues that would reflect the operation of two auditory-perceptual Gestalts, temporal proximity and pitch similarity. Indeed, the results indicated that speakers reliably used more similar pitch for words occurring in phrases sharing sequential structural dependencies, potentially obviating the need to perform hierarchical processing during speech comprehension. Pause cues were, however, less reliable, potentially indicating they are not an obligatory component of English prosody. More generally, these findings demonstrate a set of cues that may be useful for obviating the syntactic structure during the comprehension of complex spoken English sentences, however this remains to be assessed in future empirical investigation.

# Chapter 4: Multiple natural language cues assist the processing of hierarchical structure

Antony S. Trotter[1], Rebecca L.A. Frost[2], Padraic Monaghan[1,2,3]

1. Department of Psychology, Lancaster University
2. Max Planck Institute for Psycholinguistics
3. University of Amsterdam

The empirical work outlined in Chapter 4 was conducted to see whether pitch similarity and temporal proximity cues taken from natural language would facilitate the acquisition of hierarchical structure. In this study, we emphasized temporal proximity cues, and did not model the cues off the participants' native language, letting us assess whether Gestalt grouping cues (regardless of their familiarity or source) support acquisition. By comparing these results with those of Chapter 5, it provides insights into how general the influence of Gestalt cues are. This paper is currently in a draft ready for submission.

## Statement of Author Contribution

In the Chapter entitled, “Multiple natural language cues assist the processing of hierarchical structure”, the authors agree to the following contributions:

Antony S. Trotter – 70% (Writing, experimental design, data collection and analysis)

Signed: ______________________________________ Date: ______________

Dr, Rebecca L. A. Frost – 15% (Experimental design, and review)

Signed: ______________________________________ Date: 22.2.2019

Professor Padraic Monaghan – 15% (Experimental design, and review)

Signed: ______________________________________ Date: 22.2.2019

# Abstract

Recursion is considered a crucial property of human language (Hauser, Chomsky, & Fitch, 2002), and is a component of phrase structure grammars (Chomsky, 1957). Hierarchical centre-embeddings (HCEs) have therefore been taken as evidence that language is not a finite state system (Chomsky, 1957). While phrase structure may be necessary for their production, sequential processing may underlie their comprehension (Frank, Bod, & Christiansen, 2012). Under this account, listeners use surface level cues (e.g. semantic content, pitch and temporal variation) to determine the dependencies within an utterance. Here, we assessed whether including pitch cues consistent with speech (Fery & Schubö, 2010) and temporal grouping cues consistent with prior artificial language work (Hawthorne & Gerken, 2014) would facilitate the acquisition of an artificial grammar.

80 native English speakers were trained on an $A_nB_n$ grammar containing one and two levels of embedding (LoE) sequences. Participants were assigned to one of five cue conditions; baseline (distributional cues), temporal proximity (175ms pauses occurred between syntactically unrelated syllables), pitch similarity (dependent syllables occurred in the same pitch), semantic similarity (marked with phonological cues), and combined (semantic similarity + pitch + temporal). At test, participants performed a grammaticality judgement task on novel structures.

The results suggested that the additional cues did not enhance learning. Pitch (and to a lesser extent temporal) grouping cues produced higher judgement accuracy only for grammatical sequences, suggesting that listeners found them salient. However, accuracy did not increase for ungrammatical sequences, suggesting that salient, global acoustic cues masked local linguistic violations.

# 1. Introduction

Recursion is commonly regarded as a crucial property of human language (Hauser, Chomsky, & Fitch, 2002). Natural language contains several forms of recursion (de Vries, Christiansen, & Petersson, 2011). In particular, recursive centre-embedding has played a critical role in debates about the nature of human language processing: The presence of hierarchical centre-embeddings in natural language has been taken as evidence that language is not a finite state system (Chomsky, 1957; 1959). *The rat the cat the dog bit chased ate the malt* is a typical example of a centre-embedded structure with two hierarchically embedded sub-clauses, or two *levels of embedding* (LoE). As more clauses are inserted, the distance between dependent items grows, thereby increasing the difficulty for learning or remembering associations between related constituents (Lai & Poletiek, 2011). Sentences with over three LoEs have been shown to be incredibly difficult to process, even for highly proficient speakers (e.g. Bach, Brown, & Marslen-Wilson, 1986; Newmeyer, 1988). Probing the processing of hierarchically centre-embedded structures thus offers insights into the nature and complexity of human language processing.

In psycholinguistic research, humans' capacity to process linguistic structures is typically investigated using the artificial grammar learning (AGL) paradigm. AGL studies typically contain two phases; training and testing. In the training phase, participants are presented with sequences (letters or nonsense words) that are – unbeknown to participants – grammatical sequences generated by an experimental grammar. At test, participants are presented with novel sequences that conform to the rules of the grammar, or violate them. Their task is to detect which sequences adhere to the grammar, and which sequences do not. Successful learning is typically defined as an ability to classify sequences with accuracy at a level greater than chance. Artificial grammars typically conform to one of two theoretically motived rules; $A_nB_n$ or the $AB_n$.. For both types of sequence, structures are drawn from two word categories: A and B. In a $A_nB_n$ sequence, the grammar produces a sequence of As succeeded by a matching number of Bs. Under this rule, a given pair of words – e.g. $A_1B_1$ (*Be Po*) – can be inserted into another clause – $A_2B_2$ (*Da Ti*) – to make a longer sequence, $A_2A_1B_1B_2$ (*Be Da Ti Po*). Under the $AB_n$ rule, a pair of words - $A_1B_1$ (*Be Po*) - can be added to the end of another pair – $A_2B_2$ (*Da Ti*) – to make a longer sequence, $A_2B_2A_1B_1$, (*Be Po Da Ti*).

The $A_nB_n$ and $AB_n$ rules correspond to the phrase structure grammar (PSG) and finite-state grammar (FSG) levels of the Chomsky hierarchy (Chomsky, 1957; 1959). The hierarchy mathematically arranges rule systems capable of generating an infinite set of sequences by their increasing, generative power. FSGs are the weakest level of the hierarchy, which can be fully specified by transitional probabilities between a finite number of states (Hauser & Fitch, 2004). Thus, processing these sequences requires only a large enough memory to hold sequential states, and transitions between them. FSGs relate to the $AB_n$ rule as this rule type requires learners to only remember specific A-B relationships and concatenate them into longer sequences. The $A_nB_n$ rule generates PSG sequences. PSGs lie at the next level of the hierarchy, and much like FSGs, can concatenate items. Crucially, through the recursive application of the *merge* operation (Chomsky, 1995), PSGs can embed strings within other strings, resulting in phrase structures and long-distance dependencies (Berwick & Chomsky, 2016) For example, in English, complex constructions can be generated by inserting multiple relative clauses into a main clause (e.g. "*Keith, who dates Mary, who is an engineer at the factory, that is owned by the government, said he would be attending the dance alone*"). The processing mechanisms required for these complex structures are more sophisticated, requiring both an open-ended memory system, and perceptual mechanisms to recognise them (Hauser & Fitch, 2004). It is commonly suggested that PSGs, and more recently *merge* (Yang, Crain, Berwick, Chomsky, & Bolhuis, 2017) are the defining characteristic of human languages (Chomsky, 1959; Haegeman, 1991). The AGL literature has often employed both types of rule to assess what cues learners use to acquire a PSG. However, $A_nB_n$ languages are notoriously difficult to acquire, thus measuring their processing is difficult to accomplish. Several studies have, however, shed light on the issue.

In natural language, to fully parse a hierarchical centre-embedded structure (HCE), it is necessary to understand the dependencies between particular nouns and verbs. By extension, for AGL studies to be informative about language processing, it is necessary for participants to acquire the associative dependencies between particular As and Bs. To determine participants' capacity to learn such dependencies, Friederici, Bahlmann, Heim, Schubotz, and Anwander (2006) designed $A_nB_n$ and $AB_n$ sequences where particular A and B words always co-occurred. For example, whenever a given A word, e.g. "de" occurred, the B word "fo" always. Participants were assigned to $A_nB_n$ or $AB_n$ groups, and were trained on their respective grammars over 12 training blocks, each of which, presented participants with 10 grammatical sequences. Participants then heard 10 novel sequences, half of which adhered to the trained grammar, whereas the other half did not, and classified each new sequence as either grammatical or ungrammatical. Participants were given corrective feedback after each response. Learning was assessed in a follow-up session in which participants classified 160 novel sequences. Participants were able to learn both types of grammar, as evidenced by above chance classification accuracy for both groups.

However, de Vries et al. (2008) noted that as these pairings share phonological properties, a simple counting strategy could be employed to detect non-grammatical items, (e.g. if a violation sequence was $A_1A_2A_3B_3B_2A_4$, counting the number of syllables ending in "e" or "i" reveals violations). Indeed, when stimuli were constructed to prevent such a counting strategy, de Vries et al. (2008) demonstrated that learning of HCEs was no longer possible. Consequently, the relationship between Friederici et al.'s results and natural language processing is unclear. Subsequent research has suggested that HCE structures can be acquired in the laboratory under specific circumstances. Lai and Poletiek (2011; 2013) found that participants can learn specific A-B pairs when given a "starting small" training regime, whereby participants are initially trained on individual A-B pairs before receiving more complex HCE structures.

Thus, HCE structures in natural language have been shown to be difficult to interpret once a certain level of complexity has been achieved (greater than three LoEs). Further, acquiring HCEs has been found to be difficult to accomplish. An important question, then, is what helps the everyday parser to interpret these structures?

It is important to note that whilst a sentence may possess a hierarchical structure, it is not necessarily the case that this sentence will be processed hierarchically. Frank and Bod (2011) compared how well word-probability estimates generated by three kinds of probabilistic models (each incorporating different psychological mechanisms and representations) accounted for ten participants' reading-time measurements of the Dundee corpus. The first class of model was a phrase-structure grammar model, induced from large datasets of syntactic trees, and utilising hierarchical structure. The second (Markov models) and third (Echo State Networks) classes only had access to sequential structure. The phrase structure grammar models failed to estimate variance in reading time data over and above the sequential-structure models; a sentence's hierarchical structure, unlike other sources of information, did not noticeably affect the generation of expectations about upcoming words. This suggests that during comprehension, individuals may draw on non-hierarchical mechanisms to process hierarchical structure.

Frank, Bod, and Christiansen (2012) propose that speech comprehension may be driven by sequential processing. According to theory on sequential processing, to comprehend a HCE structure in a rapidly unfolding temporal context, the listener would need to rely on superficial, surface level cues to parse its dependencies, rather than processing the incoming speech as a hierarchy. For example, in the sentence, “The cat the man strokes purrs”, world knowledge can be used to parse its dependencies; cats purr, but humans do not, however, humans do stroke cats. World knowledge thus gives the dependency relationships, “man strokes x”, “cat purrs”, according to which, the meaning can be inferred. In this paper, we focus on three speech cues that may support surface level processing: Pitch, pause, and semantic similarity information (marked using phonology), which could help listeners group words in speech by providing information about words’ clausal membership.

To assess which linguistic features facilitate sequential or hierarchical processing, it is necessary to implement these cues in an experimental setting. For most AGL experiments, to create tightly controlled stimuli, the prosodic characteristics are removed from the stimuli; non-words are delivered at a constant rate, and if presented auditorily, at an even pitch and amplitude, and that dependent pairs are reinforced throughout training. In contrast, natural language is rich with cues that potentially support the processing of these structures.

Take pitch and speech rhythm, for example. Mueller, Bahlmann, and Friederici (2010) manipulated the presence of these cues in different combinations in structures with on LoE to see how they affected learning. For pitch information, each artificial HCE string had descending sentential prosody, with a pitch declination over the course of the string. Speech timing cues were present in two conditions; in the first condition, pauses were added between entire sequences, temporally bracketing grammatical sequences, whereas in the second condition, pauses also occurred between corresponding pairs, temporally bracketing strings and dependent non-words. At test, participants were successively presented with two strings, and were required to select which of the strings conformed to the rules of the grammar. The benefit of each cue was additive; participants trained with all cues were most accurate at test, selecting a higher number of grammatical strings. Thus, the presence of speech-like qualities in artificial material appears to help participants process and learn the grammatical structure of sequences with one LoE. It remains to be seen if speech-like rhythmic and pitch information will benefit processing for longer, more complex structures. Further, it is not clear if the effect of these cues can be further enhanced by adjusting their acoustic properties to more closely match natural language.

German and Dutch offer a unique environment for assessing what aspects of speech may support processing of centre-embedded structures. In English speakers, the missing verb effect is common when processing HCEs (Gibson & Thomas, 1999). The missing verb effect refers to when centre-embeddings lacking a verb-phrase are viewed as grammatical (e.g. "The patient who the nurse who the clinic hired met jack"). However, this effect is not present in speakers of German (Vasishth et al., 2010) and Dutch (Frank et al., 2015), who find the grammatical versions of these sentences easier to process. Verb-final constructions are common in German and Dutch and require the listener to track dependency relations over long distances, suggesting that experience results in language-specific processing improvements (Christiansen & Chater 2015). As a result, it is reasonable to assume that speakers of these languages may reliably employ cues that support their disambiguation.

Fery and Schubö (2010) conducted a phonetic study of pitch variation in centre-embedded relative clause production in German. In this study, participants were asked to produce HCE relative clause structures ("[(1) $_{C-1}$ *The pears* [(2) $_{C-2}$ *which at the tree* [(3) $_{C-3}$ *which green is*] (4) *hang*] (5) *are sour*]") and their non-embedded counterparts ("*The pears are sour*"), to provide a basis of comparison. The findings indicated that participants used a progressively lower pitch for each LoE, and that each constituent phrase of an LoE use a similar pitch. This results in the pitch increase seen between phrases three and four, in which pitch returns to a level similar to the noun phrase of the embedding (phrase (2) in the example). These results are summarized in Figure 1. This appears to be a departure from the cues used in Mueller et al.(2010). However, in this study, only one LoE structures were used; with the introduction of a second LoE, a pitch rise at the closure of the deepest LoE would have been appropriate. Otherwise, the pitch reductions seen in Fery and Schubö (2010) respect the trend for pitch to reduce over the course of a sentence. These results, critically, suggest that embedded clauses can bear perceptual acoustic grouping cues, assisting listeners with the interpretation of HCEs, in addition to cues signalling the onset and offset of an utterance. Given that verb-final constructions such as centre-embeddings are common in German, and require listeners to track dependency relations over along distances (Christiansen & Chater, 2015), we reasoned that the pitch prosodic cues found in this study would be good candidates for supporting the acquisition of centre-embedded structures for our native-English speaking participants, who have less experience with this syntactic structure.

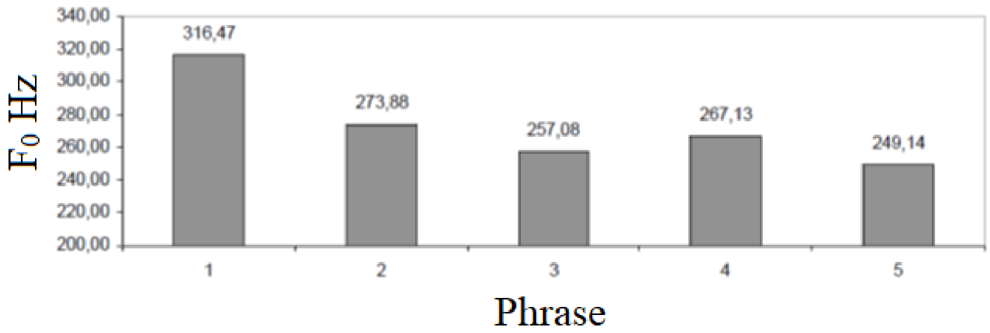


Figure 1. Peak $F_0$Hz on the first word of each phrase for HCE structures (adapted from Fery & Schubö, 2010).

Further grouping cues are also provided by rhythmic information. Hawthorne and Gerken (2014) assessed whether prosodic cues can help guide infants' learning of constituents. Nineteen-month-olds were trained on 1 (ABCDEF) or 2 clause (ABC, DEF) prosody of non-word sequences, and were subsequently tested using a modified head-turn procedure on novel grammatical (DEF, ABC) or ungrammatical (EFA, BCD) movement of the clauses from the 2-clause familiarisation phase. Pauses of 173 ms and pre-final lengthening were used to separate the two-clauses, conferring to Nespor and Vogels' (1986) prosodic hierarchy. The group trained in the 2-clause condition were able to discriminate between grammatical and ungrammatical items, even though the test sentences used a new pitch contour, indicating that the infants perceived prosodically grouped words as more constituent-like than words that straddled a prosodic boundary. In other words, the infants were able to use the rhythmic cue to derive clause membership, i.e. to group the non-words into sub-sequences. Pauses of different durations have also been found to facilitate learning of AGL sequences; for instance, short pauses (25ms) have been shown to help learning for non-adjacent dependencies (Penã et al., 2002). This pause serves as a perceptual cue to word boundaries, allowing more facility for structural processing of unfamiliar materials (de Diego Balaguer, Martinez-Alvarez, & Pons, 2016).

Segmental phonological cues have been employed in artificial grammar research to support the detection of dependencies. Friederici et al. (2006) and Bahlmann, Schubotz, and Friederici (2008) marked grammatical category membership by employing phonological cues. A and B syllables always included "e"/"i" and "o"/"u" respectively. Whilst improving accuracy on a two-alternate forced choice task, this manipulation suppressed learning of A-B dependencies by permitting the use of counting strategies. Penã et al. (2002) demonstrated learning of non-adjacent AxC structured words (where the dependency between $A_n$ and $C_n$ can separated by any x- syllable) when the syllables utilised an plosive-continuant-plosive phonological structure (e.g. “be-ra-ga”, “pu-li-ki”). In contrast, Onnis et al. (2005) found that when AxC stimuli contained a continuant-plosive-continuant (e.g. “ze-ta-vo”, “thu-gi-shu”), participants did not demonstrate learning. These studies illustrate that phonological cues to dependency structure can be used for grammatical acquisition.

In natural language, phonological similarities between dependent syllables are infrequent, however, segmental phonological cues have been shown to be useful cues for grammatical category membership. Monaghan et al. (2005) assessed 16 phonological cues the most frequent 2751 nouns and 1139 verbs in the CHILDES corpus and found several predictors that provide cues to category membership. At the word level, nouns had more syllables than verbs. Syllables in verbs had greater onset complexity and syllabic complexity than nouns, whilst nouns had more reduced syllables. Verbs ended in -ed more often than nouns. At the phoneme level, nouns had more coronal consonants than verbs, but fewer nasal consonants. Vowels in nouns were further back and higher than vowels in verbs. Phonological cues can distinguish grammatical categories, and these cues may thus support the acquisition of phrase structure grammars. In the present study, we used a segmental phonological cue – phonological similarity between dependent syllables - as an additional low-level cue to help support acquisition of non-adjacencies.

Two auditory perceptual Gestalts may be particularly relevant for the processing of HCEs; pitch similarity, and temporal proximity (Deutsch, 2013). Pitch similarity states that individuals tend to form sequential links between tones that are close in pitch, and to distinguish between those that are further apart. This suggests that the similar pitch for clauses found in Fery and Schubö (2010) would bias participants to correctly group phrasal elements. Temporal proximity states that if two tones are temporally distant, you are unlikely to create a sequential link between them. Trotter, Frost, and Monaghan (see Chapter 3) acoustically analysed a corpus of spontaneously produced active-object ("[The bear] [the girl] [is hugging] [is brown]") and passive relative clauses ("[The bear] [being hugged] [by the girl] [is brown]"). We hypothesised that pitch similarity would be highest between syntactically dependent phrases, and that pauses occurring between clauses would be longer than elsewhere in the speech. In passive relatives, pitch similarity was highest for the first two phrases ("The bear being hugged"), and the longest pause occurred between the second and third phrases, before the by-phrase ("by the girl"). This ran contrary to prediction and would not support grouping of the embedded clause. For active-object relatives, phrases in the embedded clause ("the girl is hugging") were produced using a more similar pitch and contrasted with the first phrase of the external clause, in line with our hypothesis. Pauses were also tended to be longer preceding the embedded clause. For active-object relative clauses, pitch similarity and temporal proximity therefore provide grouping information consistent with syntactic structure. Prosodic cues may facilitate the non-hierarchical processing of hierarchical structure, supporting the acquisition of phrase structure grammar.

The above studies provide the framework for implementing natural language cues in AGL research with adults. By integrating multiple sources of information, it should be possible to attain a greater understanding of their relative importance.

### *1.1 The present study*

To test the influence of natural language cues on the structural processing of HCEs, we sought to implement pitch, rhythmic, and phonological similarity cues to signal dependencies in an AGL setting. To assess their effect on learning, a baseline condition was conducted, where none of the cues were present, as well as a combined condition, which utilised all three cues. Comparing these conditions to the baseline will permit assessment of which learning conditions best support learning of the grammar, and whether a combination of these cues provides additional gains. We hypothesised that relative to baseline, each individual cue will facilitate learning, and that further, the combined cues will result in the greatest learning (Mueller et al., 2010). Additionally, we hypothesised that participants will judge novel structures correctly after increased training, and that overall, they will be less accurate with longer sequence lengths (Lai & Poletiek, 2011), due to their increased complexity.

## 2.0 Method

### *2.1 Participants*

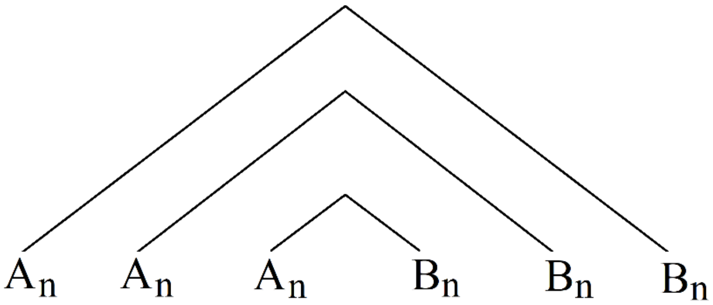
80 native English speakers ($\text{Mean}_{\text{age}}$ = 20.190, $\text{SD}_{\text{age}}$ = 3.068, $\text{n}_{\text{female}}$ = 63) participated in the study, all of whom were students at Lancaster University. Participants were randomly assigned to one of five conditions (n = 16 per condition) and received £3.50 or course credit for their participation.

### *2.2 Materials and Design*

Finite state grammar sequences were constructed following $A_nB_n$ rules (see Figure 2), producing sequences conforming to a hierarchical centre-embedded structures. Therefore, each sequence contains two word categories, A and B. Each word category contained six consonant-vowel syllables, resulting in twelve syllables per grammar. Words in both categories were monosyllabic, and were comprised of a plosive consonant (“P”, “B”, “G”, “D”, “T”, “K”) and a vowel, or vowel pairing (“a”, “e”, “i”, “o”, “u”, “oi”). The set of syllables was generated by randomly pairing a plosive with a vowel. We generated two separate languages to assess whether participants’ learning was driven by phonological factors external to the manipulations. In each version of the language, individual consonants and vowels occurred once per category, with no repetitions of consonant-vowel pairings. Therefore, this resulted in a total set of 24 syllables. *Language 1* was comprised of the following syllables: *A*: “Pe”, “Bu”, “Gi”, “Doi”, “To”, “Ka”; *B*: : “Ku”, “Ta”, “Po”, “Bi”, “De”, “Goi”. Language 2 was comprised of the following syllables: *A*: “Gu”, “Di”, “Te”, “Bo”, “Koi”, “Pa”; *B*: “Ti”, “Ge”, “Ko”, “Poi”, “Ba”, “Du”. These baseline languages were employed in the baseline, pause and pitch conditions.

Each syllable was created using the Festival speech synthesiser (Black, Taylor, & Caley, 1990). In each case, syllables were generated using the default voice, at the default rate. In addition to these default parameters, we specified the target pitch level (150Hz, 135Hz, 120Hz) using the “Default intonation”, which allows the researcher to specify the pitch at the beginning and end of the utterance. In each case, we specified both at the target pitch level. Each monosyllable lasted between 133 and 182ms (mean = 157ms, SD = 13ms). This variance resulted from differences in vowel (e.g. “e” had a shorter duration than “oi”) and consonant durations (e.g. “p” has an unvoiced onset, whereas “g” does not) that were implemented in the default voicing parameters employed by Festival.

Hierarchical centre-embedded structures were generated using the $A_nB_n$ rules. Each $A_i$ syllable was paired with a $B_i$ syllable (e.g $A_1B_1$), resulting in six grammatical pairings per language, indicated using numbered indices. To generate grammatical sequences, any $A_iB_i$ ($A_1B_1$) pairing could be inserted within any other $A_iB_i$ pairing ($A_6B_6$) to a minimum of one level of embedding (LoE; $A_6A_1B_1B_6$, hitherto LoE 1) and a maximum of 2 LoE ($A_6A_1A_3B_3B_1B_6$, hitherto LoE 2). Sequences violating the experimental grammar were generated for the test phase, in which one B syllable failed to match all A syllables in the sequence. For ungrammatical sequences, two additional constraints were used; the same B syllable could not occur more than once in the sequence, and no adjacent $A_iB_i$ violations could occur. Therefore, in LoE 1 sequences, violations always occurred in the final position ($A_6A_1B_1B_4$), and in LoE 2 sequences, violations occurred in either the fifth ($A_6A_1A_3B_3B_2B_6$) or sixth sequence positions ($A_6A_1A_3B_3B_1B_5$). Figure 2 illustrates the grammatical structure in detail, and provides examples of grammatical and ungrammatical sequences



Baseline:

| | | |
|---|---|---|
| LoE 1 G: | $A_1A_2B_2B_1$ | Pe Bu Ta Ku |
| LoE 1 U (4th Position): | $A_3A_4B_4\underline{B_1}$ | Gi Doi Bi $\underline{\text{Ku}}$ |
| LoE 2 G: | $A_4A_5A_6B_6B_5B_4$ | Doi To Ka Goi De Bi |
| LoE 2 U (5th Position): | $A_3A_2A_1B_1\underline{B_4}B_3$ | Gi Bu Pe Ku $\underline{\text{Bi}}$ Po |
| LoE 2 U (6th Position): | $A_6A_3A_2B_2B_3\underline{B_5}$ | Ka Gi Bu Ta Po $\underline{\text{De}}$ |

Phonological similarity:

| | | |
|---|---|---|
| LoE 1 G: | $A_1A_2B_2B_1$ | Pe Bu Bi Po |
| LoE 1 U (4th Position): | $A_3A_4B_4\underline{B_1}$ | Gi Doi De $\underline{\text{Po}}$ |
| LoE 2 G: | $A_4A_5A_6B_6B_5B_4$ | Doi To Ka Ku Ta De |
| LoE 2 U (5th Position): | $A_3A_2A_1B_1\underline{B_4}B_3$ | Gi Bu Pe Po $\underline{\text{De}}$ Goi |
| LoE 2 U (6th Position): | $A_6A_3A_2B_2B_3\underline{B_5}$ | Ka Gi Bu Bi Goi $\underline{\text{Ta}}$ |

Figure 2: Structure of the experimental grammars. General structure and examples of stimuli in the $A_nB_n$ PSGs. Examples of the correct and violation sequences are given for LoE 1 and LoE 2 conditions for language 1 for the baseline (baseline, pause and pitch conditions) and phonologically similar languages (phonological similarity and combined cues conditions). *G* indicates a grammatically correct sequence, and *U* indicates a sequence that violates the rules of the grammar, with the violation position stated in brackets, and underlined in the examples.

The experiment utilised a between subject design, with participants randomly assigned to one of 5 different cue conditions; *baseline* (no cues), *pause*, *pitch*, *phonological similarity*, and *combined* (pause + pitch + phonology). In each condition, cues were present over both training and testing.

In the baseline condition, the only cues that could guide learning were the frequencies with which dependent syllables in the sequences co-occurred. In the baseline, phonological similarity, and pitch conditions, 25ms inter-syllable pauses were employed, in accordance with Pena and colleagues (2002).

The pause condition employed temporal grouping cues that could highlight the dependency structure of the language; 175ms pauses occurred between levels of embedding (e.g. $A_1$ [pause] $A_2B_2$ [pause] $B_1$), in line with Hawthorne and Gerken (2014).

For the pitch condition, the initial and final syllables always used the highest and lowest pitch (150Hz, 120Hz), respecting sentence level prosody (e.g. Fery & Schubö, 2010; Mueller et al., 2010). Syllables within a LoE used the same pitch, with 15 Hz difference between levels. This pitch difference was obtained by taking the median of the pitch changes between levels of embedding presented in Figure 2 (Fery & Schubö, 2010). To verify whether the pitch manipulation was detectable by our participants, each participant in the pitch and combined cues condition (n = 32) was administered an informal pitch sensitivity test, wherein they were played two examples of a structure from each LoE. One of these sequences was canonical with the experimental pitch structure (150Hz, 135Hz, 135Hz, 120Hz) and one which was randomised (e.g. 120Hz, 150Hz, 135Hz, 120Hz), and asked whether they "sounded the same" or were different. Both sequences used the same syllables. 26 participants (81%) were able to detect the difference.

In the phonological similarities condition, syllable $A_i$ and $B_i$ would always share the same initial phoneme (e.g. *pa bu bi po*, see Figure 2 for more detail).

Finally, the combined cues condition employed all cue types.

*2.3 Procedure*

Participants were randomly assigned to one of the five experimental conditions and told that they would be presented with linguistic items that followed a sequential rule. At the start of each block, a message appeared at the centre of the screen that instructed participants to press any key to begin the familiarization phase. In each familiarisation phase, participants passively listened to 16 strings that adhered to the grammar of the language. Eight of these strings contained one level of embedding, and the remaining eight contained two levels of embedding. Stimuli were presented in a randomised order. Syllables were presented sequentially, in isolation. Whole strings were separated by 3000ms pauses.

Following familiarization, participants were presented with text informing them that the test block would begin after they pressed any key. In the testing phase, participants were presented with 16 novel sequences. After each sequence, participants performed a forced grammatical classification task; participants were required to indicate – via keyboard response – whether the sequence adhered to the underlying linguistic rules, pressing "Y" for yes, or "N" for no, after which they received corrective feedback. There were 16 trails in each testing block, with eight LoE 1 sequences, and eight LoE 2 sequences. Four of each LoE sequences were ungrammatical, and the remainder were grammatical (see Figure 3 for examples of grammatical and ungrammatical test stimuli). Test items included the same cues as training items.

In total, participants completed 12 blocks of familiarization and testing. The experiment lasted for approximately 40 minutes.

## 3. Results

### *3.1 Grammaticality Judgement Accuracy*

Overall, grammaticality judgement accuracy was similar between conditions. Exposure to baseline cues judged novel sequences with the greatest accuracy (Mean = 0.531, S.E.M. = 0.009), followed by combined cues (Mean = 0.527, S.E.M. = 0.009), phonological similarity (Mean = 0.525, S.E.M. = 0.009), pause (Mean = 0.516, S.E.M. = 0.009), and finally, pitch cues (Mean = 0.508, S.E.M. = 0.009). These descriptive statistics suggest that additional cues did not influence learning, due to performance being similar to chance (50%). Figure 3 summarises the mean per block accuracy, broken down by LoE and condition. Visual inspection of the figure suggests classification was similar between LoEs for baseline, phonological similarity and pause cues. In contrast, for pitch cues, and in the last five blocks for the combined cues condition, accuracy appears higher for LoE 1 sequences.

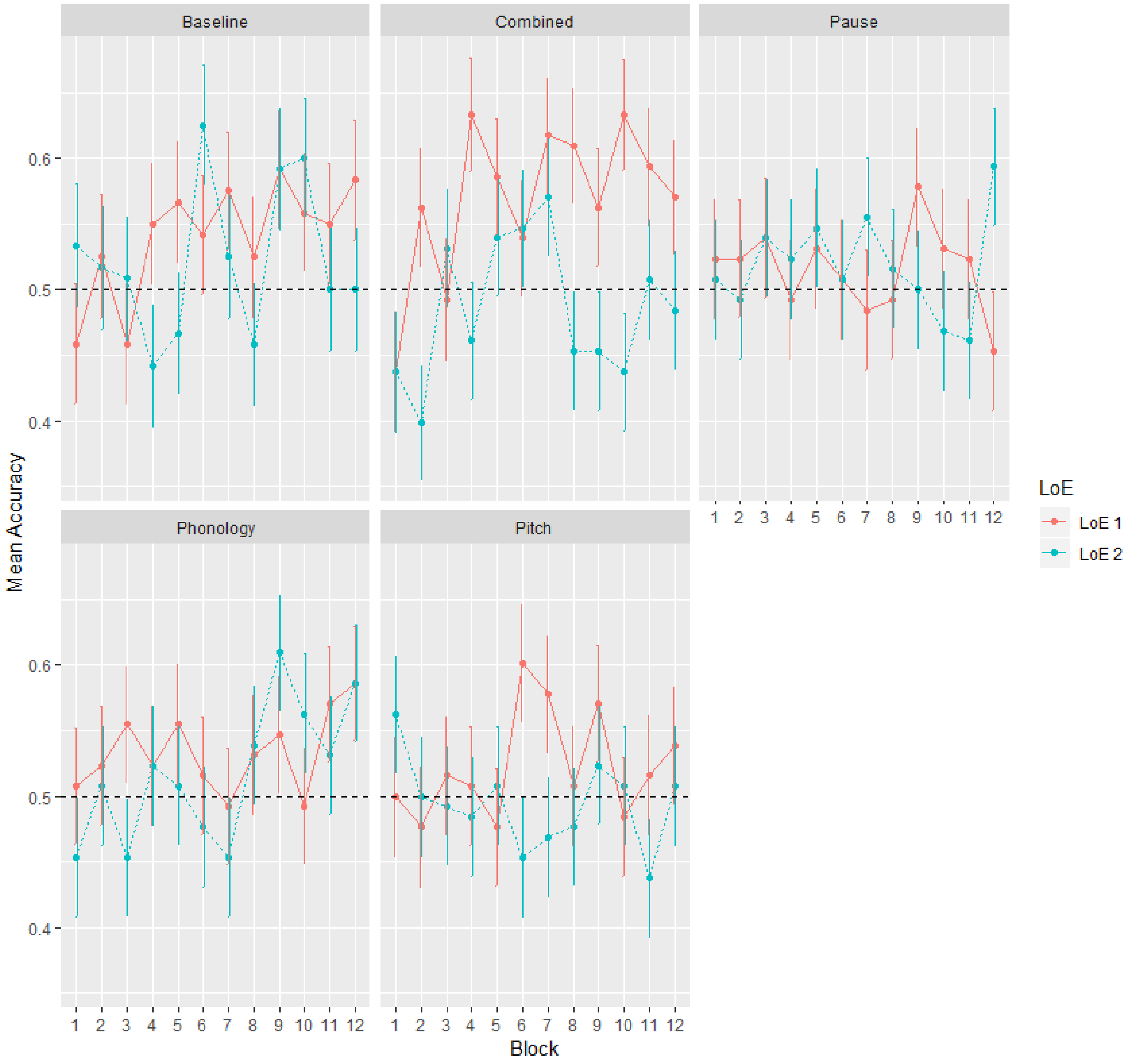


**Fig. 3**. Displays the mean classification accuracy per condition, block, and LoE. Red, solid lines and points display the means for LoE 1 sequences, blue, dashed lines and points display the means for LoE 2 sequences. Vertical coloured lines display the standard error of the mean (S.E.M.). The horizontal dashed line indicates chance performance.

To formally assess the data, we conducted a series of generalized linear mixed-effects models (GLMER) predicting the dependent variable of accuracy (correct or incorrect) with a logit-link function. Random intercepts and slopes for participants and items were included in all reported analyses. Models were built up iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). Fixed effects were retained in the final model if they resulted in significantly improved model fit in isolation, or within an interaction. Interactions were retained in the model if they significantly improved model fit. As fixed effects, we tested the effect of the cues participants were exposed to (baseline, phonological, pitch, pause, and combined), LoE (LoE 1 or LoE 2), violation type (grammatical, 5th position, final position), how much training they had received, and interactions between cue condition, training block, violation type, and LoE. To make the results more directly comparable to Trotter, Monaghan, and Frost (Chapter 5), we collapsed the 12 training blocks down to six (e.g. blocks one and two were pooled together). In Trotter, Monaghan and Frost (Chapter 5), this smoothing procedure was necessary due to model convergence issues.

Adding cue condition to the model including random effects and LoE did not improve model fit ($\chi^2(4) = 1.687$, $p = .793$), indicating that performance did not differ between each cue condition in isolation.

Adding the main effect of block to the model including random effects and LoE marginally increased model fit ($\chi^2(1) = 3.373$, $p = .066$), with classification performance increasing with more exposure.

Next, we analysed learning of different LoEs. Adding a fixed effect of LoE (LoE 1 or LoE 2) significantly improved model fit ($\chi^2(1) = 5.298$, $p = .0213$), indicating that LoE influenced participant performance, with higher accuracy for LoE 1 sequences.

Adding the main effect of violation position significantly improved model fit ($\chi^2(1)$ = 5.298, *p* = .021), reflecting that grammatical sequences were classified more accurately than ungrammatical sequences, whilst accuracy was similar across sequences including violations (see Table 1).

Adding the interaction for cue condition and LoE improved model fit ($\chi^2(4)$ = 13.974, *p* = .007), due to the interaction with the combined cue condition and LoE; relative to the baseline condition, participants trained with combined cues judged novel LoE 2 sequences less accurately.

Critically, including the two-way interaction between condition and violation position improved model fit ($\chi^2(8)$ = 58.669.373, *p* < .001), reflecting significantly higher performance than baseline for the detection of errors in the sequence final position for pitch cues, and significantly lower performance in the pause cues condition.

Including the two-way interaction between violation position and sequence length improved model fit ($\chi^2(1)$ = 16.392, *p* < .001).

The two-way interaction between block and condition did not improve model fit ($\chi^2(4)$ = 6.697, *p* = .153). Nor did the interaction between block and LoE ($\chi^2(1)$ = 0.495, *p* = .482), or the interaction between block and violation type ($\chi^2(2)$ = 3.692, *p* = .158).

Including the three-way interaction between sequence length, condition and violation position resulted in increased model fit ($\chi^2(1)$ = 19.899, *p* < .001). This reflects significantly improved classification accuracy for grammatical sequences in the pause cues condition at longer sequence lengths, relative to baseline.

Adding all three- and the four-way interaction did not improve model fit (*p*s > .05).

To summarise, participants' classification accuracy in all cue conditions was similar to baseline. Overall, mean performance was similar to chance. Participants classified grammatical sequences with above chance accuracy, and below chance accuracy for ungrammatical sequences, regardless of violation position (see figure 4), suggesting that participants were more likely to endorse any test item, harming performance on ungrammatical trials. The final model outcomes suggested that pitch cues, overall, negatively affected classification accuracy. Across all conditions, participants responded with increased accuracy with increased exposure. Although LoE was not found to contribute significantly to the final model as a main effect, the significant, negative, two-way interaction between LoE and combined cues, and three-way interaction between LoE, pause cues, and violation position indicate that LoE may have played a role in mediating performance. As this analysis raised the possibility that participants showed an overall bias in favour of grammatical sequences, we conducted an additional analysis based on signal detection theory to formally assess these claims.

Table 1

*Accuracy Final Model Outcomes*

| Fixed Effect | Estimate | Standard Err. | *z* | *p* |
|---|---|---|---|---|
| Intercept | -0.078 | 0.091 | -0.859 | .391 |
| LoE | -0.160 | 0.132 | -1.209 | .226 |
| Cue – Combined | 0.099 | 0.125 | 0.795 | .427 |
| Cue – Pause | 0.051 | 0.125 | 0.406 | .684 |
| Cue - Phonology | -0.054 | 0.125 | -0.431 | .666 |
| Cue - Pitch | -0.279 | 0.125 | -2.223 | .026* |
| 5th Position | -0.126 | 0.154 | -0.822 | .411 |
| Grammatical | 0.491 | 0.109 | 4.516 | <.001** |

| | | | | |
|---|---|---|---|---|
| Block | 0.029 | 0.011 | 2.748 | .006* |
| LoE: Cue - Combined | -0.380 | 0.183 | -2.075 | .038* |
| LoE: Cue - Pause | -0.237 | 0.182 | -1.300 | .194 |
| LoE: Cue - Phonology | 0.197 | 0.181 | 1.086 | .277 |
| LoE: Cue - Pitch | -0.109 | 0.185 | -0.059 | .953 |
| LoE: Violation – Grammatical | 0.237 | 0.172 | 1.378 | .168 |
| Cue – Combined: 5th Position | -0.307 | 0.217 | -1.415 | .157 |
| Cue – Pause: 5th Position | 0.199 | 0.211 | 0.943 | .346 |
| Cue – Pitch: 5th Position | -0.107 | 0.221 | -0.485 | .628 |
| Cue – Phonology: 5th Position | 0.042 | 0.210 | 0.201 | .841 |
| Cue – Combined: Grammatical | 0.047 | 0.149 | 0.317 | .751 |
| Cue – Pause: Grammatical | -0.315 | 0.149 | -2.130 | .033* |
| Cue – Phonology: Grammatical | 0.047 | 0.149 | 0.320 | 0.749 |
| Cue – Pitch: Grammatical | 0.417 | 0.150 | 2.783 | 0.005** |
| LoE: Cue – Combined: Grammatical | 0.336 | 0.238 | 1.413 | 0.158 |
| LoE: Cue – Pause: Grammatical | 0.546 | 0.236 | 2.314 | 0.021* |
| LoE: Cue – Phonology: Grammatical | -0.404 | 0.235 | -1.719 | 0.086 |
| LoE: Cue – Pitch: Grammatical | -0.033 | 0.239 | -0.136 | 0.892 |

Final model syntax: glmer(Accuracy ~ (1 + Condition*LoE*ViolationPosition + Block|Subject) + (1 + Condition*LoE*ViolationPosition + Block|Item) + Condition + Block + LoE + ViolationPosition + Condition:ViolationPosition + LoE:Condition + LoE:Condition:ViolationPosition, family = binomial(logit). The model analysed classification accuracy (1 vs. 0) on a total of 15360 trials.

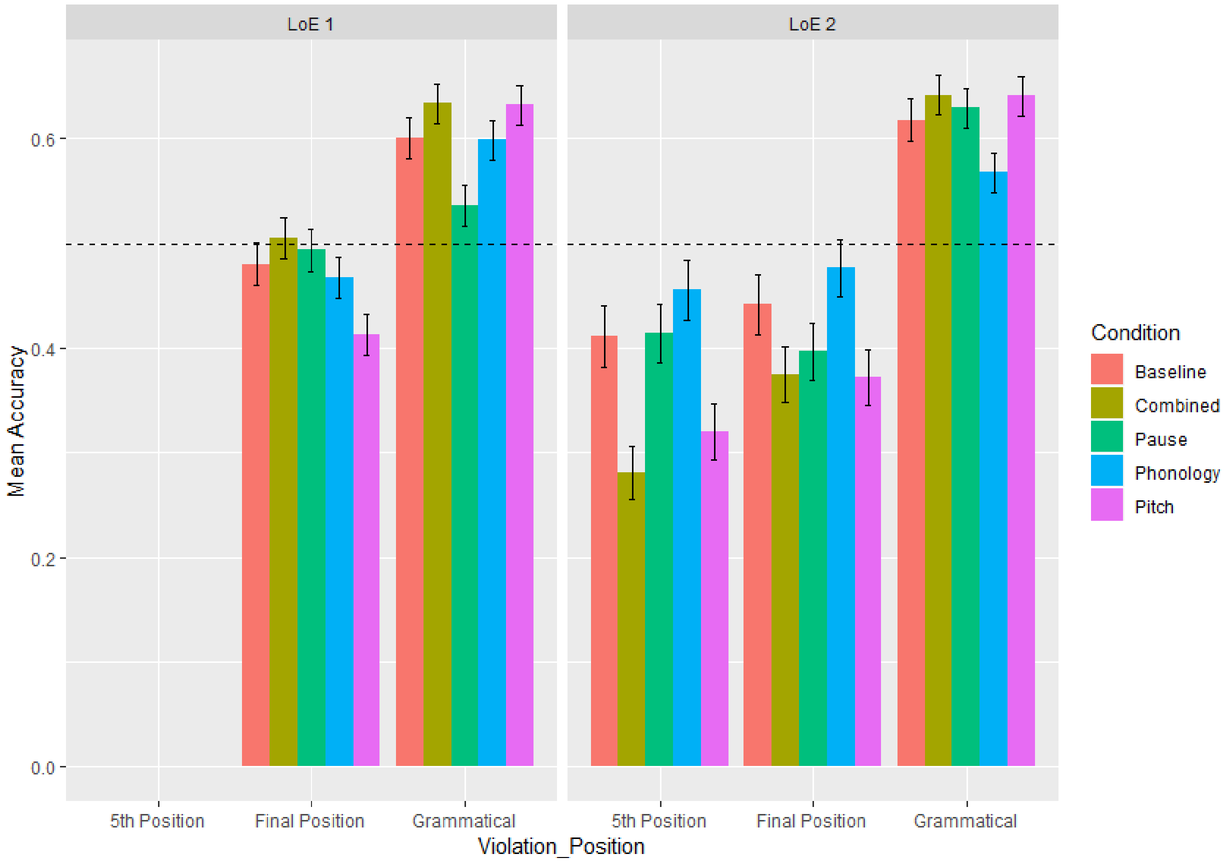


**Fig. 4**. This figure displays mean response accuracy by violation position by cue condition. The left panel illustrates the accuracy data for LoE 1 sequences (here, violations could only occur in the final position). The right panel displays the accuracy data for LoE 2 sequences. Black vertical bars indicate the standard error of the mean (SEM). The horizontal dashed line indicates chance performance.

### *3.2 Signal Detection Theory: Sensitivity Analysis*

Signal detection theory (SDT) can be employed whenever two possible stimulus types must be discriminated (Stanislaw & Todorov, 1999), in the present case, grammatical (signal trials) and ungrammatical (noise trials) stimuli. According to SDT, participants respond on the basis of the decision variable during each trial. If the decision variable is sufficiently high, the subject responds yes (grammatical), or no (ungrammatical). Correctly classifying a grammatical stimulus as grammatical is a hit, however, falsely classifying an ungrammatical stimulus as grammatical is termed a false alarm. SDT argues that participants' decision variable will be affected by prior input, therefore, the decision variable will elicit a distribution of values across grammatical and ungrammatical trials. The hit rate is the proportion of the signal distribution that exceeds the criterion, and false alarm rate is the proportion of noise distribution that exceeds the criterion. Using the hit and false alarm rate allows researchers to derive two aspects of participants' performance; their *sensitivity* to the signal, and their *response bias*. In the present paper, we employed the non-parametric measures of sensitivity A', and the response bias of A', *b*, given by equations (1) and (2) below, in accordance with Zhang and Mueller's (2005) correction:

$$(1)\ A' = \begin{cases} \frac{3}{4} + \frac{H-F}{4} - F(1-H) & if\ F \leq 0.5 \leq H\ ; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{F}{4H} & if\ F \leq H < 0.5\ ; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{1-H}{4(1-F)} & if\ 0.5 < F \leq H\ . \end{cases}$$

$$(2)\ b = \begin{cases} \frac{5-4H}{1+4F} & if\ F \leq 0.5 \leq H\ ; \\ \frac{H^2+H}{H^2+F} & if\ F < H < 0.5\ ; \\ \frac{(1-F)^2+(1-H)}{(1-F)^2+(1-F)} & if\ 0.5 < F < H. \end{cases}$$

In the present study, we computed A' and b for each participant per block and LoE. A' values of 0.5 are taken to mean that participants are unable to distinguish signal from noise, while b values of 1 indicate no response bias, with values greater than 1 indicating a bias towards no responses, while those less than 1 indicating towards yes responses.

To formally assess participants' sensitivity, we conducted a series of generalized linear mixed-effects models (LMER) predicting the dependent variable of A'. As both A' and b are computed using all trials within a block (1 – 12), we were unable to include by-items random intercepts and slopes in this analysis. By-subjects intercepts and slopes are retained in these models. Models were built up iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). Fixed effects were retained in the final model if they resulted in significantly improved model fit in isolation, or within an interaction. Interactions were retained in the model if they significantly improved model fit. As fixed effects, we tested the effect of the cues participants were exposed to (baseline, phonological, pitch, pause, and combined), testing block (1 – 12), and LoE (LoE 1 or LoE 2), and interactions between cue condition, block, and LoE.

Adding the effect of cue condition to the baseline model did not improve model fit ($\chi^2(7) = 4.250$, $p = .373$), indicating that sensitivity did not differ on the basis of cue condition in isolation.

The addition of block, however, significantly improved model fit ($\chi^2(1) = 6.086$, $p = .014$), suggesting that sensitivity increased across blocks. However, in the final model, this factor failed to significantly differ from 0, indicating that the variance explained by block in isolation can be attributed to its interaction with other factors.

The addition of LoE further improved model fit ($\chi^2(1) = 18.374$, $p < .001$), indicating that response sensitivity differed strongly on the basis of LoE (see figure 5 for greater detail).

The addition of each two-way interaction failed to improve model fit (*p*s > .05). However, the addition of the three-way interaction between cue condition, block, and LoE significantly improved model fit ($\chi^2(5) = 13.470$, $p = .019$). Table 2 below presents the final model outcomes, while figure 5 presents the three-way interaction in greater detail.

Table 1

*Sensitivity Final Model Outcomes*

| Fixed Effect | Estimate | Standard Err. | *t* |
|---|---|---|---|
| Intercept | 0.415 | 0.042 | 9.908** |
| Cue – Combined | -0.020 | 0.049 | -0.416 |
| Cue – Pause | -0.064 | 0.049 | -1.303 |
| Cue - Phonology | -0.072 | 0.049 | -1.470 |
| Cue - Pitch | -0.067 | 0.049 | -1.368 |
| Block | 0.006 | 0.003 | 1.717 |
| LoE | -0.072 | 0.035 | -2.070* |
| Cue – Baseline: Block: LoE | 0.0001 | 0.006 | 0.018 |
| Cue – Combined: Block: LoE | -0.009 | 0.006 | -1.390 |
| Cue – Pause: Block: LoE | 0.004 | 0.006 | 0.678 |
| Cue – Phonology: Block: LoE | 0.012 | 0.006 | 1.909 |
| Cue – Pitch: Block: LoE | -0.006 | 0.006 | -0.985 |

Model Syntax: A’ ~ (1 + Cue*Block*LoE|Subject) + Cue + Block + LoE + Cue:Block:LoE

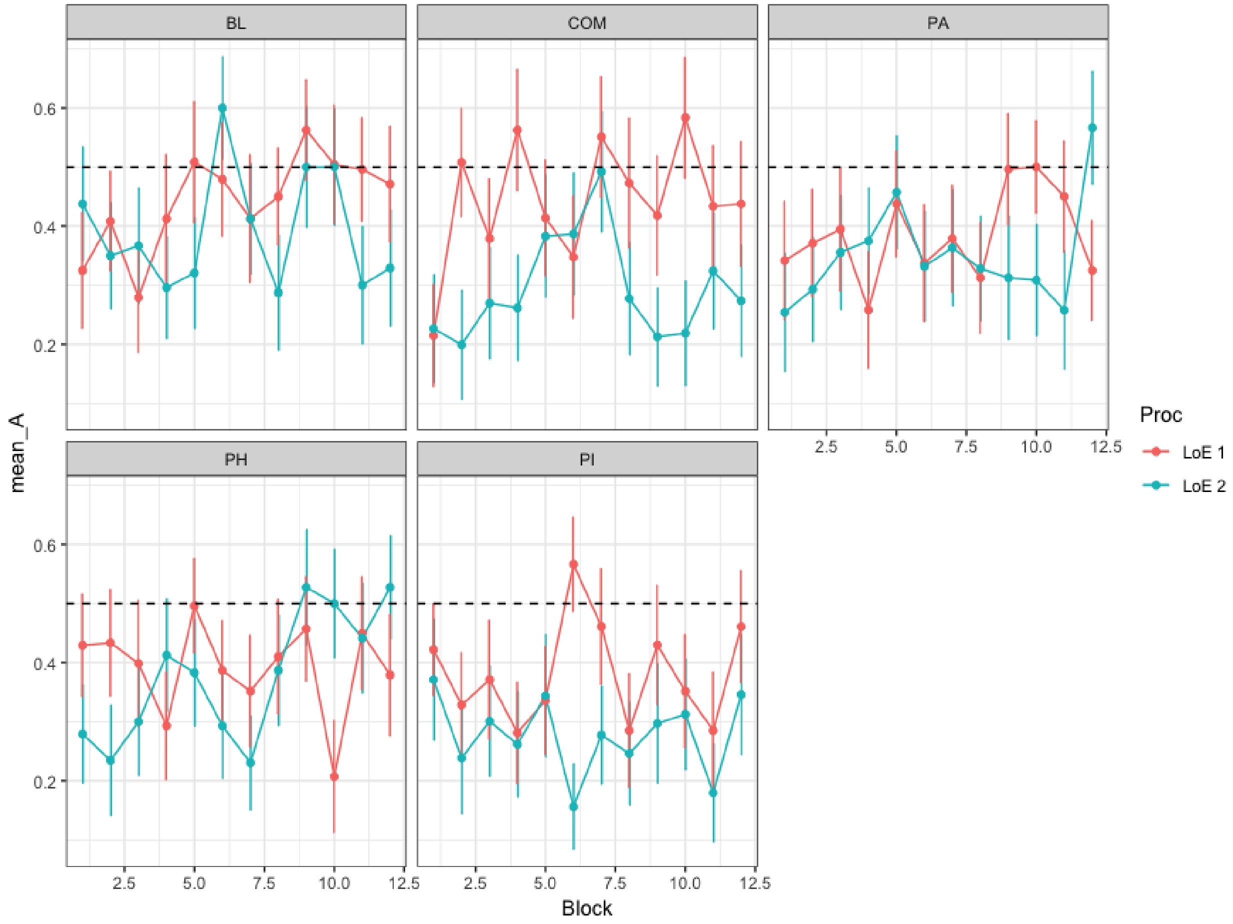


**Fig. 5.** Mean A' per block, split by LoE and Condition. Error bars display the standard error of the mean. 0.5 indicates that signals cannot be distinguished from noise, an is indicated by the dashed horizontal line.

### *3.3 Signal Detection Theory: Response Bias*

To formally assess participants response bias, we conducted a series of generalized linear mixed-effects models (LMER) predicting the dependent variable of b. By-subjects intercepts and slopes are retained in these models. Models were built up iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). Fixed effects were retained in the final model if they resulted in significantly improved model fit in isolation, or within an interaction. Interactions were retained in the model if they significantly improved model fit. As fixed effects, we tested the effect of the cues participants were exposed to (baseline, phonological, pitch, pause, and combined), testing block (1 – 12), and LoE (LoE 1 or LoE 2), and interactions between cue condition, block, and LoE.

Adding the effect of cue condition to the baseline model did not improve model fit ($\chi^2(4) = 3.522$, $p = .475$), indicating that cue condition in isolation did not affect response bias.

Adding the effect of block did not improve model fit ($\chi^2(1) = 0.699$, $p = .403$), indicating that participant bias did not change over the course of training.

Adding the effect of LoE, however, significantly improved model fit ($\chi^2(1) = 27.554$, $p < .001$), reflecting that on average, participants were more likely to provide yes responses to LoE 2 sequences. However, in the final model, the effect of LoE was not significant, indicating that the variance explained by LoE in isolation can be attributed to its interaction with other factors.

Adding the two-way interaction between cue condition and block did not improve model fit ($\chi^2(4) = 1.955$, $p = .744$), indicating that response bias did not change by conditions on the basis of block.

Including the two-way interaction between cue condition and LoE, however, produced significantly improved model fit ($\chi^2(4) = 11.718$, $p = .019$), indicated that participants showed a greater difference in response bias between levels of embedding (see figure 6 for greater detail). This was particularly striking for the combined cues and pause cues conditions, where response bias was much larger between LoE 1 and 2, indicating that longer sequences in these conditions greatly increased participants' response bias, irrespective of training.

Adding the two-way interaction between block and LoE, however, failed to improve model fit ($\chi^2(1) = 3.096$, $p = .079$), indicating that increased training did not statistically affect response bias.

Finally, adding the three-way interaction between cue condition, block, and LoE did not improve model fit ($\chi^2(5) = 3.889$, $p = .566$).

Table 1

*Response Bias Final Model Outcomes*

| Fixed Effect | Estimate | Standard Err. | *t* |
|---|---|---|---|
| Intercept | 0.622 | 0.066 | 9.381** |
| Cue – Combined | 0.091 | 0.092 | 0.982 |
| Cue – Pause | -0.020 | 0.092 | -0.213 |
| Cue - Phonology | -0.021 | 0.092 | -0.228 |
| Cue - Pitch | -0.122 | 0.092 | -1.326 |
| LoE | -0.114 | 0.076 | -1.513 |
| Cue – Combined: LoE | -0.269 | 0.105 | -2.554* |
| Cue – Pause: LoE | -0.083 | 0.106 | -0.785 |
| Cue – Phonology: LoE | 0.042 | 0.106 | 0.398 |
| Cue – Pitch: LoE | 0.042 | 0.106 | 0.398 |

Model Syntax: Response Bias ~ (1 + Cue*LoE|Subject) + Cue + LoE + Cue:LoE

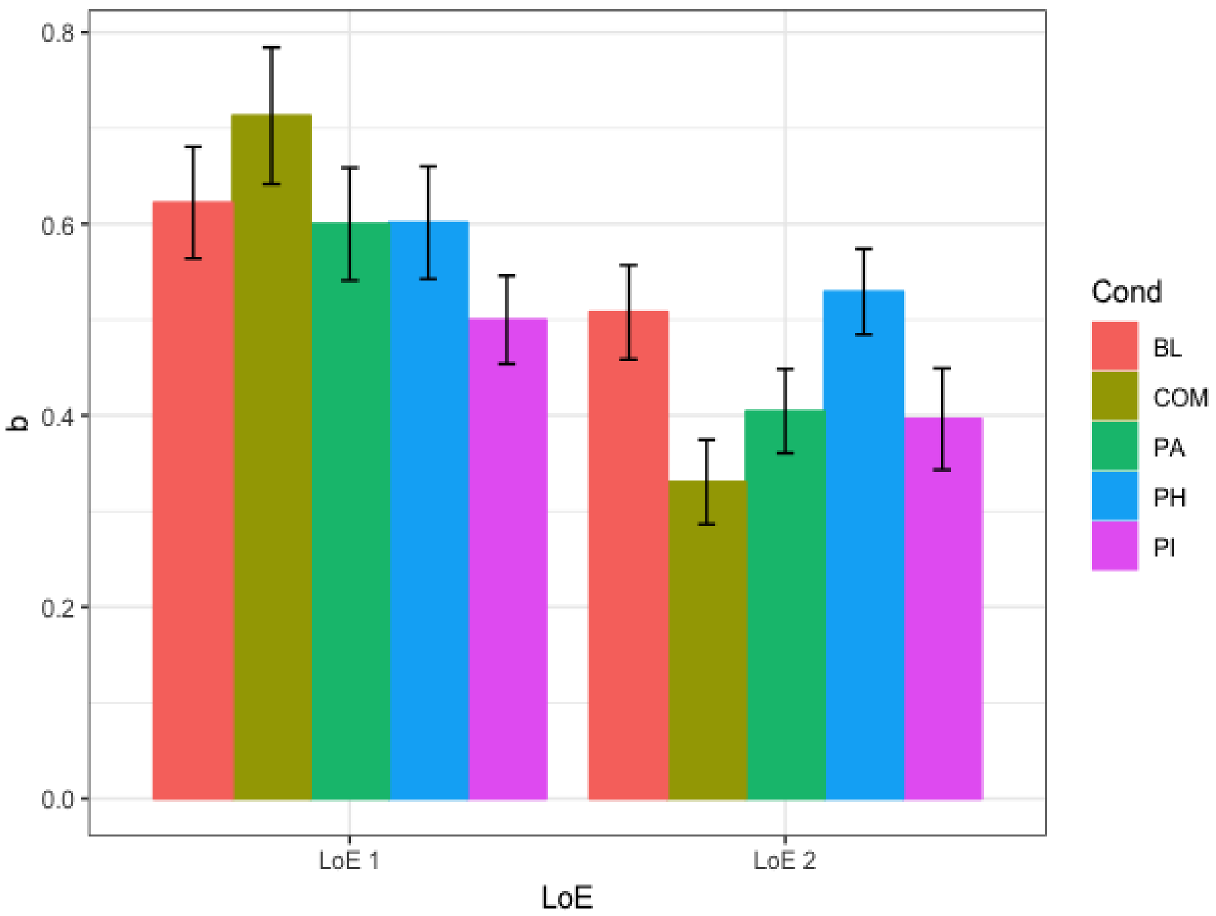


**Fig. 6**. Mean response bias by condition and LoE. Error bars display the standard error of the mean. b values closer to 0 indicate a greater bias towards yes responses.

## 5. Discussion

The aim of this study was to assess whether acoustic cues modelled on speech production data would facilitate the acquisition of hierarchically centre-embedded structures. Frank, Bod and Christiansen (2012) argue that the processing of speech in real-time may be sequential, with individuals relying on low-level, surface level cues to initially parse a sentence, and subsequently assign a syntactic structure based on this parse. Natural speech contains a rich set of cues from which phrasal groupings may be computed. Two likely candidates are pitch (Fery & Schubö, 2010; Trotter, Frost and Monaghan, Chapter 3) and temporal (Trotter, Frost and Monaghan, Chapter 3) cues sufficient to compute phrasal groupings via the pitch similarity and temporal proximity auditory-perceptual Gestalts (Deutsch, 2013). The speed of auditory processing would lend itself well to forming an initial parse of incoming speech, supporting comprehension. To test this claim, we implemented pitch cues consistent with those contained in HCEs based on Fery and Schubö (2010), and pause cues consistent with Hawthorne and Gerken (2015), and phonological similarity cues following Pena and colleagues (2002; see also Friederici et al., 2006). We predicted that the addition of each cue would result in increased learning (reflected in higher grammatical classification accuracy) with greater exposure, when compared to a baseline condition, which contained no cues other than co-occurrence statistics.

The results of this artificial grammar learning study did not fully support these predictions. Participant accuracy improved over training, suggesting learning occurred. However, a subsequent analysis using SDT indicated that sensitivity to the grammatical structure did not improve over learning, suggesting that either participants accuracy can be explained by statistical noise, or alternatively, that accuracy on trials not assessed under SDT (correctly classifying ungrammatical structures), may explain this effect. The main effect of cue condition did not reach significance in any of the analyses, suggesting that participants were unable to use pitch, pause, and phonological cues to group dependent elements in HCEs. The interaction between cue condition and block did not reach significance across any of the analyses, indicating that where learning did occur, learning was independent from temporal, pitch and phonological cues. In terms of accuracy, participants classified grammatical strings more accurately than ungrammatical strings. The response bias analysis indicated that this largely reflects a strong tendency towards yes responses across all conditions, that was especially pronounced for the combined cues condition. The significant interaction between the pitch cues condition and grammatical sequences suggested that pitch cues biased participants towards accepting any test item, all of which adhered to the pitch structure outlined in Fery and Schubö (2010). The response bias analysis supports this conclusion; in the pitch cues condition, bias was high across both LoEs, which was highest for LoE 2 structures. Conversely, pause cues elicited lower classification accuracy than baseline for grammatical LoE 1 strings, a finding that was not easily accounted for by the SDT analysis, pause cues elicited a similar response bias to baseline for LoE 1 strings. However, pause cues did produce relatively lower sensitivity, particularly in the intermediate blocks. Given these results, it seems a likely conclusion that whilst additional acoustic cues did not greatly affect learning, they were effective at capturing attention, producing response biases.

Why should these acoustic cues produce such robust response bias? One possibility is a tension between local, linguistic structure, and global, acoustic structure. Cues were present in both training and test structures. As a result, they were predictive of grammaticality in 75% of cases (the remaining 25% being ungrammatical stimuli, where they were used unreliably). Whenever pitch cues were present (in both pitch and combined cues), response bias was prevalent, suggesting participants found them highly salient. This in turn raises the possibility that participants were *overly* reliant, or more accurately, actively misled by them; local linguistic violations were obfuscated by global pitch structure, present over all test stimuli. Thus, the salience of pitch cues led participants to be more likely to endorse any test stimulus, resulting in higher than baseline accuracy for grammatical strings, and below chance accuracy for all kinds of grammatical violations. The SDT analysis would seem to support this; participants did not become more sensitive to the underlying structure than in the baseline condition, despite the increased accuracy. The resulting incorrect responses for ungrammatical strings, and subsequent corrective feedback would still lead to trial-and-error learning, explaining why classification accuracy was not perfect for grammatical sequences. Thus, overall it appears that pitch cues were salient, and did alter performance, however this may have inhibited learning.

In the pause cues condition, the sensitivity results suggested that participants did not provide a salient cue to global structure, with participant being largely insensitive to the underlying grammatical structure. However, they did elicit greater response bias than baseline for LoE 2 sequences, suggesting that they, at the least, captured some degree of attention, reflected by the three-way interaction with LoE and grammaticality in the accuracy analysis. Overall, this suggests that participants were largely not able to use temporal cues to detect dependencies, and therefore, we failed to replicate the findings of Mueller et al. (2010), where additional pauses between clauses between clauses of LoE 1 HCEs improved learning.

The phonological similarity condition did not differ from baseline on the basis of grammaticality, or LoE for either accuracy, or sensitivity. Thus, we can conclude that phonological similarity did not support the processing of phrase structure. This result fails to replicate Friederici and colleagues' (2006) findings. This could be due to the manner in which we tested knowledge of grammatical structure. In the present study, ungrammatical sequences did not violate the count of As and Bs, but rather the precise link between particular A and B syllables. As a result, we assessed specific knowledge of particular dependencies, as opposed to surface level properties of the structure.

The combined cues condition behaved similarly to both the pause and pitch cues conditions. Response bias was higher for LoE 2 strings, with greater sensitivity and accuracy for LoE 1 strings. The response bias difference was significant in its model, suggesting that the overlap of pause, pitch, and phonological cues interrupted response bias for LoE 1 strings, supported by the significant two-way interaction with LoE and condition. Temporal proximity cues have been previously shown to be able to be able to overpower pitch cues. Hamoui and Deutsch (2010) conducted a grouping preference study, where temporal proximity and pitch similarity suggested different groupings of the same sequence. Each sequence was comprised of twelve tones, where pitch similarity was high between tones one to four, five to eight, and nine to twelve, resulting in three groups of four tones. Pauses of differing lengths were inserted after every third tone, suggesting four groups of three tones. With short pauses, participants relied heavily on pitch similarity to group the sequences. However, as temporal distance increased, participants came to rely more heavily on temporal proximity to make their grouping decisions. Thus, the interaction of temporal proximity and pitch similarity in our study may have led participants to preferentially weight pause cues in their judgements, preventing response bias of the same magnitude observed for pitch cues. While this theory may account for the departures from the pitch results, it does not fully explain why the results depart from the pause results.

The role of multiple, interacting cues in language acquisition is contentious. Yu and Ballard (2007) posit that when multiple cues are present, their benefit is additive. This claim is supported by the authors' computational work, which sought to model infant word-referent mapping data. They constructed multiple models employing different cues; a distributional cues only model, distributional plus attention-based cues (e.g. gaze), distributional plus prosodic cues (where key words were highlighted with pitch cues), and a unified model using all cue types. The results indicated that the unified model outperformed all other models, supporting an additive account. Similarly, the intersensory redundancy hypothesis (Bahrick, Lickliter, & Flom, 2004) states that the overlap of multiple cues on a linguistic structure increases its salient. Further, the correlation of multiple cues is unlikely to occur by chance, suggesting a reliable relationship. Both of these accounts therefore suggest that the combined cues condition should have the greatest judgement accuracy, which the results did not support. On the other hand, these theories account well for Mueller et al.'s (2010) results. These theories, however, may not fully account for environmental noise, or the reliability of cues, which may have been problematic in the current study.

Rencently, Monaghan (2017) has proposed the theory of degeneracy. Here, multiple cues for language structure facilitate learning, as they provide a network of overlapping cue types that are resistant to their environmental variation. The novel aspect of the degeneracy theory is that it argues that noisy cues produce more robust learning: The variable presence of cues prevents the learner from selectively attending to a single cue, forcing them to make maximum use of the environment. Thus, environmental noise should produce more robust learning, allowing the language user to rely on many cues, or a smaller subset in any given situation. In a cross-situational learning task Monaghan, Brand, Frost and Taylor (2017) tested these claims, by varying the extent to which prosodic, gestural, and distributional cues were available to the participant, and found that when cues were present 75% of the time, learning was greatest. Thus, for learning a small set of words, in the presence of competitor objects, a degree of environmental variability supported word-referent mappings.

Formally distinguishing between these accounts is beyond the scope of the present study. However, the results do not conform to an additive effect of multiple cues; the accuracy and sensitivity data for the combined cues condition in part resembled both the results for the pause and pitch cues conditions but failed to outperform either. Similarly, under the intersensory redundancy hypothesis, it would be likely that the correlation between pitch, pause and phonological similarity data would provide correlating information, resulting in increased saliency, and greater learning. The degeneracy account suggests that the 75% reliability of each cue would lead to participants relying on the sum set of cues; if participants found pitch cues unreliable, participants may shift their attention to pause or phonological cues. This may explain as to why response bias and accuracy in the combined cues condition partially resembled both the pause and pitch cues conditions. However, the lack of probabilistic cues in the design of the present study prevents us from distinguishing between any of these accounts. Future artificial grammar learning research should consider implementing cues in their design which allow formal assessment of these theories.

We aimed to assess the claim that including surface-level acoustic and phonological speech cues would facilitate the acquisition of an artificial, hierarchical centre-embedded grammar. This was based on the idea that if hierarchical sentences can be processed sequentially, then individuals must compute groupings based on low-level perceptual biases. The results from the study were inconclusive, participants were not able to use each of these cues to support detecting the structure of the sequences in a short training regime. However, we demonstrated that pitch similarity improved classification accuracy for grammatical sequences of both LoEs and pause and combined cues supported classification accuracy for grammatical LoE 2 sequences. In future work, we suggest a higher-powered replication would help to both elucidate these effects, and better establish their reliability. Furthermore, on the basis of the supplemental analysis presented in Chapter 2, it may be wise to consider reducing the overall size of the vocabulary, as in this study, increased vocabulary sizes tended to produce smaller effect sizes.

# Chapter 5: Auditory-perceptual Gestalts affect the acquisition of hierarchical structure

Antony S. Trotter[1], Padraic Monaghan[1, 2], Rebecca L. A. Frost[3]

1. Department of Psychology, Lancaster University
2. University of Amsterdam
3. Max Planck Institute for Psycholinguistics

Following on from Chapter 4, this study assessed whether cues based on our participants' native language (Chapter 3) would facilitate the acquisition of syntax to a greater extent than those based on non-native cues (Chapter 4). This paper is currently in a draft ready for submission.

## Statement of Author Contribution

In the Chapter entitled, “Auditory-perceptual Gestalts affect the acquisition of hierarchical structure”, the authors agree to the following contributions:

Antony S. Trotter – 70% (Writing, experimental design, data collection, and analysis)

Signed: ________________________________ Date: ____________

Professor Padraic Monaghan – 15% (Experimental design, analysis, and review)

Signed: ________________________________ Date: 22.2.2019

Dr, Rebecca L. A. Frost – 15% (Experimental design, and review)

Signed: ________________________________ Date: 22.2.2019

## Abstract

Hierarchical centre-embeddings (HCEs) in natural language have been taken as evidence that language is not a finite state system (Chomsky, 1957). Recently, it has been argued that sequential processing drives their comprehension (Frank, Bod, & Christiansen, 2012). Sequential accounts state listeners employ surface level cues (e.g. semantic content, pitch and temporal variation) to determine the dependencies within an utterance. The results of a speech corpus study (Trotter, Frost, & Monaghan, Chapter 3) suggest that HCEs have pitch and temporal cues that could support dependency detection. English speakers produce the phrases of the embedded clause in a similar pitch that is distinct from the main clause, and pause prior to the onset of the embedded clause. Here, we assessed whether incorporating these cues would enhance learning in an artificial grammar learning study.

64 native English speakers were trained on an $A_nB_n$ grammar containing one and two levels of embedding (LoE) sequences. Participants were assigned to one of four cue conditions: baseline (distributional cues), temporal proximity (111ms pauses occur between LoEs), pitch similarity (dependent syllables occur at the same pitch) and combined (pause and pitch cues). At test, participants performed a grammaticality judgement task on novel structures.

Results indicated that overall, cues did not support learning. However, participants in the pitch cues condition showed lesser response bias, and greater sensitivity to the grammatical structure for LoE 2 structures. Temporal proximity did not result in greater than chance performance. Our results suggest that pitch similarity cues facilitate dependency detection in HCEs.

# 1. Introduction

Recursion is claimed to play a crucial role in human language, allowing for an infinite number of meanings to be expressed through the combination of a finite number of words (Hauser, Chomsky, & Fitch, 2002). A key focus of this research is sentences that contain multiple hierarchically organized centre-embeddings (HCEs). These structures are offer insights into how language processing can handle sentences generated by finite state, or phrase structure grammars. These grammars are levels of the Chomsky (1957; 1959) hierarchy, which classifies rule systems capable of generating an infinite set of sequences by defining increasing constraints on possible structures. FSGs are the weakest level of the hierarchy and can be fully specified by transitional probabilities between a finite number of states (Hauser & Fitch, 2004). To process sequences generated by an FSG simply requires a large enough memory stack to hold sequential states and the transitions between them, in order to concatenate them into longer sequences. PSGs are the next level of the hierarchy. They can similarly concatenate items, but can, crucially, embed strings within other strings, resulting in phrase structures and long-distance dependencies (e.g. HCEs, "[$_{c-1}$ The cat [$_{c-2}$ the man stokes] purrs]"). Generating and processing these complex structures requires more sophisticated mechanisms; an open-ended memory system and additional perceptual mechanisms are necessary to retain and recognise dependency relationships between elements separated by intervening words (between *cat* and *purrs*) (Hauser & Fitch, 2004). It is commonly suggested that PSGs are a crucial component of human language (Chomsky, 1959; Haegeman, 1991). Probing the processing differences between sequences generated by these two grammars provides evidence about the complexity of human language processing.

Research has shown that HCEs with more than three levels of embedding (LoE) are challenging to process, even for expert speakers (e.g. Bach, Brown, & Marslen-Wilson, 1986; Hudson, 1996; Newmeyer, 1988). Take, for example, this three LoE statute from s.1 of the British Road Traffic Act (1972), “A person [$_1$ who, [$_2$ when riding a cycle, [$_3$ not being a motor vehicle,] on a road or other public place,] is unfit to ride through drink or drugs,] shall be guilty of an offence.” This example illustrates that as more clauses are inserted, the distance between non-adjacent, dependent elements grows, in turn increasing the difficulty of learning or remembering associations between related constituents (Lai & Poletiek, 2011). This suggests that the ability to process PSGs has limits. Recent computational work (Ferrer-i-Cancho, 2015; Ferrer-i-Cancho & Gomez-Rodriques, 2016) suggests that as dependency lengths increase, the probability of producing non-adjacent, HCE structures decreases. In a large sample of syntactic trees taken from the Stanford and Prague corpora, Ferrer-i-Cancho and Gomez-Rodriquez (2016) found a positive correlation between the sum of dependency lengths in a sentence, and the number of crossings a sentence would contain (see figure 1 for examples). When dependency lengths increase, so too does the difficulty of associating constituent elements, increasing the probability that a sentence will include one or more crossing dependency. This would suggest that the complexity of HCEs is limited in natural language. Given the challenges associated with processing HCEs, it is important to ask what cues present in natural language might support individuals to acquire and process them.

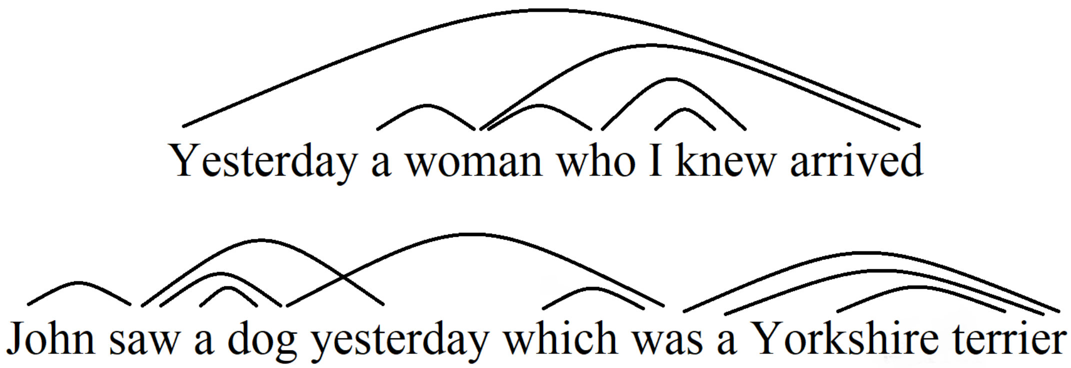

Yesterday a woman who I knew arrived
John saw a dog yesterday which was a Yorkshire terrier

Fig. 1. The top sentence illustrates a sentence without (top) and with (bottom) crossing dependencies. The sum of dependency lengths, d, in the top sentence is 16, number of crossings, c, is 0. In the bottom sentence, d = 18, c = 1.

A flexible method for investigating the processing of linguistic structure is the artificial grammar learning (AGL) paradigm. Typical AGL studies include two phases; training and testing. During training, participants are presented with sequences (of letters or nonsense words) that are – unbeknown to participants – grammatical sequences generated by an experimental grammar. At test, participants are presented with novel sequences that conform to the rules of the grammar, or violate them. Participants typically perform a grammatical judgement task on each sequence. Successful learning is typically defined as above chance classification of sequences as being either grammatical or ungrammatical.

A key set of AGL studies have distinguished participants' ability to learn sequences generated using an $A_nB_n$ or an $AB_n$ structure. Both employ two word categories, A and B. The $A_nB_n$ rule produces a sequence of As followed by a matching number of Bs. Any pair of words – $A_iB_i$ – can be centre-embedded into another pair – $A_jB_j$ – to produce a longer sequence - $A_jA_iB_iB_j$. Critically, the $A_nB_n$ rule corresponds to the PSG level of the Chomsky hierarchy. In contrast, the $AB_n$ rule follows a right-attachment rule, such that any pair in the language – $A_iB_i$ – can be concatenated with another – $A_jB_j$ - to produce a longer sequence - $A_iB_iA_jB_j$. The $AB_n$ rule is thus a FSG. Crucially, assessing the processing differences between PSGs and FSGs provides evidence on how fully participants' linguistic processing can be specified by each level of the Chomsky hierarchy. $A_nB_n$ languages have proven difficult to acquire, so measuring their processing is difficult to accomplish.

To properly parse a HCE, it is necessary to determine the dependencies between particular As and Bs. In reference to a typical HCE ("[$_{A1}$ The boy] [$_{A2}$ the girl] [$_{B2}$ chases] [$_{B1}$ runs]"), correspondences between individual As and Bs can be conceptualised as being between nouns and verbs. To determine whether specific dependencies can be acquired, Bahlmann, Schubotz and Friederici (2008), and Friederici, Bahlmann, Heim, Schubotz and Anwander (2006) designed $A_nB_n$ and $AB_n$ sequences where $A_i$ ("de") and $B_i$ ("fo") always co-occurred; if "de" was present in a sequence, "fo" always appeared in the relevant B position. Word category membership was marked by the vowel (A words always paired a plosive with "e", B words paired plosives with "i"). Participants were assigned to $A_nB_n$ or $AB_n$ groups, and received 12 blocks of training. Training blocks first presented participants with 10 grammatical sequences, followed by a grammatical classification task on 10 novel sequences (half of which were grammatical). After each response, participants were provided with corrective visual feedback on the accuracy of the response. Learning was assessed by performance on a grammatical classification task on 160 novel sequences (80 grammatical). Participants classification accuracy was above chance for both rules.

However, de Vries (2008) noted that the $A_iB_i$ pairings shared phonological properties; a simple counting strategy could be employed to detect grammatical violations. Specifically, the violation sequence $A_1A_2A_3B_3B_2A_4$ can be detected by counting the number of syllables ending in "e" or "i". When stimuli were constructed to prevent a counting strategy, de Vries et al. (2008) demonstrated that HCE learning was no longer possible.

Under specific circumstances, though, acquisition of HCE structure has been demonstrated in empirical research. Lai and Poletiek (2011; 2013) found that participants could learn particular A-B dependencies when trained with a “starting small” regime, whereby participants are initially trained on isolated A-B pairs, before moving up to more complex sequences where A-B pairs are centre-embedded in others; i.e. if particular dependencies are highlighted. Further, Peña et al. (2002) showed that participants’ acquisition of an $A_nxB_n$ grammar (where co-occurring A-B pairings are separated by any intervening x syllable), is assisted when they are provided with 25ms pauses between strings. Grammatical acquisition was measured using a classification task in which participants selected between novel words and part-words (e.g. in the training sequence, $A_1xB_1A_2xB_2A_3xB_3$, $A_1xB_1$ is a word, $xB_1A_2$ and $B_2A_3x$ would be part-words). The authors suggested two reasons why this should be the case; that inter-syllabic pauses made the stimuli more speech-like, thus triggering language-like computations, and that it explicitly brackets sequences, highlighting dependencies.

Perruchet, Tyler, Galland and Peereman (2004) dispute whether Pena et al.’s (2002) result purely reflected non-adjacent dependency learning. In two experiments, they probed the extent to which participant accuracy on the forced choice task could be explained by other sources of information in the materials. In the first, they preserved the AxB structure of materials, whilst removing statistically specified dependency relationships (and A could pair with any B), and trained participants both with and without pauses. At test, participants noted down any words they perceived in the speech stream. Considering only trisyllabic words, 72.35% of responses in the no-pause group adhered to the AxB pattern. Given the absence of statistical regularities and temporal bracketing, this suggests that participants were able to use factors not controlled for in Peña et al.’s

(2002) materials to segment the speech stream into AxB words, and further, that the novel word ($A_i$x$B_i$) vs. part-word (x$B_i$$A_j$) forced choice task fails to demonstrate grammatical acquisition.

In a second experiment, Perruchet et al. (2004) replicated Pena et al.'s (2002) study. Critically, however, they adjusted an additional comparison between rule words ($A_i$x$B_i$) and scrambled words (initial and final syllables drew from any word family, e.g. $B_i$x$A_j$, $A_i$x$B_j$), allowing assessment of specific dependency learning. Participants selected rule words over scrambled words more often than chance, indicating that dependency learning did occur. However, the effect size was significantly smaller than Pena et al.'s (2002), suggesting that Pena and colleagues' results can be mostly explained by positional information. While the learning of non-adjacent dependencies can occur, AGL experiments need to carefully construct tests to assess whether participants have acquired specific dependency information, i.e. to fully demonstrate acquisition of an artificial grammar.

It is important to note that whilst a sentence can be hierarchically structured, it does not necessarily follow that individuals will process it hierarchically. Frank and Bod (2011) compared the word-probability estimates from three probabilistic language models – embedded with different psychological mechanisms and representations – against reading-time measurements of the Dundee corpus (comprised of 2368 sentences). Three classes of model were implemented. The first class - PSG models - was induced from treebanks, and utilised hierarchical structure. The second - Markov models – and third – Echo State Networks – only had access to sequential structure. Critically, the PSG model failed to estimate variance in reading time data above all of the sequential structure models. Unlike other sources of information, hierarchical structure did not noticeably affect the generation of expectations about upcoming words.

Frank, Bod, and Christiansen (2012) have therefore suggested that the comprehension of hierarchical structure is driven by sequential processing. Under the sequential processing theory, to comprehend a HCE structure in a rapidly unfolding temporal context, the listener would need to rely on superficial surface level cues to determine its dependencies. For example, in the statute from the British Road Traffic Act (1972), world knowledge can be used to determine the syntactic relationships. Bicycles can neither be inebriated, nor can they be guilty of an offence, though they often are ridden on roads. Humans, however, can ride bicycles, imbibe drink and drugs, feel the appropriate effects, and commit criminal offences. This generates the basic units "person riding bike" "(if) person is drunk", "(then) person is guilty". Subsequently, you assign a syntactic structure informed by this semantic parse.

Semantic cues are not the sole cue available in speech for supporting the detection of dependencies; human speech is rich with prosodic cues that may trigger auditory processing biases that provide grouping information to clausal membership. For the present study, two processing biases are particularly important; pitch similarity and temporal proximity (Deutsch, 2013). The former states that individuals tend to group sequential sounds that are similar in pitch, and to distinguish between those distinct in pitch. The latter states that you are likely to group sequential sounds that occur closer together in time and distinguish between tones that are more temporally distant. If speech contains cues consistent with these Gestalt mechanisms, it may provide the processor a way of processing hierarchical structure non-hierarchically.

Trotter, Frost, and Monaghan (Chapter 3) conducted an acoustic analysis of a corpus of spontaneously produced active-object ("[The bear] [the girl] [is hugging] [is brown]") and passive relative clauses ("[The bear] [being hugged] [by the girl] [is brown]"). We hypothesised that pitch similarity would be highest between syntactically dependent phrases, and that pauses occurring between clauses would be longer than elsewhere in the speech. Contrary to our prediction, for passive relatives, pitch similarity was highest between the first two phrases of the sentence ("The bear being hugged"), and the longest pause preceded the by-phrase. In contrast, in line with our hypotheses, for active object relatives, pitch similarity was highest between the phrases of the embedded clause ("the girl is hugging"), and the embedded clause was preceded by a pause longer than elsewhere in the speech. Thus, for active-object relative clauses, pitch similarity and temporal proximity provide grouping information consistent with syntactic structure. In terms of sequential processing, prosodic information may therefore, potentially support rapidly determining dependencies, enabling rapid interpretation of speech. However, finding the presence of these cues does not prove that they are useful for comprehension, sequential or otherwise.

Trotter, Monaghan, and Frost (Chapter 4) assessed whether acoustic grouping cues would support the acquisition of HCE structure in an AGL study. Participants were randomly assigned to one of five cue conditions: Baseline (only distributional cues); Pitch similarity (dependent syllables occurred in a similar pitch); Temporal proximity (lengthened pauses occurred between LoEs); Semantic similarity, marked by phonological cues (dependent syllables began with the same plosive, e.g. Ba Du De Bo); and Combined (pitch similarity + temporal proximity + phonological cues). These cues were present over training and testing. Participants received 12 blocks of training and testing. In each training block, participants were presented with 16 grammatical structures (8 LoE 1, 8 LoE 2). At test, participants were presented with 16 novel sequences, with eight of each LoE, half of which were grammatical. The results indicated that participants were unable to use temporal proximity or phonological cues to group HCEs; grammaticality judgement accuracy did not differ from baseline. Pitch cues were salient, resulting in higher than baseline judgement accuracy for grammatical sequences, though lower than chance accuracy for ungrammatical sequences. This suggests that participants were responding positively to the global, acoustic pitch structure, reducing sensitivity to local, linguistic violations; pitch structure may have masked linguistic violations. Finally, the combined cues partially resembled both the pitch and pause cues conditions, suggesting an interaction; participants had higher accuracy when judging LoE 2 grammatical sequences (similar to the pitch condition), but failed to differ from baseline for LoE 1 sequences (similar to the pause condition).

While the results of these studies are intriguing, the cues used in this study may have been problematic. Notably, the pitch similarity cues were based on the production data of German native speakers (Fery & Schübo, 2010). These cues were salient, evidenced by the bias towards accepting any string adhering to the acoustic structure, however, it remains possible that the native-English speaking sample's lack of exposure to these pitch cues failed to facilitate proper grouping of dependent elements. Participants were also unable to use temporal proximity cues. These pause cues were taken from prior AGL work assessing acquisition of clausal membership (Hawthorne & Gerken, 2014), leading to the question of whether they were salient in the context of HCEs. To address these concerns, in the present study, we assessed whether stimuli based on a corpus of native English-speech (Trotter, Frost, & Monaghan, Chapter 3) would result in different outcomes. More specifically, the corpus results suggested shorter pauses between LoEs, and a higher pitch for each LoE. We anticipate that this may increase the saliency of the pitch cues, but potentially reduce the saliency of pause cues.

An important question regarding rhythmic and tonal information is how the cues interact. Yu and Ballard (2007) suggest that the effect of cues is additive. They assessed this theory by conducting a computational modelling study of human infant word-referent mapping data. They constructed several models which employed a range of cues; (1) distributional cues, (2) distributional cues and attention-based cues (e.g. gaze), (3) distributional cues and prosodic cues (highlighting of key words with a higher pitch range, overall pitch, and an exaggerated pitch contour), and (4) a unified model using distributional, gaze and prosodic cues. Their findings indicated that the unified model outperformed all other models, supporting an additive account. Another model that explains the role of multiple, interacting cues is the intersensory redundancy hypothesis (Bahrick, Lickliter, & Flom, 2004). This account suggests that when multiple cues highlight the same linguistic structure, they make it salient and suggest that the cues are informative. At the same time, it suggests that the relationship is not random; correlated cues are unlikely to occur by chance, as opposed to the correspondence between a single cue and a structural feature of the language. Correlated cues thus increase in saliency and become increasingly attended to over learning. However, these account do not make explicit predictions about environmental variation. For example, long pauses can occur at syntactic boundaries suggesting they are a useful cue to structure (e.g. Cooper, Paccia, & Lapointe, 1978), however, this is not always the case, as around 50% of long pauses occur at non-syntactic boundaries (Fernald & McRoberts, 1996). How do learners account for the fact that cues are not always present, or reliable?

Environmental variability central to Monaghan's (2017) theory of degeneracy. Here, multiple cues for language structure facilitate learning, as they provide a network of overlapping cue types that are resistant to their environmental variation. The novel aspect of the degeneracy theory is that it argues that noisy cues produce more robust learning: The variable presence of cues prevents the learner from selectively attending to a single cue, forcing them to make maximum use of the environment. Thus, environmental noise should produce more robust learning, allowing the language user to rely on many cues, or a smaller subset in any given situation. In a cross-situational learning task Monaghan, Brand, Frost and Taylor (2017) tested these claims, by varying the extent to which prosodic, gestural, and distributional cues were available to the participant, and found that when cues were present 75% of the time, learning was greatest. Thus, for learning a small set of words, in the presence of competitor objects, a degree of environmental variability supported word-referent mappings. Only degeneracy offered a potential explanation for Trotter, Frost, and Monaghan's (Chapter 4) results, as multiple cues resulted in performance that did not exceed any individual cue, and partially resembled the performance of each. Thus, in the current study, it is of interest as whether cues modelled on participants' native language experience will prove more reliable individually, and whether the noisy overlap of these cues will result in preferential performance.

The present study is primarily an attempt to assess whether implementing natural language cues in line with participants' native language experience (c.f. Trotter, Frost, & Monaghan, Chapter 3) will result in better learning than when they are taken from German data (c.f. Trotter, Monaghan, & Frost, Chapter 4). The current AGL study thus employs three cue conditions based on our corpus data; pitch similarity (where syllables within an LoE use the same pitch, with the first and last word using the highest and lowest pitch, respectively), temporal proximity (where longer bracketing pauses occur between LoEs), and a combined condition. By comparing these conditions to baseline, we can assess which cue type best facilitates learning of hierarchical structure. Including the combined cues condition allows us to ask whether a combination of cues produces greater learning. In relation to our prior AGL study, we removed the phonological similarity condition, due to potential confounds it introduces, allowing a purer contrast between the individual speech cues, and the combined condition. We hypothesised that: (1) relative to baseline, each cue will improve learning; (2) participant performance will improve with more training; (3) participants will be less accurate with longer stimulus lengths, due to increased complexity.

## 2. Method

### *2.1 Participants*

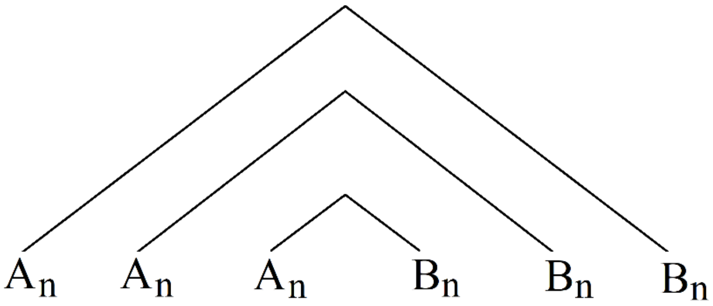
64 native English speakers ($\text{Mean}_{\text{Age}}$ = 19.143, $\text{SD}_{\text{Age}}$ = 2.928, $n_{\text{female}}$ = 50) participated in the study, resulting in 16 participants per condition. All participants were students at Lancaster University.  Participants received £3.50 or course credit for their participation.

### *2.2 Materials and Design*

Finite state grammar sequences were constructed following $A_nB_n$ rules (see Figure 2), producing sequences conforming to a hierarchical centre-embedded structures. Therefore, each sequence contains two word categories, A and B. Each word category contained six consonant-vowel syllables, resulting in twelve syllables per grammar. Words in both categories were monosyllabic, and were comprised of a plosive consonant (“P”, “B”, “G”, “D”, “T”, “K”) and a vowel, or vowel pairing (“a”, “e”, “i”, “o”, “u”, “oi”). The set of syllables was generated by randomly pairing a plosive with a vowel. We generated two separate languages to assess whether participants’ learning was driven by phonological factors external to the manipulations. In each version of the language, individual consonants and vowels occurred once per category, with no repetitions of consonant-vowel pairings. Therefore, this resulted in a total set of 24 syllables. *Language 1* was comprised of the following syllables: *A*: “Pe”, “Bu”, “Gi”, “Doi”, “To”, “Ka”; *B*: : “Ku”, “Ta”, “Po”, “Bi”, “De”, “Goi”. Language 2 was comprised of the following syllables: *A*: “Gu”, “Di”, “Te”, “Bo”, “Koi”, “Pa”; *B*: “Ti”, “Ge”, “Ko”, “Poi”, “Ba”, “Du”. These baseline languages were employed in the baseline, pause and pitch conditions.

Each syllable was created using the Festival speech synthesiser (Black, Taylor, & Caley, 1990). In each case, syllables were generated using the default voice, at the default rate. In addition to these default parameters, we specified the target pitch level (180Hz, 165Hz, 150Hz) using the “Default intonation”, which allows the researcher to specify the pitch at the beginning and end of the utterance. In each case, we specified both at the target pitch level. To ensure each syllable had the target pitch value, we subsequently assessed this in Praat (Version 6.0.13; Boersma, Paul & Weenink, 2016), and corrected the pitch contour when necessary. Each monosyllable lasted between 133 and 182ms (mean = 157ms, SD = 13ms). This variance resulted from differences in vowel (e.g. “e” had a shorter duration than “oi”) and consonant durations (e.g. “p” has an unvoiced onset, whereas “g” does not) that were implemented in the default voicing parameters employed by Festival.

Hierarchical centre-embedded structures were generated using the $A_nB_n$ rules. Each $A_i$ syllable was paired with a $B_i$ syllable (e.g $A_1B_1$), resulting in six grammatical pairings per language, indicated using numbered indices. To generate grammatical sequences, any $A_iB_i$ ($A_1B_1$) pairing could be inserted within any other $A_iB_i$ pairing ($A_6B_6$) to a minimum of one level of embedding (LoE; $A_6A_1B_1B_6$, hitherto LoE 1) and a maximum of 2 LoE ($A_6A_1A_3B_3B_1B_6$, hitherto LoE 2). Sequences violating the experimental grammar were generated for the test phase, in which one B syllable failed to match all A syllables in the sequence. For ungrammatical sequences, two additional constraints were used; the same B syllable could not occur more than once in the sequence, and no adjacent $A_iB_i$ violations could occur. Therefore, in LoE 1 sequences, violations always occurred in the final position ($A_6A_1B_1B_4$), and in LoE 2 sequences, violations occurred in either the fifth ($A_6A_1A_3B_3B_2B_6$) or sixth sequence positions ($A_6A_1A_3B_3B_1B_5$). Figure 2 illustrates the grammatical structure in detail, and provides examples of grammatical and ungrammatical sequences



Language 1:

| | | |
|---|---|---|
| LoE 1 G: | $A_1A_2B_2B_1$ | Pe Bu Ta Ku |
| LoE 1 U (4th Position): | $A_3A_4B_4B_1$ | Gi Doi Bi Ku |
| LoE 2 G: | $A_4A_5A_6B_6B_5B_4$ | Doi To Ka Goi De Bi |
| LoE 2 U 1 (5th Position): | $A_3A_2A_1B_1B_4B_3$ | Gi Bu Pe Ku Bi Po |
| LoE 2 U 2 (6th Position): | $A_6A_3A_2B_2B_3B_5$ | Ka Gi Bu Ta Po De |

Figure 2: Example structures from language 1. General structure and examples of stimuli in the $A_nB_n$ PSGs. G and U indicate grammatical and ungrammatical sequences, respectively.

The experiment utilised a between subject design, with participants randomly assigned to one of four different cue conditions; *baseline* (no cues), *pause*, *pitch*, and *combined* (pause + pitch). In each condition, cues were present over both training and testing.

In the baseline condition, the only cues that could guide learning were the frequencies with which dependent syllables in the sequences co-occurred. In the baseline, phonological similarity, and pitch conditions, 5ms inter-syllable pauses were 5ms, reflecting the 25th percentile of non-critical pauses in Trotter, Frost, and Monaghan (Chapter 3).

The pause condition employed temporal grouping cues that could highlight the dependency structure of the language; 111ms pauses occurred between levels of embedding (e.g. $A_1$ [pause] $A_2B_2$ [pause] $B_1$). This duration reflects the mean inter-clause pause duration found in Trotter, Frost, and Monaghan (Chapter 3).

For the pitch condition, syllables within an LoE used a similar pitch, with 15Hz difference between levels. The first and last syllable of the sequence always occurred with the highest and lowest pitch respectively (LoE 1; $Pa_{180Hz}$ $Te_{165Hz}$ $Ko_{165Hz}$ $Du_{150Hz}$: Loe 2; $Pa_{180Hz}$ $Te_{165Hz}$ $Doy_{150Hz}$ $Bi_{150Hz}$ $Ko_{165Hz}$ $Du_{150Hz}$). The sequence initial pitch (180Hz) and pitch reduction between phrases reflect the results of Trotter, Frost, and Monaghan (Chapter 3) for active-object relatives. To verify whether the pitch manipulation was detectable by our participants, each participant in the pitch and combined cues condition (n = 32) was administered an informal pitch sensitivity test, wherein they were played two examples of a structure from each LoE. One of these sequences was canonical with the experimental pitch structure (180Hz, 165Hz, 165Hz, 150Hz) and one which was randomised (e.g. 150Hz, 165Hz, 180Hz, 165Hz), and asked whether they “sounded the same” or were different. 27 participants (84%) were able to detect the difference.

Finally, the combined cues condition employed both pitch and pause cues.

### *2.3 Procedure*

Participants were randomly assigned to one of the five experimental conditions and told that they would be presented with linguistic items that followed a sequential rule. At the start of each block, a message appeared at the centre of the screen that instructed participants to press any key to begin the familiarization phase. In each familiarisation phase, participants passively listened to 16 strings that adhered to the grammar of the language. Eight of these strings contained one level of embedding, and the remaining eight contained two levels of embedding. Stimuli were presented in a randomised order. Syllables were presented sequentially, in isolation. Whole strings were separated by 3000ms pauses.

Following familiarization, participants were presented with text informing them that the test block would begin after they pressed any key. In the testing phase, participants were presented with 16 novel sequences. After each sequence, participants performed a forced grammatical classification task; participants were required to indicate – via keyboard response – whether the sequence adhered to the underlying linguistic rules, pressing "Y" for yes, or "N" for no, after which they received corrective feedback. There were 16 trails in each testing block, with eight LoE 1 sequences, and eight LoE 2 sequences. Four of each LoE sequences were ungrammatical, and the remainder were grammatical (see Figure 3 for examples of grammatical and ungrammatical test stimuli). Test items included the same cues as training items.

In total, participants completed 12 blocks of familiarization and testing. The experiment lasted for approximately 40 minutes.

## 3. Results

### *3.1 Grammatical Judgement Accuracy*

Overall, grammaticality judgement accuracy was relatively similar between conditions. Participants in the combined cues condition had the lowest overall accuracy (Mean = 0.486, SEM = 0.009), followed by the pitch condition (Mean = 0.501, SEM = 0.009), pause (Mean = 0.511, SEM = 0.009), and baseline (Mean = 0.511, SEM = 0.009). Overall judgement accuracy was therefore around chance in each condition, suggesting participants were unable to acquire the experimental grammar. Figure 3 below displays the per-block mean accuracy (and standard error of the mean) for each condition, split by LoE. Visual inspection of this figure suggests that grammaticality judgement accuracy differed on the basis of testing block, and LoE, with consistently lower accuracy for LoE 1 sequences relative to LoE 2 sequences. At LoE 2, accuracy appears to differ from chance in several blocks for the pause and pitch cues conditions, suggesting that acoustic cues may facilitate processing in longer sequences.

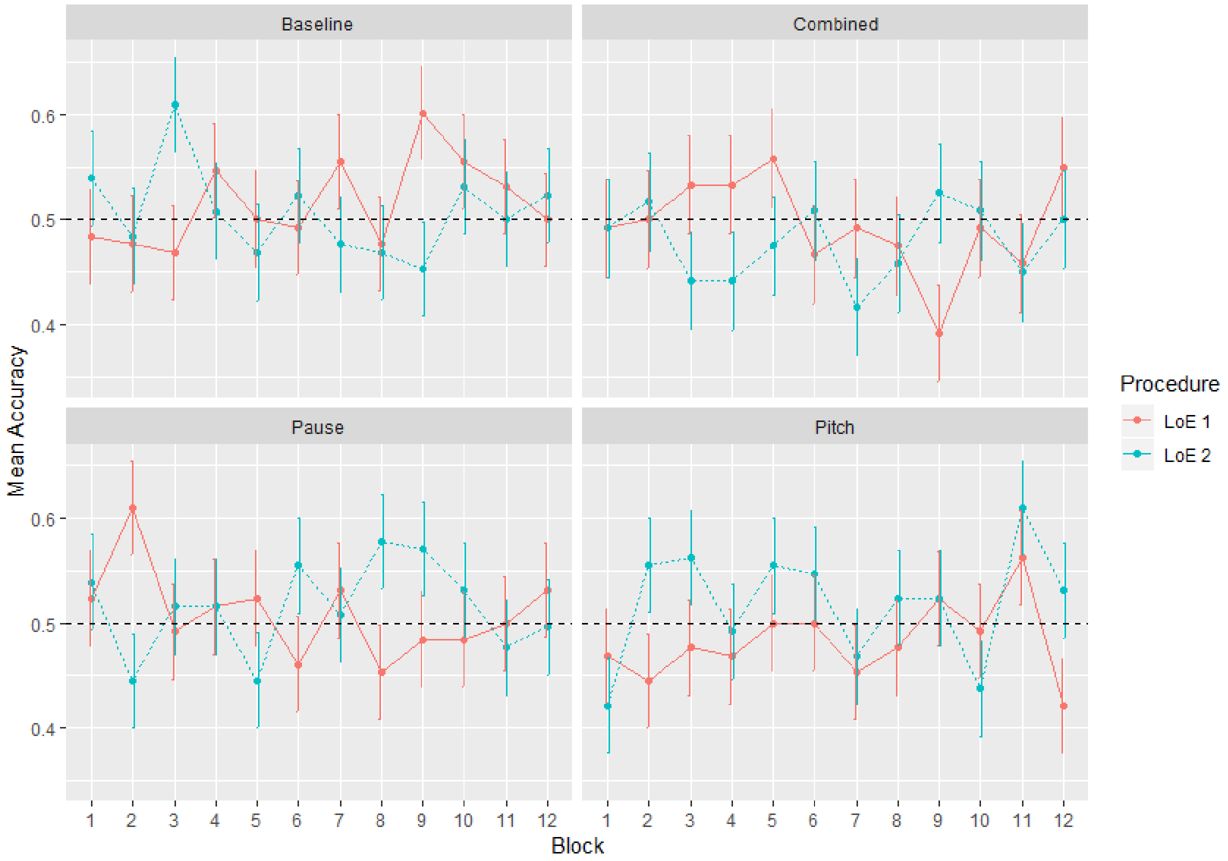


Fig. 3. Displays the mean judgement accuracy per condition, block, and LoE. Red, solid lines and points display the means for LoE 1 sequences, blue lines and points display the means for LoE 2 sequences. Vertical coloured lines display the standard error of the mean (S.E.M.). The horizontal dashed line indicates chance performance.

To analyse these effects, we conducted a series of generalised linear mixed-effects models (GLMER) predicting the dependent variable of accuracy (correct vs. incorrect) with a logit-link function. Random intercepts and slopes for participants and items were included in all reported analyses. Models were built up iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). Fixed effects were retained in the final model if they resulted in significantly improved model fit in isolation, or within an interaction. Interactions were retained in the model if they significantly improved model fit.

As fixed effects, we tested the effect of cue condition (baseline, pause, pitch, and combined), the amount of training they had received, violation type (Grammatical, 5th position, final position), and all interactions between these fixed factors. For the purposes of these analyses, we collapsed the 12 test blocks down to six (i.e. blocks one and two were pooled together). This was necessary due to model convergence issues; when 12 blocks were used as a main effect, models largely failed to converge, and were hence uninterpretable. Implementing this smoothing procedure resulting in convergent models that are interpretable.

First, we analysed the effect of cue condition on participant accuracy. The effect of cue condition did not improve model fit ($\chi^2(3) = 4.778$, $p = .188$), indicating that different cue types did not result in greater accuracy over and above distributional information.

Next, we added the effect of test block. This did not improve model fit ($\chi^2(1) = 0.071$, $p = .789$), indicating that performance did not improve based on the amount of training participants had received; participants did not demonstrate learning.

Adding the effect of LoE failed to improve model fit ($\chi^2(1) = 0.114$, $p = .735$), indicating that participants were similarly accurate with both longer and shorter sequences.

Following this, we added the main effect of violation position, which resulted in a significant improvement in model fit ($\chi^2(2) = 187.44$, $p < .001$). This reflected the tendency for participant accuracy to be higher for grammatical sequences, and similar for violation sequences. The final model outcomes are presented in Table 1.

Further adding in the two-way interaction between block and cue condition did not improve model fit ($\chi^2(3) = 1.912$, $p = .591$), indicating that participants in the pause, pitch and combined cues conditions failed to improve more than the baseline condition over training.

Crucially though, including the two-way interaction between cue condition and violation position further improved model fit ($\chi^2(6) = 15.312$, $p = .018$), reflecting the tendency for participants to be more accurate at correctly judging non-final violation sequences in the pitch cues condition relative to the other conditions. Overall, participant accuracy did not differ from baseline on the basis of prosodic information, however, pitch cues facilitated the detection of grammatical violations for LoE 2 sequences in non-final positions (see Figure 3 for more detail).

Including the two-way interaction between violation position and sequence length improved model fit further ($\chi^2(1) = 10.438$, $p = .001$), indicating that participants more accurately judged grammatical sequences with two levels of embedding.

No further interactions significantly improved model fit (*p*s > .05)

The significantly increased performance for grammatical sequences suggests that participants exhibited response bias. As a result, we proceeded to conduct an analysis employing signal detection theory, to better assess participants' sensitivity to the grammatical structure, and to quantify the extent to which participant performance reflects response bias.

Table 1

*Accuracy Final Model Outcomes*

| Fixed Effect | Estimate | Standard Err. | *z* | *p* |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Intercept | -0.187 | 0.118 | -1.588 | .112 |
| 5th Position Violation | -0.228 | 0.206 | -1.084 | .279 |
| Grammatical | 0.507 | 0.150 | 3.372 | .001** |
| Cue – Combined | -0.152 | 0.153 | -0.998 | .318 |
| Cue – Pause | -0.046 | 0.142 | -0.326 | .745 |
| Cue – Pitch | -0.075 | 0.135 | -0.555 | .579 |
| LoE | -0.136 | 0.090 | -1.515 | .13 |
| 5th Position Violation: Cue - Combined | 0.239 | 0.251 | 0.951 | .342 |
| Grammatical: Cue - Combined | 0.036 | 0.197 | 0.182 | .855 |
| 5th Position Violation: Cue - Pause | 0.208 | 0.248 | 0.838 | .402 |
| Grammatical: Cue – Pause | 0.0411 | 0.187 | 0.220 | .826 |
| 5th Position Violation: Cue – Pitch | 0.507 | 0.236 | 2.145 | .032* |
| Grammatical: Cue – Pitch | -0.06 | 0.181 | -0.331 | .741 |
| Grammatical: LoE | 0.297 | 0.126 | 2.365 | .018* |

Model Syntax: acc ~ (1 + ViolationPosition + Cue + SequenceLength + ViolationPosition:Cue + ViolationPosition:SequenceLength|Participant) + (1 + ViolationPosition + Cue + SequenceLength + ViolationPosition:Cue + ViolationPosition:SequenceLength|Item) + + ViolationPosition + Cue + SequenceLength + ViolationPosition:Cue + ViolationPosition:SequenceLength. This model analysed accuracy data for trials of both LoEs ($N_{trials}$ = 12288).

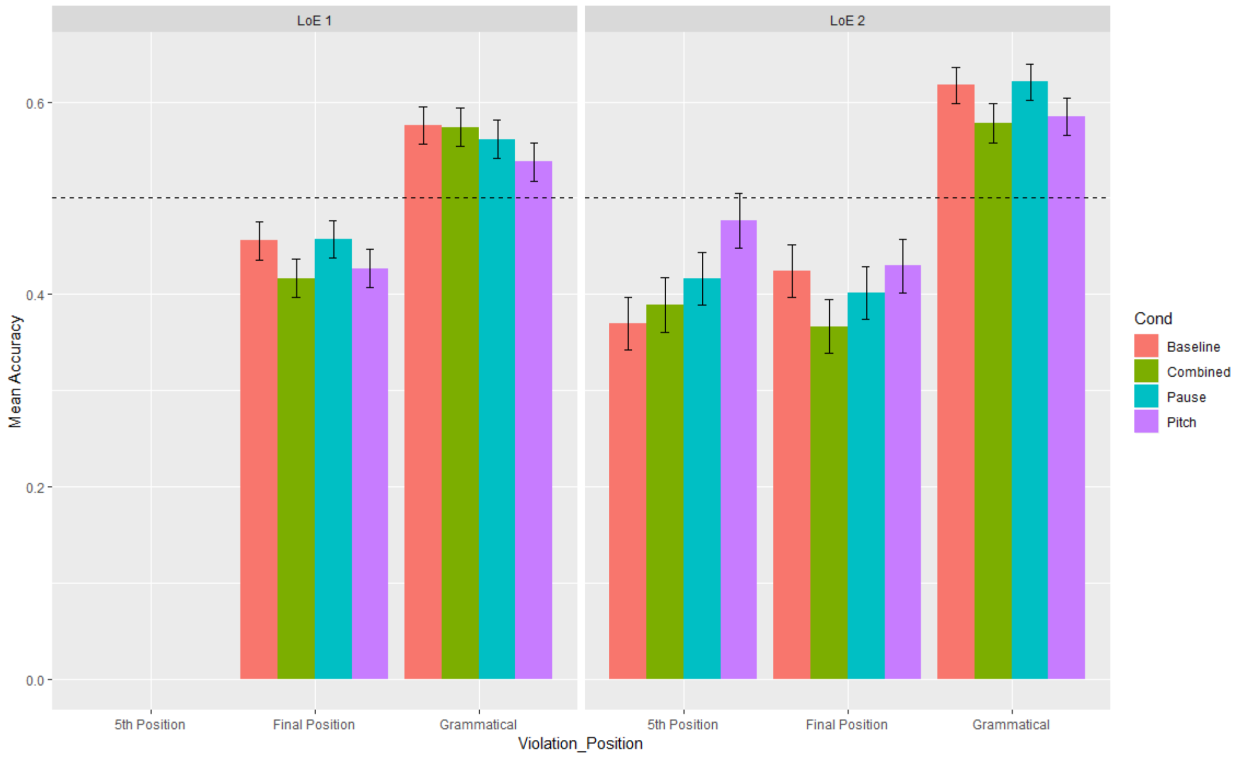


Fig. 4. Mean response accuracy by error position, broken down by cue condition. The left panel illustrates the accuracy data for 1 LoE sequences (here, errors could only occur in the sequence final position). The right panel illustrates the accuracy data for 2 LoE sequences. Black bars indicate the standard error of the mean (SEM). The horizontal dashed line indicates chance performance.

### *3.2 Signal Detection Theory: Sensitivity analysis*

Signal detection theory (SDT) can be employed whenever two possible stimulus types must be discriminated (Stanislaw & Todorov, 1999), in the present case, grammatical (signal trials) and ungrammatical (noise trials) stimuli. According to SDT, participants respond on the basis of the decision variable during each trial. If the decision variable is sufficiently high, the subject responds yes (grammatical), or no (ungrammatical). Correctly classifying a grammatical stimulus as grammatical is a hit, however, falsely classifying an ungrammatical stimulus as grammatical is termed a false alarm. SDT argues that participants' decision variable will be affected by prior input, therefore, the decision variable will elicit a distribution of values across grammatical and ungrammatical trials. The hit rate is the proportion of the signal distribution that exceeds the criterion, and false alarm rate is the proportion of noise distribution that exceeds the criterion. Using the hit and false alarm rate allows researchers to derive two aspects of participants' performance; their *sensitivity* to the signal, and their *response bias*. In the present paper, we employed the non-parametric measures of sensitivity A', and the response bias of A', *b*, given by equations (1) and (2) below, in accordance with Zhang and Mueller's (2005) correction:

$$(1)\ A' = \begin{cases} \frac{3}{4} + \frac{H-F}{4} - F(1-H) & if\ F\ \leq 0.5\ \leq H\,; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{F}{4H} & if\ F\ \leq H\ < 0.5\,; \\ \frac{3}{4} + \frac{H-F}{4} - \frac{1-H}{4(1-F)} & if\ 0.5 < F\ \leq H\,. \end{cases}$$

$$(2)\ b = \begin{cases} \frac{5-4H}{1+4F} & if\ F\ \leq 0.5\ \leq H\,; \\ \frac{H^2+H}{H^2+F} & if\ F\ < H\ < 0.5\,; \\ \frac{(1-F)^2+(1-H)}{(1-F)^2+(1-F)} & if\ 0.5 < F < H. \end{cases}$$

In the present study, we computed A' and b for each participant per block and LoE. A' values of 0.5 are taken to mean that participants are unable to distinguish signal from noise, while b values of 1 indicate no response bias, with values greater than 1 indicating a bias towards no responses, while those less than 1 indicating towards yes responses.

To formally assess participants' sensitivity, we conducted a series of generalized linear mixed-effects models (LMER) predicting the dependent variable of A'. As both A' and b are computed using all trials within a block (1 – 12), we were unable to include by-items random intercepts and slopes in this analysis. By-subjects intercepts and slopes are retained in these models. Models were built up iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). Fixed effects were retained in the final model if they resulted in significantly improved model fit in isolation, or within an interaction. Interactions were retained in the model if they significantly improved model fit. As fixed effects, we tested the effect of the cues participants were exposed to (baseline, pitch, pause, and combined), testing block (1 – 12), and LoE (LoE 1 or LoE 2), and interactions between cue condition, block, and LoE.

Adding the effect of cue condition to the baseline model did not improve model fit ($\chi^2(3) = 5.888$, $p = .117$), suggesting that in isolation, cue condition did not produce differences in participants' sensitivity to the grammatical signal.

The addition of block to the model also failed to improve model fit ($\chi^2(1) = 0.4815$, $p = .488$), suggesting that increased exposure to the grammar did not improve sensitivity.

Adding the effect of LoE also failed to improve model fit ($\chi^2(1) = 0.017$, $p = .897$), suggesting that sensitivity did not differ between LoEs.

Subsequently adding the two-way interaction between cue condition and block failed to improve model fit ($\chi^2(3) = 1.478$, $p = .687$); additional training did not affect the sensitivity of participants differently across conditions.

Adding the two-way interaction between cue condition and LoE, however, significantly improved model fit ($\chi^2(3) = 8.332$, $p = .040$); participants in the pitch cues condition showed a greater difference in sensitivity than baseline between LoE 1 and LoE 2 sequences, with greater sensitivity for the LoE 2 sequences (see figure 5 for greater detail).

Adding the two-way interaction between block and LoE failed to improve model fit ($\chi^2(1) = 0.013$, $p = .911$), suggesting that increased training did not result in increased sensitivity for particular LoEs.

Finally, including the three-way interaction between cue condition, block, and LoE did not produce improved model fit ($\chi^2(4) = 2.964$, $p = .563$).

**Table 2**

*Sensitivity Final Model Outcomes*

| Fixed Effect | Estimate | Standard Err. | *p* |
|---|---|---|---|
| Intercept | 0.398 | 0.027 | 14.836*** |
| Cue – Combined | -0.052 | 0.039 | -1.347 |
| Cue – Pause | -0.017 | 0.038 | -0.456 |
| Cue – Pitch | -0.073 | 0.038 | -1.921 |
| LoE | -0.047 | 0.037 | -1.269 |
| Cue – Combined: LoE | -0.004 | 0.037 | -0.067 |
| Cue – Pause: LoE | 0.049 | 0.052 | 0.933 |
| Cue – Pitch: LoE | 0.130 | 0.052 | 2.477* |

Model Syntax: A’ ~ (1 + Cue*LoE|Subjec) + Cue + LoE + Cue:LoE

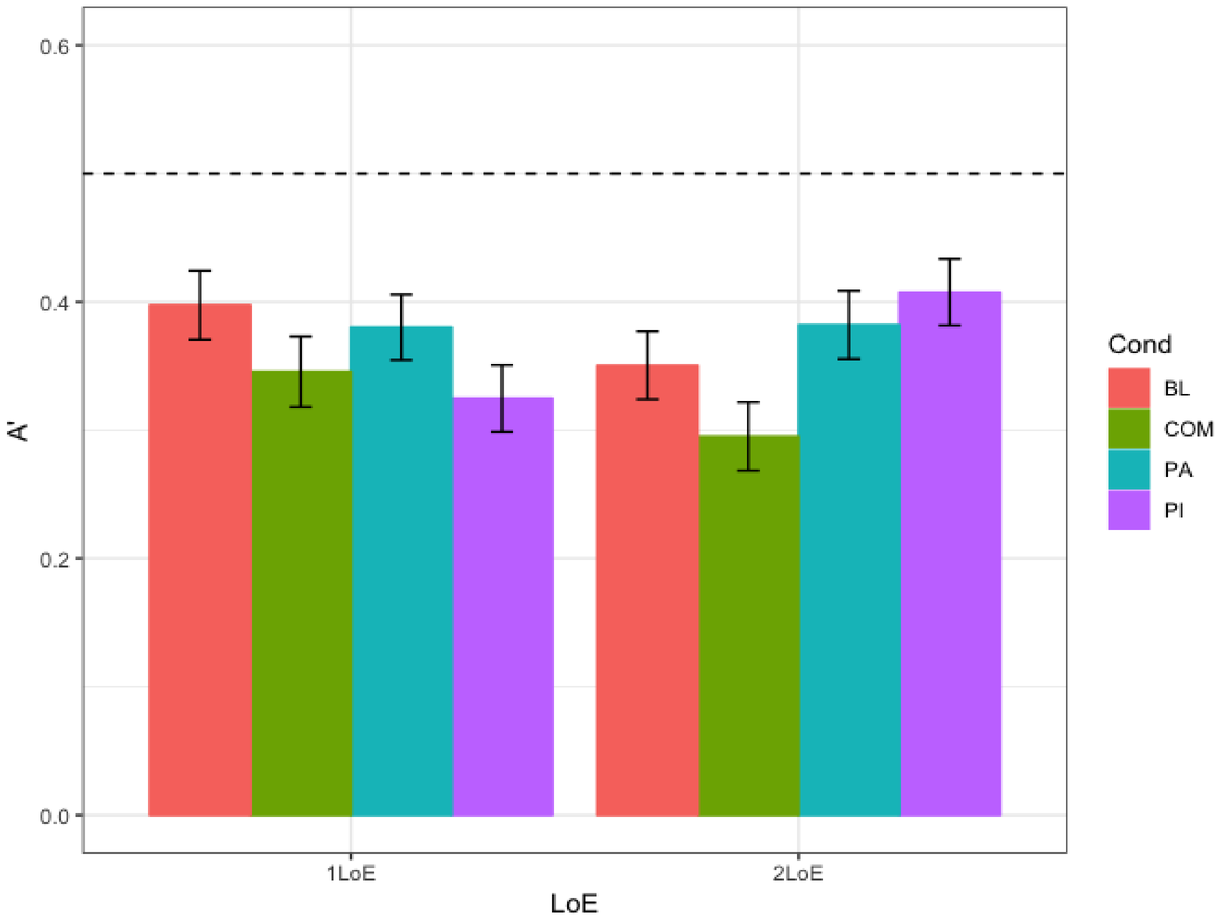


**Fig. 5**. Mean A’ value by Condition and LoE. Error bars indicate the SEM, and the horizontal dashed line indicates the threshold at which participants are unable to distinguish signal and noise, indicating they were unable to detect the underlying signal. Values above .5 indicate the ability to distinguish signal from noise.

### *3.3 Signal Detection Theory: Response bias*

To formally assess participants response bias, we conducted a series of generalized linear mixed-effects models (LMER) predicting the dependent variable of b. By-subjects intercepts and slopes are retained in these models. Models were built up iteratively, adding in fixed effects and interactions sequentially, and performing likelihood ratio tests after the addition of each new fixed effect term and interaction (following Barr, Levy, Scheepers, & Tily, 2012). Fixed effects were retained in the final model if they resulted in significantly improved model fit in isolation, or within an interaction. Interactions were retained in the model if they significantly improved model fit. As fixed effects, we tested the effect of the cues participants were exposed to (baseline, phonological, pitch, pause, and combined), testing block (1 – 12), and LoE (LoE 1 or LoE 2), and interactions between cue condition, block, and LoE.

Adding the effect of cue condition to the baseline model did not improve model fit ($\chi^2(3) = 3.517$, $p = .319$), suggesting that response bias was equal across all cue conditions.

Adding the effect of block did not improve model fit ($\chi^2(1) = 0.142$, $p = .706$), suggesting that response bias did not change significantly across training.

The addition of LoE, however, significantly improved model fit ($\chi^2(1) = 5.340$, $p = .020$); response bias was greater for LoE 2 structures.

Adding the two-way interaction between cue condition and block did not improve model fit ($\chi^2(3) = 2.878$, $p = .411$); response bias in each condition was not affected by increased exposure to the experimental grammar.

Including the two-way interaction between cue condition and LoE significantly improved model fit ($\chi^2(3) = 8.833$, $p = .032$); response bias differed between conditions. Overall, participants were biased towards yes responses in all conditions, with a greater response bias for LoE 2 sequences. Notably, however, there was a significant reduction in response bias for LoE 2, relative to LoE 1 sequences in the pitch cues condition (see figure 6 for greater detail), similar to the effects found for participants' sensitivity above.

Including the two-way interaction between block and LoE failed to improve model fit ($\chi^2(1) = 0.848$, $p$ = .357), indicating that increased exposure to the grammar did not differently affect response bias at each LoE.

Finally, including the three-way interaction between cue condition, block, and LoE did not improve model fit ($\chi^2(4) = 3.551$, $p$ = .470).

Table 3

*Sensitivity Final Model Outcomes*

| Fixed Effect | Estimate | Standard Err. | *p* |
|---|---|---|---|
| Intercept | 0.605 | 0.057 | 10.536*** |
| Cue – Combined | -0.091 | 0.083 | -1.097 |
| Cue – Pause | 0.036 | 0.081 | 0.438 |
| Cue – Pitch | -0.126 | 0.081 | -1.557 |
| LoE | -0.166 | 0.079 | -2.254 |
| Cue – Combined: LoE | 0.037 | 0.106 | 0.349 |
| Cue – Pause: LoE | 0.011 | 0.104 | 0.105 |
| Cue – Pitch: LoE | 0.268 | 0.104 | 2.564* |

Model Syntax: Response Bias ~ (1 + Cue*LoE|Subject) + Cue + LoE + Cue:LoE

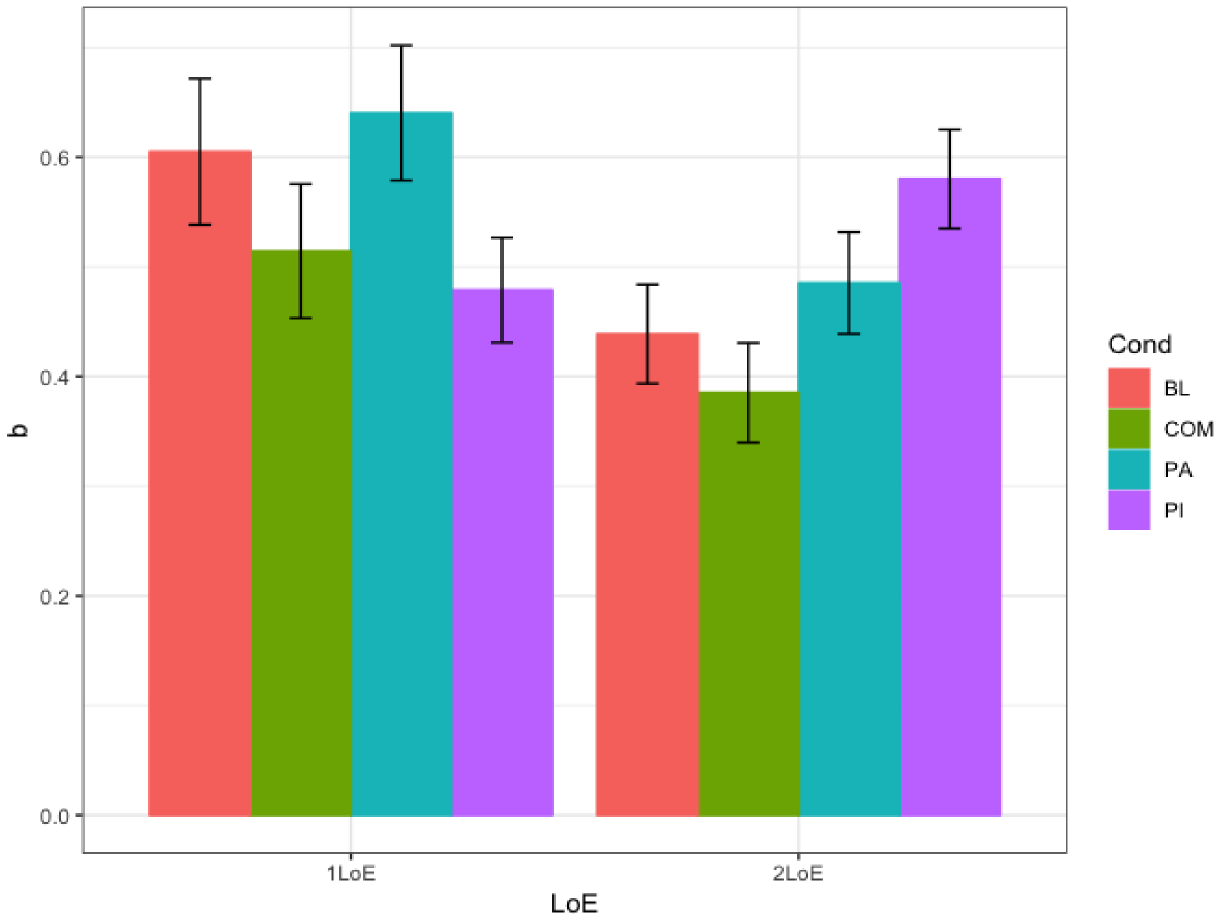


**Fig. 6**. Mean response bias by condition and LoE. Error bars display the standard error of the mean. b values closer to 0 indicate a greater bias towards yes responses.

## 4. Discussion

The main aim of this study was to assess whether acoustic cues modelled on the speech production data of our participants' native language would facilitate the acquisition of hierarchically centre-embedded structures. Frank et al., (2012) argue that the processing of speech in real-time may be sequential, with individuals relying on superficial surface level cues to form an initial parse of a sentence, and subsequently assigning a syntactic structure based on this parse.

Natural speech contains a rich set of prosodic cues from which phrasal groupings may be computed, specifically, pitch similarity and temporal proximity (Trotter, Frost, & Monaghan, Chapter 3). Respectively, these state that groupings are more likely to be made between tones when they are more similar in pitch, and that tones that occur closer together in time are more likely to be grouped together. We implemented pause and pitch cues consistent with these findings in the current AGL study and assessed their individual effects relative to two comparison conditions; baseline (where there were no cues, only frequency of co-occurrence) and combined cues (where both cue types were present). We predicted that; relative to baseline, each cue will improve learning; judgement accuracy will improve with more training; and that participants will be less accurate with longer sequence lengths, due to their increased complexity relative to shorter sequences.

The results of the behavioural study did not support these predictions. No individual or combined cue condition resulted in better performance than baseline, and notably, performance was close to chance in all conditions. Participants' judgement accuracy did not increase over blocks, suggesting participants were unable to acquire the experimental grammar. LoE did not affect participant performance, suggesting that LoE 1 and LoE 2 sequences are similarly difficult to acquire. Regardless of cue condition, participants were more accurate at correctly classifying grammatical than ungrammatical sequences, suggesting an increased likelihood of participants endorsing any sequence as grammatical. There were significant interactions in the model that suggest cue condition did facilitate grammaticality judgements for violations highlighted by the tonal and temporal structure participants were exposed to. Violations in the 5th position of LoE 2 sequences were more likely to be correctly judged in the pitch cues condition.

We also conducted an analysis based on Signal Detection Theory (SDT; Stanislaw & Todorov, 1999) to assess the extent to which participant performance could be attributed to response bias, and how sensitive they were to the underlying grammatical structure. The sensitivity analysis revealed that participants were unable to distinguish the grammatical structure in any of the conditions. Only the two-way interaction between cue condition and LoE reached significance, revealing that participants in the pitch cues condition were significantly more sensitive to the grammatical structure in LoE 2, relative to LoE 1 structures. The response bias analysis largely agreed with our interpretation of the accuracy data; across all conditions, participants were biased towards judging any sequence as grammatical across all conditions, and were more biased with LoE 2 sequences. A significant two-way interaction with cue condition, intriguingly revealed that participants in the pitch cues condition were less biased with LoE 2 structures, similar to the sensitivity results. Overall, whilst the results suggest participants were not sensitive to the underlying grammatical structures, the suggest an intriguing role for pitch cues.

How do we account for this pattern of results? First, let us consider the null effect of cue condition. In isolation, no cue type produced greater than baseline accuracy, replicating the results of Trotter, Frost and Monaghan (Chapter 4), suggesting that pitch similarity and temporal proximity cues do not largely affect grammaticality judgements in artificial language tasks.

We failed to replicate the finding that pitch similarity cues resulted in higher than baseline accuracy at both LoEs for grammatical structures (Trotter, Frost and Monaghan, Chapter 4). In this prior study, we suggested that the inclusion of pitch cues over both training and testing resulted in tension between global, acoustic structure, and local, linguistic structure. Due to the salience of pitch cues, participants became overly reliant upon them, biasing responses to each test structure as grammatical, increasing judgement accuracy for grammatical structures, and lower than chance accuracy for ungrammatical structures, which was also supported by a response bias analysis. In contrast, in the present study, pitch cues resulted in greater than baseline accuracy when classifying grammatical violations in the fifth sequence position, greater sensitivity and lower response bias for LoE 2 structures. What aspect of the pitch condition could account for these findings?

In two LoE sequences, errors could appear in the fifth ($A_1A_2A_3B_3\underline{B_4}B_1$), or sixth sequence positions ($A_1A_2A_3B_3B_2\underline{B_4}$). In the latter case, the error is sequence final; recency effects suggest that this error type should be more salient and easily detected. While this seems intuitive, dependent syllables in the sequence initial and final positions occur at the highest and lowest pitch level, respectively, to respect the descending pitch declination over sentences in natural language (e.g. Mueller et al., 2010); pitch is not salient for these dependencies. On the other hand, items in the fifth position have the same pitch as those occurring in the second, increasing their salience with pitch similarity. It may be the case that where pitch similarity is highest, it is easier to detect grammatical violations.

Unfortunately, we did not include violations in the deepest LoE for either sequence length – the third positions for one LoE sequences, and the fourth for two LoE sequences – which would allow us to formally assess this claim. We did not include these adjacent violations, as it would have introduced a test confound; adjacent violations would affect the transitional probabilities of adjacent pairs acquired during training. As a result, our measure of learning would then additionally (or perhaps primarily) reflect frequency of co-occurrence, instead of sensitivity to long-distance dependencies in the structure of the language. Within the scope of the present study, we suggest that pitch cues are useful for acquisition. While pitch cues may not helped participants to find the underlying grammatical structure, in LoE 2 structures, it did help to reduce response bias, and elicited gains in sensitivity, hence it was clearly salient.

Across all three analyses, pause cues did not significantly affect performance. In comparison to Trotter, Frost and Monaghan (Chapter 4), pause cues produced higher accuracy for grammatical sequences which was not apparent here, though this was not apparent in the sensitivity and response bias analysis. The different pattern of results may be attributable to the reduced pause duration in this study; the pauses may have been less explicit, or simply less salient. Alternatively, given that the results of Trotter, Frost and Monaghan (Chapter 4) were not conclusive, that English native speakers may be somewhat insensitive to pitch cues. This assumption received some support; Seidl (2007) found that 6-month-old English acquiring infants were sensitive to prosodic boundaries in the absence of pause cues (experiment 2), however, they were insensitive to boundaries when pitch cues were removed (experiment 3). In contrast, Männel and Friederici (2009) found that German acquiring infants require prosodic boundaries at prosodic boundaries to elicit the closure positive shift – an event-related potential that is reliably evoked at the close of a prosodic phrase. The insensitivity of English speakers to pause cues may therefore reflect language-specific factors: German has a larger number of inflections and a flexible word order, suggesting that the functional demands on pitch may be greater for English speakers in highlighting phrase structure (Männel and Friederici, 2009). This is however troubling both for a Gestalt processing account, and does not fully align with experimental observations in English.

Notably, Kraljic and Brennan (2005) and Snedeker and Trueswell (2003) found that participants designated as instructors reliably produced durational prosodic cues (pauses, final lengthening) to their partner who was required to perform actions based on these accounts, and that these cues supported listeners' performance. In Kraljic and Brennan's (2005) study, these cues were reliably produced whether or not there was an intended audience. As such, we can raise the question of why these durational cues should not affect performance here. In each of these studies, utterances varied in duration. In natural language, pauses not only reflect planning, but also physical constraints on the vocal system, notably, participants requiring to breath. Therefore, pause duration will reflect the overall duration of the phrase. In the present study (and Trotter, Frost, and Monaghan, Chapter 4), this is clearly not the case, with extended pauses occurring after each syllable. Therefore, experience could lead these pauses to be statistically related to disfluency and to be ignored. This study has no available method of testing this assumption, however. In future work, we would recommend that pause durations are also computed as reflecting the overall duration of the phrase to account for this. Notably, however, this directly conflicts with the results of Mueller, Bahlmann, and Friederici (2010), who found that inter-syllabic pauses positively affect acquisition of HCE structures, an effect this study failed to replicate.

There was a null effect of combined cues – where participants were exposed to both pause and pitch cues – both in isolation, and when moderated other factors. If both pause, and pitch cues increased grammaticality judgement accuracy for fifth sequence position violations, should combined cues not elicit the greatest accuracy? A reading consistent with the intersensory redundancy hypothesis (Bahrick et al., 2004), or an additive account (Yu & Ballard, 2007) suggests this should be the case. The theory of degeneracy (Monaghan, 2017), argues that a noisy cue environment best supports learning, preventing learners from becoming overly reliant on individual cues. In the combined condition, cues were present over training and testing and always co-occurred. Thus, in 25% of cases, both cue types are unreliable. Overlapping cues should result in a broad attentive focus; if participants uniquely relied on the salient pitch cues and received a high amount of corrective feedback, they should shift their focus onto pause cues, and vice versa. Given that participants were less able to use pause cues for detection of fifth sequence position violations, shifting between cue types might result in performance no different from baseline.

Degeneracy may then provide an account for this pattern of results. However, to explicitly test this hypothesis would require a replication of this study with variable rates of cue reliability, and probabilistic cues and is thus beyond the scope of this paper.

In summary, in the present study, we aimed to assess the claim that including pitch similarity and temporal proximity cues based on participants' native language experience would facilitate acquisition of an artificial language. This was based on the idea that if hierarchical sentences can be processed sequentially, then individuals may compute dependencies based on low-level perceptual biases. The results from this study suggested a particularly salient role for pitch cues, based on the relative benefit it produced in accuracy, sensitivity and response bias for LoE 2 structures. Thus, we suggest that individuals can use pitch similarity to support phrasal grouping of hierarchically structured speech. To verify these effects, however, will require additional replications, potentially in a higher-powered study. Additionally, implementing these cues in different test paradigms using online measures (e.g. eye-tracking) would allow future studies to assess the influence of these cues on processing.

# Chapter 6: Gaze behaviour in the visual world suggests auditory-perceptual Gestalts facilitate the comprehension of hierarchical structure

Antony S. Trotter[1] & Padraic Monaghan[1,2, 3]

1. Department of Psychology, Lancaster University, UK
2. Department of Linguistics, University of Amsterdam, NL
3. Max Planck Institute for Psycholinguistics, Nijmegan, NL

The results of Chapter 2 raised the question of to what extent the results of Chapters 4 and 5 could be attributed to the use of a reflection-based task. Chapter 6 thus represents an important extension of these studies; the use of processing-based measures, and including real linguistic content. This paper is presented as a draft ready for submission.

## Statement of Author Contribution

In the Chapter entitled, “Gaze behaviour in the visual world suggests auditory-perceptual Gestalts facilitate the comprehension of hierarchical structure”, the authors agree to the following contributions:

Antony S. Trotter – 80% (Writing, experimental conceptualisation and design, data collection, and analysis)

Signed: ______________________________ Data: ____________

Professor Padraic Monaghan – 20% (Experimental conceptualisation and design, analysis, and review)

Signed: P Mon______________________________ Data: 21/2/19

# Abstract

Speech comprehension relies upon rapidly being able to process its dependencies. Recent proposals suggest this is driven by sequential processing (Frank, Bod, & Christiansen, 2012), with low-level statistical correspondences supporting dependency detection. The Gestalt principles of temporal proximity and pitch similarity are particularly relevant to the comprehension of phrasal clauses. The former states listeners will group sequential words if they occur closer together in time. The latter states that listeners will group sequential words if they are similar in pitch. Trotter, Frost, and Monaghan (Chapter 3) found that for spontaneously produced hierarchical centre-embedded structures (HCEs), phrases within the embedded clause are similar in pitch, and are preceded by a lengthened pause. Passives differ; a longer pause and pitch reduction occurs after the verb phrase of the embedded clause. These results suggest that in speech, temporal proximity and pitch similarity provide grouping cues for tracking dependencies in HCEs. This study assesses if these cues are useful in comprehension.

Using the visual world paradigm, we analysed participants' (n = 64) gaze behaviour in response to active and passive relative clauses, whilst they viewed scenes containing four potential targets. Prosodic structure was manipulated to be congruent with active or passive cues in Trotter, Frost, and Monaghan (Chapter 3), or two control structures. Pitch similarity results indicated that - regardless of form - cues supporting the grouping of the embedded clause facilitate processing. Temporal proximity cues consistent with syntactic structure did not facilitate processing, instead results suggested a general benefit of increased processing time.

## 1. Introduction

The hierarchical structure of language has remained a key focus of psycholinguistics. Chomsky (1957; 1959) argued that due to the presence of hierarchical dependencies in language, the human language processor must minimally conform to the phrase-structure grammar level of the Chomsky hierarchy. The Chomsky hierarchy embeds rule systems capable of generating an infinite set of sequences by defining increasing constraints on possible structures. The weakest level of the hierarchy is the finite-state grammar, which can be fully specified by transitional probabilities between a finite number of states (Hauser & Fitch, 2004). To process a finite-state grammar sequence requires only a large enough memory stack to hold sequential states – and the transitions between them – to concatenate them into longer sequences. Phrase structure grammars lie at the next level of the hierarchy. Similarly, they can concatenate items, but can additionally embed strings within other strings, resulting in complex phrase structures, and long-distance dependencies. The key focus of research into phrase-structure grammars has been the hierarchical centre-embedded structure. The processing mechanisms necessary to generate and process these complex structures are more sophisticated, requiring an open-ended memory system, in addition to the perceptual mechanisms to recognise them (Hauser & Fitch, 2004). This has led to a fruitful research tradition investigating the processing differences between sequences generated by these two grammars, providing evidence about the mechanisms of human language processing, and the perceptual mechanisms that support the processing of phrase structure.

Hierarchical centre-embeddings (HCEs) with more than three levels of embeddings are challenging to process, even for expert speakers (e.g. Bach, Brown, & Marslen-Wilson, 1986; Hudson, 1996; Newmeyer, 1988). Consider this example from the British Road Traffic Act (1972), “A person [$_1$ who, [$_2$ when riding a cycle, [$_3$ not being a motor vehicle,] on a road or other public place,] is unfit to ride through drink or drugs,] shall be guilty of an offence.” This is a typical example of a three LoE construct. It illustrates that as more clauses are inserted, relating dependent elements becomes more difficult. As more clauses are embedded, the distance between syntactically dependent elements increases. As dependency lengths increase, the difficulty of associating the related constituents increases (Lai & Poletiek, 2011). This suggests that human’s ability to process sequences generated by a PSG is limited.

Given the difficulties associated with processing hierarchical centre-embeddings, it is important to question how they are processed. Frank, Bod, and Christiansen (2012) have proposed that the comprehension of hierarchical structures is achieved through sequential processing, though production may be drive by hierarchical processing. This argument is supported by recent computational work. Frank and Bod (2011) compared how well word-probability estimates generated by three kinds of probabilistic language models relating to different psychological mechanisms and representations accounted for the reading time measurements (based on the eye-tracking data of 10 participants) of the Dundee corpus (2368 sentences). The first class of model was a PSG model, induced from syntactic trees, and utilised hierarchical structure. The second – Markov models – and third – Echo State Networks – classes only had access to sequential structure. The results indicated that the PSG model failed to estimate variance in reading time data over and above each of the sequential-structure models, suggesting that a sentence’s hierarchical structure - unlike other sources of information - did not noticeably affect the generation of expectations about upcoming words.

Under the sequential processing theory, to comprehend a hierarchical structure in a rapidly unfolding temporal context, the listener relies upon superficial, surface level cues to parse its dependencies, rather than processing the upcoming words in a hierarchy. Returning to our example from the British Road Traffic Act (1972), world knowledge can be used to determine the relationships between related elements. Bicycles, as inanimate objects, are unable to consume drink or drugs, ride themselves, or be guilty of an offence. Humans, however, can ride bicycles in various locations, imbibe drink and drugs, feel their effects, and be found guilty of criminal offences. This suggests the basic units "person riding bike", "(if) person is drunk", "(if) on road", "(then) person is guilty". Thus, by employing world knowledge, the listener can determine the dependency structure of the incoming sentence without have to explicitly process the words in a hierarchy.

Semantic cues are not the only surface level cue available for relating dependent elements; human speech is rich with prosodic cues that may trigger auditory processing biases that provide information to clausal structure. There are two biases in auditory processing that are particularly relevant to the present study; pitch similarity, and temporal proximity. The former states that individuals tend to group together sounds that are close in pitch, and to distinguish between those that are further apart, while the latter states that if two sounds are temporally distant, you are unlikely to create a link between them (Deutsch, 2013). Trotter, Frost, and Monaghan (Chapter 3) conducted a speech analysis of a small corpus of spontaneously produced active-object ("[The bear] [the girl] [is hugging] [is brown]", see figure 1 for a formal syntactic representation) and passive relative clauses ("[The bear] [being hugged] [by the girl] [is brown]"). We hypothesized that pitch similarity would be highest between syntactically dependent phrases, and that pauses occurring between clauses would be longer than elsewhere in the speech. For active-object structures, phrases in the embedded clause ("[the girl] [is hugging]") were spoken in a more similar pitch and contrasted in pitch with the first phrase of the external clause. Pauses also tended to be longer preceding the embedded clause (i.e., before "[the girl]"), however, their duration was highly variable. Thus, for active-object relative clauses, pitch similarity and temporal proximity potentially provide grouping information consistent with syntactic structure. In terms of sequential processing, prosodic information may therefore support rapidly forming an initial parse by providing reliable grouping information. For passives, however, results indicated that pitch similarity was highest between the first two phrases ("[The bear] [being hugged]"), and the longest pause occurred preceding the next phrase ("[by the girl]"). This result would be consistent with a "good enough" processing account (Ferreira, 2003); by this point in the sentence, enough information had been provided to find the referent of the sentence, and thus the first two phrases are tonally and temporally grouped together.

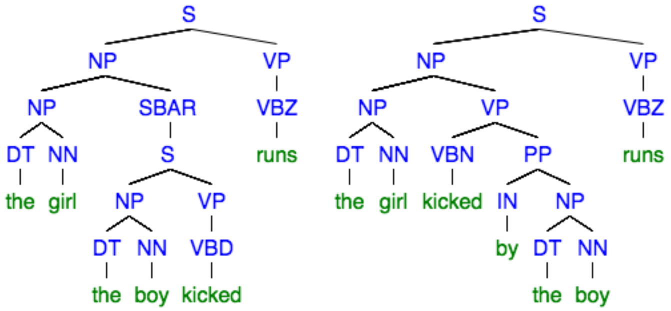


**Fig. 1.** Syntactic trees for reduced active-object (right) and passive (left) relative clauses.

The results for active-object relative clauses suggest that prosody may provide a means through which individuals could rapidly determine the dependencies of incoming speech, consistent with the assumptions of sequential processing accounts. However, finding the presence of these cues in a speech corpus may be merely an artefact of speech production demands. Their presence in speech alone does not prove that they are useful for comprehension, sequential or otherwise.

To determine if a cue facilitates processing, it is necessary to implement it in a controlled experimental setting. A flexible method for investigating the processing of linguistic structure is the artificial grammar learning (AGL) paradigm. In this paradigm, the researcher constructs artificial language fragments, and examines participants' learning of sequences composed of these fragments. These studies contain a training and test phase. During training, participants are presented with sequences that, unbeknownst to participants, adhere to either a PSG or FSG. In the testing phase, participants are exposed to novel sequences generated by the grammar - that are either grammatical or ungrammatical - and perform a classification task. Exposure to the distributional statistics of the language during training allows participants to acquire the rules of the experimental grammar. Using a between-groups design, AGL studies can implement different cues to assess which cues improve grammatical learning over conditions where only the distributional statistics of a language are available to learners.

Trotter, Frost and Monaghan (Chapter 4) implemented three different cue types; pitch similarity, temporal proximity cues, and pitch similarity cues based on spoken German, and prior work investigating clausal segmentation in infants (Hawthorne & Gerken, 2014). The results indicated that although there was learning overall, no cue type improved learning over baseline. Participants were more accurate at identifying grammatical structures than the baseline condition, resulting in above chance performance. Participants trained with pause cues were also more accurate with grammatical structures, however, only for sequences with two (as opposed to one) levels of embedding. Across both conditions, performance was qualitatively worse on ungrammatical sequences, though this contrast was insignificant. As the acoustic cues were present over both training and testing, this suggested that they globally increased the plausibility of accepting all structures, leading to higher performance on grammatical structures. This was confirmed by a subsequent analysis of sensitivity and response bias. This may have reflected a tension between global acoustic cues, and local linguistic cues. In natural language, this conflict may be resolved by prosodic variation over multiple syntactic structures (e.g. Trotter, Frost, & Monaghan, Chapter 4), preventing a specific prosodic structure from being seen as a reliable grammatical cue. Basing the prosodic information on corpus data of native German speakers (Fery & Schubö, 2010) raises the question of whether it failed to facilitate the English-speaking sample's performance.

To address whether basing the acoustic cue conditions on participants' native language experience would improve grammaticality judgement performance, Trotter, Monaghan, and Frost (Chapter 5) replicated this study, implementing temporal proximity and pitch similarity cues consistent with an English-speaking corpus (Trotter, Frost, & Monaghan, Chapter 3). Accordingly, temporal proximity cues were shortened (111 vs 175 ms), reducing their salience, pitch similarity cues had more variance, increasing salience, however, phonological similarity cues were not included here. An analysis of sensitivity and response bias suggested that exposure to pitch similarity cues improved participants' sensitivity and reduced response bias for sequences including two levels of embedding, reflecting higher classification accuracy for grammatical violations in the fifth sequence position (e.g. $A_1A_2A_3B_3\underline{B_5}B_6$). Why should this be the case? In sequences including two levels of embedding, violations could occur in the fifth or sixth sequence position. Dependent syllables occurring in first and final positions occurred in the highest and lowest pitch, respecting the descending pitch contour of spoken language (see Mueller, Bahlmann, & Friederici, 2010). In contrast, syllables in the fifth and second positions occur at the same pitch; the salience of this dependency is boosted by pitch similarity. As participants' grammaticality judgements of sequences with violations in the final position did not improve over baseline, it suggests that participants were sensitive to pitch grouping cues. Whilst temporal proximity did improve accuracy over blocks, participants were less able to effectively use temporal proximity to process hierarchical structure.

It is important to note, however, that the use of the AGL paradigm has notable drawbacks. Primarily, it is an explicit, reflection-based measure; learning is assessed with the ability to correctly classify novel sequences, with each assessment conducted after learning blocks, and each response following sequence exposure. These measures can be contrasted with implicit, processing-based mechanisms, such as looking times. The two task types may tap into different mechanisms; processing-based tasks appear more effective for assessing processing-based learning, such as the online processing of speech, or acquisition of grammatical structure (Christiansen, 2018; Frizell, O'Neill, & Bishop, 2017; Isbilen et al., 2018). To provide a complete picture of the role of auditory-perceptual Gestalt grouping cues in speech perception, it is necessary to assess their effect in processing-based tasks.

A useful processing-based task for assessing spoken language is the visual world paradigm. Cooper (1974) developed the framework for what is now known as the visual world paradigm. In this study, participants viewed scenes whilst listening to short narratives. The results revealed that listeners' gaze was drawn to objects that were mentioned or were associated to the text. Importantly, participants' eye movements were tightly time-locked to the text; 90% of fixations to critical objects were triggered either while the corresponding word was spoken, or 200ms after word offset. The general set-up of a contemporary visual world comprehension paradigm is straightforward (Huettig, Rommers, & Meyer, 2011). On each trial, participants hear an utterance while viewing an experimental display, during which their eye movements are recorded. A popular version of the paradigm uses displays composed of line drawings, or semi-realistic scenes shown on a computer screen, and sentences that describe or comment upon the scene (e.g. “The boy will eat the cake”, Altmann & Kamide, 1999; “The uncle of the girl who will taste the beer is from France”, Kamide, 2012). Usually, the display contains the object mentioned in the utterance (the target), and distractor objects that are unmentioned. In another version of the paradigm, the displays are sets of objects laid out on a workspace (e.g. Snedeker & Trueswell, 2003), or shown as line drawings on a computer screen (e.g. Allopenna et al., 1998). Less commonly, displays comprise potential agents and patients mentioned in the utterance (e.g. Knoeferle & Crocker, 2007). Utilising semi-realistic scenes allows researchers to assess how listeners’ perception of the scene and their knowledge about the scenes and events affect their incremental understanding of the spoken utterances (Huettig et al., 2011). If displays of objects are used, the impact of world knowledge is disrupted, which renders them well suited for studying the activation of conceptual and lexical knowledge associated with individual words.

The visual world paradigm has successfully been employed to assess the role of prosodic cues on the comprehension of speech input. Dahan, Tanenhaus, and Chambers (2002) assessed whether pitch accenting can bias participants to look to a new item in a display, relative to when a word is deaccented. Accented words refer to a decrease followed by a rapid increase in pitch on the accented vowel, whereas de-accented refers to a simple, slower increase in pitch. In this study, participants were instructed to move an object around a display (e.g. “Put the candle below the triangle… Now put the candle above the square”, in a display including a candle, a candy, a square and a triangle). Listeners were able to use pitch cues predictively: When hearing words with an accented vowel, participants tended to look at a new item (the candle), whereas when hearing words with a deaccented vowel, listeners tended to look at the previously mentioned item (the candy).

Similarly, Watson, Tanenhaus, and Gunglogson (2008) demonstrated that pitch accenting can be used to identify contrast referents. Here, the question of interest was whether pitch accents on critical vowels of phonological competitors (e.g. camel/candle) would bias fixations towards the new (or contrastive) item, or the given item. Participants heard a series of commands (e.g. Click on the camel and the dog. Move the dog to the right of the square. Now, move the *camel/candle* below the triangle). In the final command the first vowel of the critical word would (underlined in the example), either had a sharp rise to the speaker's maximum pitch (H*) – assumed indicate an element is new to the discourse (*candle*) (Dahan, Tanenhaus, & Chambers, 2002) – or an initial pitch drop, followed by subsequent rapid increase in pitch, which has been argued to signal contrast between a previously given item relative to salient competitors (*camel*) (Pierrehumbert & Hirschberg, 1990). The results indicated that participants were rapidly able to use the pitch accent to direct their eye movements; when exposed to an L + H* accented vowel, fixations increased to contrast members (the candle), while decreasing to the new referent (the camel). For the H* accent, fixations increased to all potential referents with names consistent with the input, regardless of they were contrast or discourse new. The authors concluded that the domains of the pitch accents overlapped, with L + H* being specific, and H* being compatible with both new and contrast referents. In both studies, there is evidence for listeners being able to make use of pitch cues to bias visual attention.

Snedeker and Trueswell (2003) conducted a visual world study investigating the role of pauses and other duration cues in speech comprehension. This study differs in a critical way from the above; the visual world design here involved participant interaction. The speaker had to instruct a listener to perform an action on an array of objects in front of them. The array contained several items, however, the critical items were an animal holding an instrument (e.g. a frog holding a flower), and a corresponding separate animal and instrument (e.g. an empty-handed frog, and a large flower). Speakers were given a modifier (the experimenter picks up the flower and taps the frog) or an instrument (the experimenter touches the frog holding the flower) demonstration of the command, “Tap the frog with the flower”. The timing of speakers’ speech differed by which demonstration they received. Instrument speakers paused for a shorter duration following “Tap”, lengthened “frog”, and paused for a longer duration between “frog” and the with phrase. Modifier speakers paused for a longer duration following “tap”. The authors conducted a windowed analysis of listeners’ gaze behaviour in response to the commands, using two regions: 200-500ms following the onset of the direct object noun, and 200-800ms following the onset of the prepositional object. In the direct object noun window, instrument prosody resulted in participants looking equally to the frog holding the flower and the empty-handed frog; both were considered likely candidates. Modifier prosody, however, resulted in listeners mostly looking toward the frog holding the flower. In the prepositional object window, instrument prosody resulted in more looks to the flower. In the same time period, modifier prosody produced more looks to the frog with the flower. Taken together, the results suggest that participants can use pause (and other durational) cues to rapidly eliminate competitors during an unfolding utterance. However, it is notable that in this study, speakers who were unaware of the syntactic ambiguity did not produce the disambiguating prosodic cues.

In contrast to Snedeker and Trueswell (2003), Kraljic and Brennan (2005) found that speakers produced disambiguating temporal cues regardless of whether they were aware of the ambiguity or not. Here, a similar paradigm to Snedeker and Trueswell’s (2003) was used. However, it differed on several key dimensions; speakers were provided with diagrammatic instructions that could produce a goal (*put the dog in the basket on the star,* “in the basket” could be used to specify a particular dog) or a goal interpretation (to put the dog into the basket, and move the combined object to the star), and arrays of objects could be ambiguous (the display would contain both a frog holding a flower, and a separate frog and flower) or unambiguous (the display contained only a frog with a flower), and further, that both participants acted as speakers and matchers. Across three experiments they found that speakers produced syntactically driven temporal boundaries (taken as the duration of the noun phrase and subsequent pause), with a larger boundary following the first noun phrase for a goal interpretation, and shorter in the modifier condition. Using eye-tracking, it was found that prosodic cues rapidly biased participants' gaze to the correct object. Crucially, speakers produced the prosodic boundaries regardless of whether the display was ambiguous, whether the speaker was able to detect the ambiguity – notably contrasting Snedeker and Trueswell's (2003) findings - and whether the speaker had previously been a matcher, i.e. whether the speaker was aware of the matcher's needs. Taken together, these results suggest that prosodic cues are generated during the production process, likely driven by syntactic factors, and are not guided by audience design.

To test the influence of temporal proximity and pitch similarity on the structural processing of hierarchical structure, we sought to implement the cues found in the corpus analysis of Trotter, Frost and Monaghan (Chapter 3) in a VWP paradigm study. The study was split into two conditions, to test separately the role on comprehension of temporal proximity and pitch similarity. To assess the effect of each cue type, participants listened to reduced active-object and reduced-passive relative clauses, while viewing scenes comprising four separate interactions between agents and patients. For each syntactic form, participants were exposed to four different prosodic conditions; (1) active-congruent, (2) passive-congruent, (3) a no boundary control (where there were no pitch changes, or increased pause duration between phrases), and (4) a two boundary control (pitch changes/lengthened pauses occurred in both active and passive congruent locations). The gaze behaviour elicited by each syntactic-prosodic condition - and comparisons between them - will allow us to assess whether prosodic cues can facilitate the processing of hierarchical speech. If exposed to an active sentence, and there is no difference in the number of target fixations between active-congruent and passive-congruent prosody, it suggests that syntax-specific prosodic groupings provide no processing benefit. If neither of these differ from the high variance control – or there are more looks to the target in the high variance control – it would suggest that it is only the presence of prosodic variation (and not Gestalt grouping stratagems) that facilitates processing. Finally, if Gestalt grouping cues fail to result in more target fixations than the low variance control, then it suggests that the addition of prosodic grouping cues has no processing benefit.

## 2. Method

### *2.1 Participants*

64 self-reported native English speakers ($M_{Age}$ = 20.859, SD = 2.487, $n_{female}$ = 50) participated in the study (32 per condition), all of whom were students at Lancaster University. Participants received £6.50 or course credit for their participation.

### *2.2 Stimuli and design*

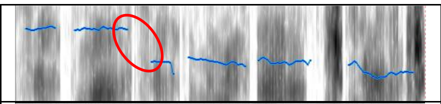
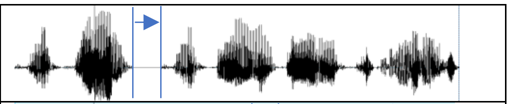
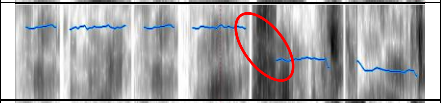
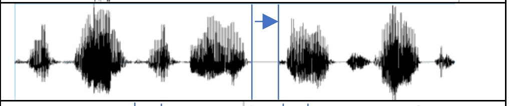
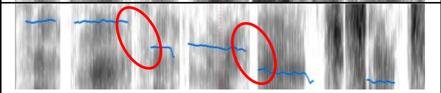
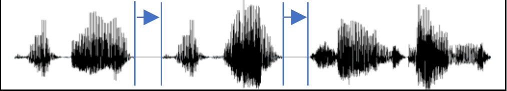
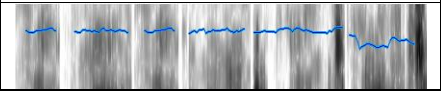
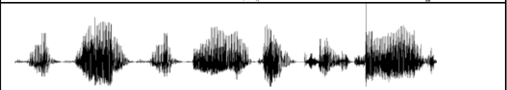
This experiment comprised two conditions. The first – pitch dynamics – assessed the role of pitch dynamics in the absence of temporal grouping cues. The second – temporal dynamics – assessed the role of temporal dynamics in the absence of pitch grouping cues. Each study used a 2 (active vs. passive syntax) x 2 (present vs. absent pitch/pause boundary following phrase 1) x 2 (present vs. absent pitch/pause boundary following phrase) within-subjects design. This combination resulted in eight experimental conditions, in which prosody was manipulated to be congruent or incongruent with active-object (hierarchical centre-embedded) or reduced passive relative clause constructions. For each syntactic form, there were four cue conditions; active prosody, where there was a lengthened pause following the first phrase of the sentence, passive prosody, where a lengthened pause or pitch reduction followed the second phrase of the sentence, - the embedded noun phrase in actives, or the embedded, agent verb phrase for passives -, a no cues control where no lengthened pauses, or pitch reductions were present, and a second, both cues control, where lengthened pauses or pitch reductions occurred after both the first and second phrases. Table 1 displays examples for active syntactic sentences across all critical conditions, with pitch contours for the pitch similarity conditions, and waveforms for the temporal proximity conditions.

Each condition was comprised of eight experimental sentences, resulting in 64 critical trials. For each condition, we also included eight control sentences with a more obvious syntactic structure (active, “The girl chases the boy and he runs”; passive, “The running boy was chased by the girl”). As a result, participants were exposed to a total of 128 experimental trials. In each experimental trial, participants had to identify the target image of the auditorily presented sentence in the presence of three distractor images. Participants were only able to make their decision after the offset of the sentence.

Table 1

*Pitch and temporal dynamics by sentence structure*

| Prosody | Pitch Similarity | Temporal Proximity |
|---|---|---|

| | | |
|---|---|---|
| Active |  |  |
| Passive |  |  |
| Control All Cues |  |  |
| Control No Cues |  |  |

The boy  the girl  chases  runs
The boy  chased bythegirl runs

*2.3 Sentence generation*

The stimuli were created using the Festival speech synthesiser (Black, Taylor, & Caley, 1990). Each word was generated in isolation, using the default voice, at the default rate, with target pitch level set using the “default intonation” function, with pitch set at two intervals, the starting and closing pitch. Following this, we assessed the pitch contour and duration of each word in Praat (Version 6.0.13; Boersma, Paul & Weenink, 2017). The resulting contours were then manually flattened to result in words with a mean pitch at the target level to ensure only the pitch manipulations of interest could influence participant performance. To ensure each word was equivalent within a class, we extracted the $F_0$Hz value of each word at 5ms intervals and ran t-tests between each word. In cases where $p < .05$, we adjusted the pitch contour using Praat, until $p > .05$. Next, we matched the duration of each word within a grammatical category (e.g. each transitive verb would have the same duration) for each syntactic form, and thus across conditions, each stimulus within each form always had the same duration (except in the study focusing on pause cues, where they differed between prosodic conditions). This was achieved by lengthening vowels within each word (e.g. in runs, “u” would be extended). Finally, we generated full sentences using Audacity version 2.1.2 (Audacity Team, 2016) to combine the audio files for each word, with a 5ms inter-word pause.

In the pitch similarity condition, prosodic cues were defined as a pitch reduction of 15Hz reduction in the mean $F_0$Hz. This is highlighted on the pitch contours in Table 1, showing a single reduction in the active and passive prosodic conditions, whereas there are two reductions in the all cues control. In each case, utterances started at 180Hz. Therefore, in active and passive structures, after the first pitch reduction, each word had a mean value of 165Hz, and in the all cues control condition, after the second pitch reduction, each word had a mean pitch of 150Hz. Cast instead as semitone distance from middle C, there was a 1.51 semitone difference between utterance onset, and words following the first pitch reduction. Between the second and the first, there was a distance of 1.61 semitones. Critically, humans are able to detect a minimal difference of 0.8 semitones (Dowling & Harwood, 1986), making these changes well above the detectable threshold. The no cues control contained no pitch reductions. In the temporal proximity condition, pause cues lasted 111ms, and are illustrated in table 1 with blue vertical lines where the duration is highlighted with a horizontal blue arrow. These pauses were generated as a silent period in Audacity. Similar to the pitch condition, the active and passive conditions contained a single lengthened pause, the all cues control contained two, and the no cues control included no lengthened pauses (see Table 1 for greater detail). The pause duration and pitch reduction were based on the results of Trotter, Frost and Monaghan (Chapter 3). The pitch reduction is the mean of the largest inter-phrase $F_0$Hz reduction, i.e. between the main clause verb phrase and the embedded clause noun phrase for actives, and the embedded verb and noun phrase for passives. The pause duration represented the mean of the inter-phrase pauses in the same locations.

*2.4 Display composition*

In each experimental trial, participants had to identify the target image relating to the auditorily presented sentence in the presence of three distractor images. Each image showed one agent and one patient performing actions. Distractors were generated to contain one of three violations; 1) agent-verb violation, in which the agent performs a different action; 2) patient-verb violation, where the patient performs a different action; and 3) role-reversal, where the patient of the sentence performed the agents' role (see figure 1 below). We used two characters in this study, a boy and a girl, each of whom could perform eight transitive verbs (punch, kick, greet, mock, cheer, ignore, beg, applaud) and eight intransitive verbs (run, walk, kneel, grin, squat, sit, sneak, crawl). Each experimental scene was generated using Moho Studio 12 (Smith Micro Software, 2016). Here, default characters were modified, and then their skeletal models were manipulated to create poses for each action. Prior to running the study, we asked 5 participants to provide labels for each of these actions, to assess whether each was visually distinct from one another, and whether the actions could be identified. If an image failed to be distinct and identifiable, a new version of the image was created, and new participants were asked to provide labels for the set. This was repeated twice, resulting in our final set of images. The resulting images were then placed on a scene at the same Y-coordinate and matched at an equal X-coordinate from the edge of the scene. In each trial, four scenes were combined in the manner seen in figure 1.

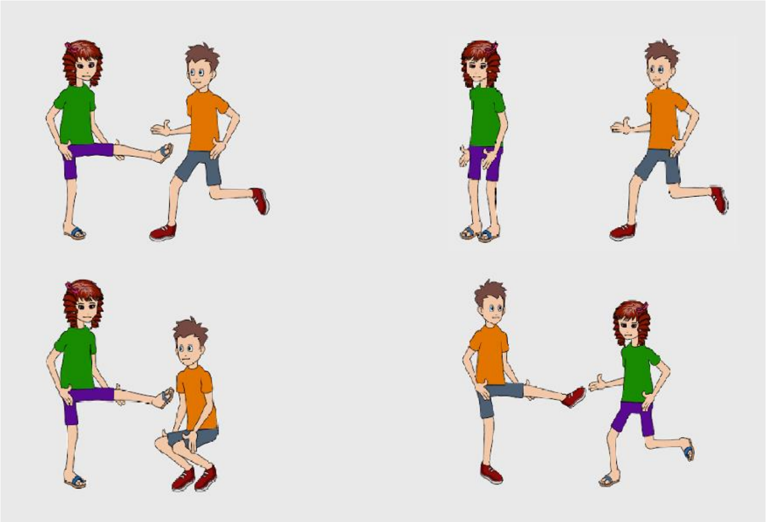


**Fig. 1**. Example experimental display, for the sentence, "The boy the girl kicks runs". The top left image is the target. The top right displays an agent-verb violation, in which the agent ("the girl") performs a different action on the patient ("ignores"). The bottom left displays a patient-verb violation, where the patient performs a different action ("squats"). The bottom right displays an agent-patient role reversal, where the agent ("the girl") becomes the patient, and vice versa, but the actions for the agent and patient remain the same.

### *2.5 Procedure*

Participants were randomly assigned to the experimental conditions. For each condition, participants viewed 128 experimental items, split into eight blocks. On each trial, participants viewed four images; the target image, an agent-verb distractor, patient-verb distractor, and an agent-patient role reversal distractor (see figure 1). The location of the target image and each distractor type was counter-balanced, with each occurring an equal number of times within each location. Each block comprised 16 trials, with eight critical sentences (four of each syntactic structure, with two of each prosodic condition) and eight control sentences (again, with four of each syntactic structure, with two of each prosodic condition). Sentences were presented in a random order.

The test session started with a familiarization session, in which participants were presented with each action and its verbal label in isolation, to ensure participants could subsequently identify each action. Following this, the eye-tracking study began. Eye movements were tracked using a Tobii X60 remote desktop tracker sampling at 60Hz. The distance was held constant between 55 and 60cm. The eye-tracker was calibrated prior to each experimental block.

Participants were asked to listen carefully to the sentences, and to not move their eyes away from the screen. We utilized a look-and-listen task (Huettig, Rommers, & Meyer, 2011); participants were not given specific viewing instructions. Each trial was structured as follows: First, a central fixation cross in the centre of the screen for 500ms. The cross then disappeared, and then the experimental display appeared for a 5s preview. This was included due to the complexity of the display (four events, with an agent and patient interacting), allowing participants to categorise each image. Next, the display disappeared, and were replaced with a second 500ms fixation cross. Next, the cross disappeared and was replaced with the experimental display, and the experimental sentence played. At sentence offset, the display remained, and participant indicated which of the displayed scenes was the target using the keyboard. Each participant was presented with all 128 items. The eye-tracking experiment, including calibration, took approximately 40 minutes. The data from participants' left and right eyes were analysed in terms of fixations. Fixations were coded as directed to the target, one of the three distractor types, or elsewhere. Further, we analysed participants' accuracy data. However, we were unable to assess participants' reaction time data due to computer recording error.

## 3. Results

### *3.1 Response Accuracy*

To analyse how participants' comprehension was affected by prosodic cues, we conducted a series of generalized linear mixed effects models (GLMER) predicting the dependent variable of response accuracy (correct vs. incorrect; 1 vs. 0), using a logit link function. As fixed effects, we tested the effect of experimental condition (pause vs. pitch cues), prosodic condition (active, passive, no cues control, all cues control, see Table 1), syntax (active vs. passive), trial number (1 – 128), and the interactions between these fixed factors. Random intercepts and slopes for subjects and items were included in all reported analyses.

The models were built up incrementally, adding in fixed effects and performing likelihood ratio tests after the addition of each new fixed term (following Barr, Levy, Scheepers, & Tily, 2013). Fixed effects were retained in the model if they resulted in a significant improvement of model fit in isolation, or as part of an interaction that improved model fit. Interaction terms were retained in the model if they improved model fit.

First, we analysed the effect of experimental condition on response accuracy. Including condition did not improve model fit ($\chi^2(1) = 0.109$, $p = .742$), indicating that participants responded with similar accuracy when exposed to both pitch similarity, or temporal proximity cues.

Next, we added the effect of syntax, which did not improve model fit ($\chi^2(1) = 0.569$, $p = .451$), indicating that participants responded with similar accuracy to both active and passive syntactic forms.

Following this, we analysed the effect of prosodic condition. The addition of prosodic condition did not improve model fit ($\chi^2(3) = 3.495$, $p = .321$), indicating that the prosodic grouping cue participants were exposed to did not affect overall comprehension.

The addition of trial number, however, did improve model fit ($\chi^2(1) = 185.330$, $p < .00001$), indicating that participant accuracy improved across the experimental session (see Table 2 for final model outcomes). This suggests that participants became more proficient at understanding the synthesised speech across the paradigm, or simply became more proficient at the experimental task.

Subsequently including the two-way interaction between condition and syntax resulted in a marginal improvement in model fit ($\chi^2(1) = 3.513$, $p = .06$). This reflected a trend towards higher performance with active structures in the pitch cues conditions than in the pause cues condition (see figure 2)

Including the two-way interactions between experimental condition and prosody ($\chi^2(3) = 1.120$, $p = .772$), experimental condition and trial ($\chi^2(1) = 0.794$, $p = .373$), syntax and prosody ($\chi^2(3) = 4.6222$, $p = .202$), and prosody and trial ($\chi^2(3) = 0.304$, $p = .959$) did not improve model fit.

Further, including the three-way interactions between condition, syntax and prosody ($\chi^2(9) = 13.757$, $p = .131$), condition, syntax and trial ($\chi^2(3) = 6.608$, $p = .085$), condition, prosody and trial ($\chi^2(7) = 4.140$, $p = .764$), and syntax, prosody and trial ($\chi^2(7) = 6.261$, $p = .510$) did not improve model fit.

Finally, including the four-way interaction between condition, syntax, prosody and trial did not improve model fit ($\chi^2(9) = 13.757$, $p = .131$).

The best-fitting model therefore included the main effects of experimental condition (Pause vs. Pitch), syntax (Active vs. passive), trial number, and the two-way interaction between cue condition and syntax.

**Table 2**

GLMER model outcomes for response accuracy

| Fixed Factor | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | 0.053 | 0.292 | 0.180 | .857 |
| Condition – Pitch | -0.065 | 0.341 | -0.190 | .850 |
| Syntax – Active | 0.008 | 0.225 | 0.033 | .974 |
| Trial Number | 0.015 | 0.001 | 13.302 | < .00001*** |
| Condition – Pitch: Syntax - Active | 0.304 | 0.161 | 1.892 | .059 |

Model Syntax: Accuracy ~ (1 + Condition*Syntax + Trial Number) + (1 + Condition*Syntax + Trial Number|Item) + Condition + Syntax + Condition:Syntax, family = binomial(logit)

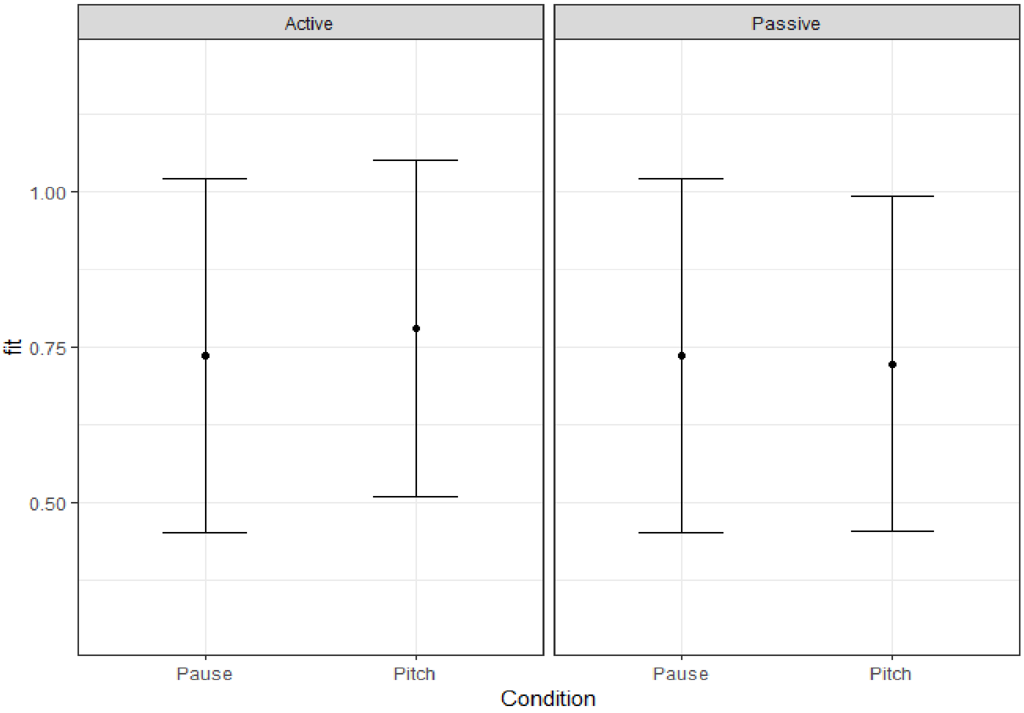


Fig. 2. Model estimates of probability of making a correct response split by syntactic form and condition. Points indicate the model estimate, black bars illustrate the 95% confidence interval of the estimate.

### *3.2 Gaze Behaviour: Pitch Similarity Cues*

In the following analyses, active and passive structures are analysed separately, due to the differences in duration between the two structures. Passive structures were longer than active structures due to the presence of the agent by-phrase (e.g. “The boy chased by the girl” vs. “The boy the girl chases”). If structures were to be analysed together, the time series would have to be scaled first, reducing the interpretability of any interactions with cue condition.

#### *3.2.1 Active Structures*

In these analyses, active structures were split into 3 critical analysis windows, to allow us to assess whether the presence or absence of a prosodic boundary affected fixation behaviour at different stages in incremental sentence comprehension. Analysis window 1 was taken 200ms following the onset of the active congruent pitch change, and the following 300ms (see Table 3 for the onset and offset of each analysis window by syntactic form). Analysis window 2 spanned 200ms after the onset of the embedded verb phrase, consistent with passive congruent pitch change, and lasted until the onset of the final verb. Analysis window 3 spanned 200ms after the onset of the final verb until its offset (1620 – 1770ms). Our analyses focussed on these time windows, as previous research estimates that the time needed to program and execute an eye movement can be as great as 150ms (see e.g., Matin, Shao, & Boff, 1993), allowing to assess whether participants became more likely to fixate the target in particular prosodic conditions. The offset of each time-window was selected such that it matched the offset of the phrase, meaning that the presence vs. absence of the subsequent pitch change did not influence fixation behaviour.

Figure 3 illustrates the mean proportion of fixations by time for the whole active trial, with the analysis regions marked.

**Table 3**

Analysis window onsets and offsets by syntactic form

| Syntax | Window | Onset (ms) | Offset (ms) |
|---|---|---|---|
| Active | 1 | 710 | 1020 |
| Active | 2 | 1220 | 1420 |
| Active | 3 | 1620 | 1770 |
| Passive | 1 | 710 | 915 |
| Passive | 2 | 1115 | 1415 |
| Passive | 3 | 1812 | 1962 |

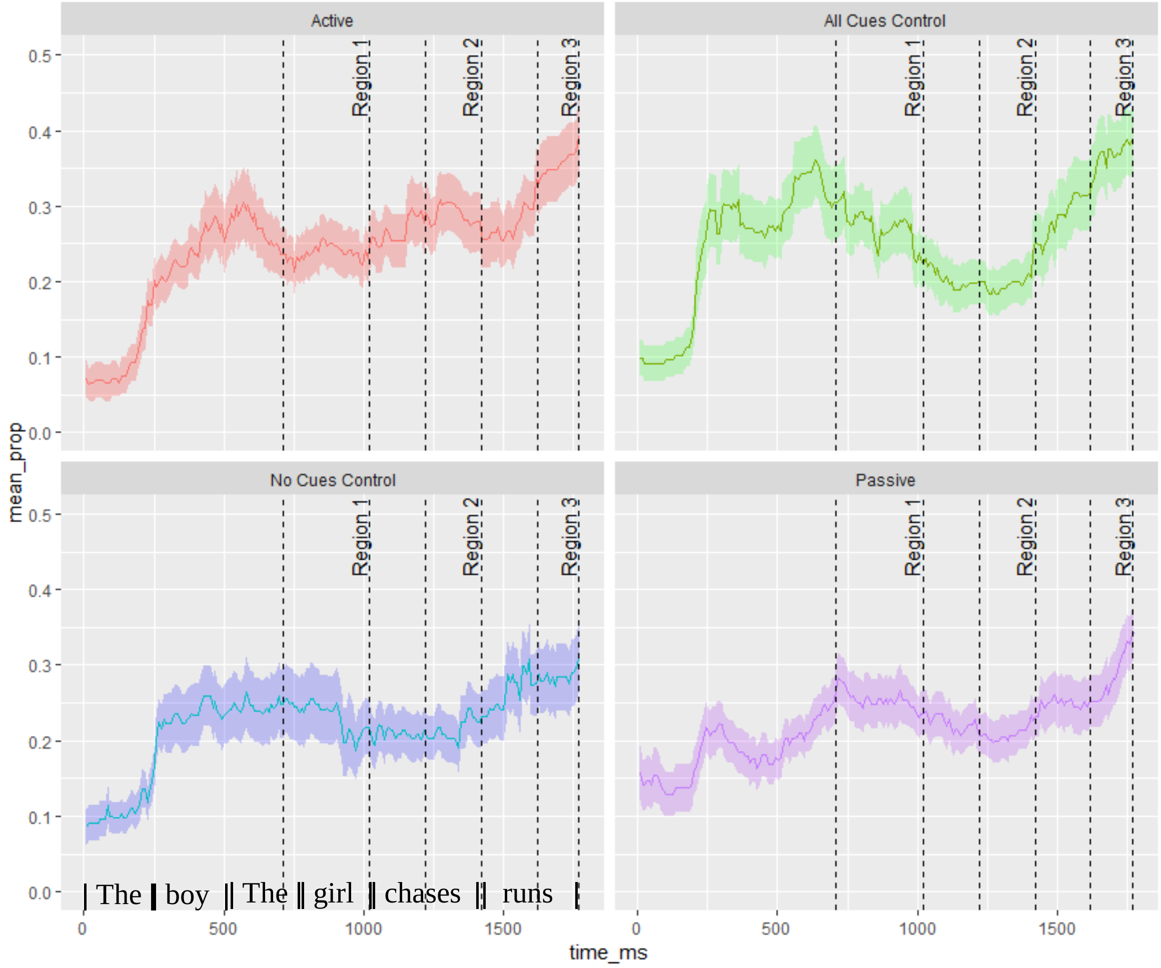


**Fig. 3**. Mean proportion of looks to target by prosodic condition, for active structures. The coloured line illustrates the mean, and the shaded area surrounding it illustrates the standard error. Analysis regions are indicated by dashed vertical lines, where the region label is at the region offset. At the base of the bottom-left panel, an example sentence is provided.

#### *3.2.1.1 Region 1: The embedded noun phrase*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a 4-level factor, wherein active was taken as the baseline predictor), the time-bin in which the fixation took place ($\text{Median}_{\text{bin duration}}$ = 8.372 ms, region duration = 310 ms, resulting in 37 time bins, coded as a numeric predictor), and the interactions between these fixed factors.

The models were built up incrementally, adding in fixed effects and performing likelihood ration tests after the addition of each new fixed term (following Barr, Levy, Scheepers, & Tily, 2013).

First, we analysed the effect of time bin on fixations to target vs. elsewhere. The effect of time bin significantly improved model fit ($\chi^2(1) = 9.960$, $p$ = .002), indicating that participants, overall made *fewer* looks to target as the determiner and noun phrase of the embedded clause unfolded (see Table 4 for final model outcomes). However, the estimate provided by the best-fitting model suggests that this change did not significantly differ from 0.

Next, we analysed the effect of prosodic condition. The addition of condition significantly improved model fit ($\chi^2(3) = 62.912$, $p < .00001$), indicating the probability of fixating was affected by the presence vs. absence of a prosodic boundary. The estimates provided by the final model, suggest that only the all cues control produced a lower number of target from fixations from the active cues condition, though this difference was not a significant contrast.

Finally, we included the interaction of prosodic condition and time bin in the analysis, which resulted in significant improvements in model fit ($\chi^2(3) = 8.821$, $p = .031$), reflecting exposure to the all cues control condition resulted in more looks to target over time, while the passive prosodic and no cues control conditions resulted in fewer looks to target over time, relative to the active prosodic condition.  Whilst the addition of the interaction term increased model fit, the estimates for the individual levels of the variables do not significantly differ from 0, and so the interaction is due to the effects of each prosodic condition diverging from one another over time.

The best-fitting model therefore included both the main effects of time bin and prosodic condition, and their interaction.

**Table 4**

GLMER Model outcomes for target fixation likelihood in analysis region1

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -1.056 | 0.392 | -2.690 | .007** |
| Time-bin | -0.0005 | 0.0004 | -1.429 | .153 |
| Prosody – No Cues Control | 0.339 | 0.483 | 0.703 | .482 |
| Prosody – All Cues Control | -0.769 | 0.460 | -1.671 | .095 |
| Prosody – Passive | 0.149 | 0.450 | 0.331 | .741 |
| Time-bin: Prosody – No Cues Control | -0.0008 | 0.0005 | -1.523 | .127 |
| Time-bin: Prosody – All Cues Control | 0.0008 | 0.0005 | 1.410 | .159 |

| | | | | |
|---|---|---|---|---|
| Time-bin: Prosody – Passive | -0.0002 | 0.0005 | -0.461 | .645 |

Model syntax: Target Fixation ~ (1 + Prosody:Time-bin | Subject) + (1 + Prosody: Time-bin| Item) + Prosody + Time-bin + Prosody: Time-bin, family = binomial(logit)

*3.2.1.2 Region 2: The embedded verb phrase*

We conducted a series of generalised linear mixed-effects models (GLMER) predicting the dependent variable of fixations to target vs. other (1 vs. 0) using a binomial distribution, using a logit link function, split by analysis window. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a 4-level factor, wherein active was taken as the baseline predictor), the time-bin in which the fixation took place (1 to 24, coded as a numeric predictor), and the interactions between these fixed factors. The analysis proceeded here in the same way for the first analysis. Random intercepts and slopes for participant and item were included in all reported analyses.

First, we analysed the effect of time bin on fixations to target vs. elsewhere. The effect of time bin did not improve model fit ($\chi^2(1) = 1.326$, $p = .250$). However, in the final model, time-bin was retained as a predictor, as it contributed to a significant interaction (see table 5 for final model outcomes), and indicated that on average, participants tended to make fewer looks to target, an effect primarily driven by the reduction seen in the active cues condition.

Next, we analysed the effect of prosodic condition. The addition of condition significantly improved model fit ($\chi^2(3) = 195.13$, $p < .00001$), reflecting that the active condition, overall, elicited more looks to target than the other cue conditions over the verb phrase of the embedded clause.

Finally, we included the interaction of prosodic condition and time bin in the analysis, which resulted in significant improvements in model fit ($\chi^2(3) = 10.385$, $p = .016$). This reflects the downward trend for looks to target in the active cues control, while in contrast, the all cues control condition elicited a small increase in looks to target over time (see figure 4 for greater detail).

The best-fitting model thus included the following fixed effects; time bin, prosodic condition, and the two-way interaction between time bin and prosodic condition.

**Table 5**

GLMER model outcomes for target fixation likelihood in analysis region 2

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | 0.610 | 0.979 | 0.623 | .533 |
| Time-bin | -0.002 | 0.001 | -2.216 | .027* |
| Prosody – No Cues Control | -4.325 | 1.401 | -3.088 | .002** |
| Prosody – All Cues Control | -4.312 | 1.382 | -3.120 | .002** |
| Prosody – Passive | -4.278 | 1.336 | -3.201 | .001** |
| Time-bin: Prosody – No Cues Control | 0.003 | 0.001 | 2.519 | .012* |
| Time-bin: Prosody – All Cues Control | 0.003 | 0.001 | 2.625 | .009** |
| Time-bin: Prosody – Passive | 0.003 | 0.001 | 2.632 | .008** |

Model syntax: Target Fixation ~ (1 + Prosody:Time-bin | Subject) + (1 + Prosody: Time-bin| Item) + Prosody + Time-bin + Prosody: Time-bin, family = binomial(logit)

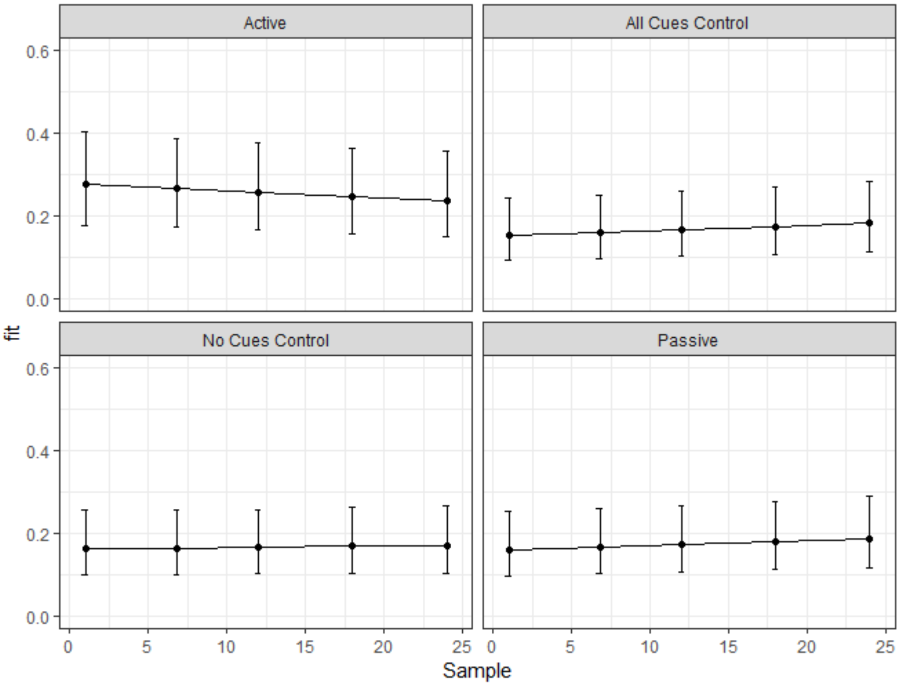


**Fig. 4**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

*3.2.1.3 Region 3: The main clause verb*

We conducted a series of generalised linear mixed-effects models (GLMER) predicting the dependent variable of fixations to target vs. other (1 vs. 0) using a binomial distribution, using a logit link function, split by analysis window. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a 4-level factor, wherein active was taken as the baseline predictor), the time bin in which the fixation took place (1 – 18, coded as a numeric predictor), and the interactions between these fixed factors. The analysis proceeded here in the same way for the first analysis. Random intercepts and slopes for participant and item were included in all reported analyses.

First, we analysed the effect of time bin on fixations to target vs. elsewhere. The effect of time bin significantly improved model fit ($\chi^2(1) = 11.758$, $p = .0006$), indicating that participants made more looks to target as the final verb phrase unfolded (see Table 6 for the final model outcomes). In the final model outcomes, however, time bin did not significantly differ from 0, demonstrating that the variance explained by time bin was moderated by cue condition.

Next, we added the main effect of prosodic condition, which significantly improved model fit ($\chi^2(3) = 85.633$, $p < .00001$). This reflected the tendency for participants to initially make fewer looks to target at the onset of the analysis window (active = 0.31, passive = 0.18), though by the end of the window, the two conditions are equivalent.

Finally, we included the interaction between prosodic condition, and time bin, which significantly improved model fit ($\chi^2(3) = 10.021$, $p = .018$), reflecting that participants made more looks to target in the passive prosodic over time, relative to the active congruent condition (see figure 5 for greater detail).

Thus, the final model included the following terms; time bin, prosodic condition, and their two-way interaction.

**Table 6**

GLMER model outcomes for target fixation likelihood in analysis region 3

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -2.233 | 1.859 | -1.201 | .230 |
| Time-bin | 0.001 | 0.001 | 0.613 | .540 |
| Prosody – No Cues Control | -0.779 | 2.638 | -0.296 | .768 |
| Prosody – All Cues Control | -0.014 | 2.539 | -0.005 | .996 |
| Prosody – Passive | -6.802 | 2.527 | -2.691 | .007** |
| Time-bin: Prosody – No Cues Control | 0.0002 | 0.002 | 0.148 | .882 |
| Time-bin: Prosody – All Cues Control | 0.0001 | 0.002 | 0.079 | .937 |
| Time-bin: Prosody – Passive | 0.004 | 0.001 | 2.593 | .009** |

Model syntax: Target Fixation ~ (1 + Prosody*Time-bin | Subject) + (1 + Prosody* Time-bin| Item) + Prosody + Time-bin + Prosody: Time-bin, family = binomial(logit)

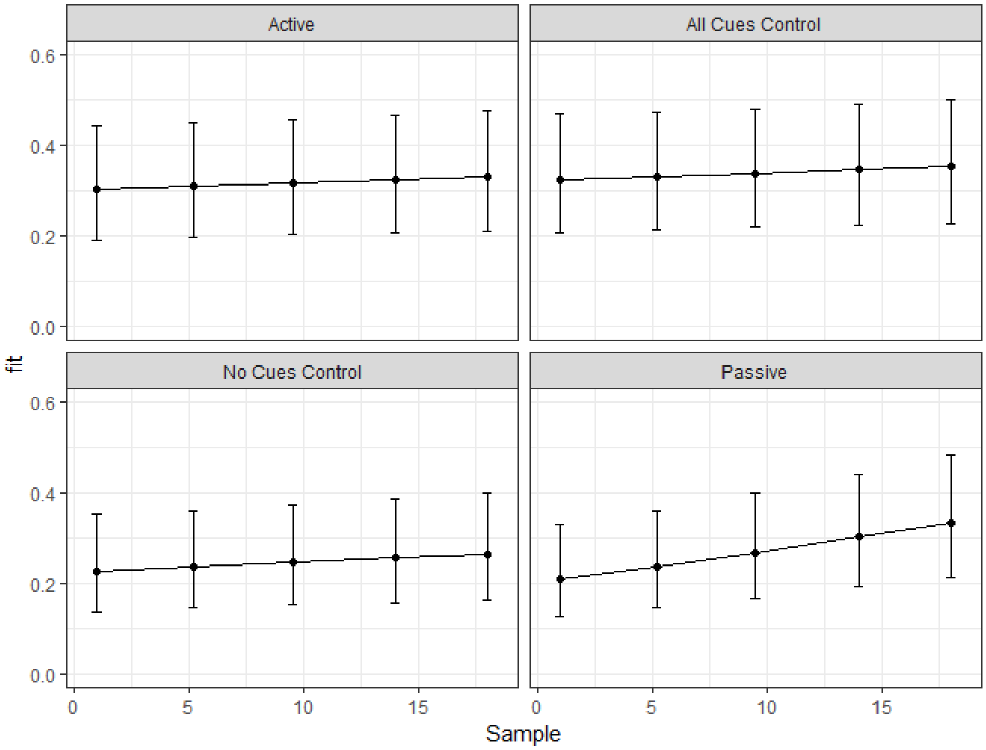


**Fig. 5**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

### *3.2.1.4 Summary*

For active structures, no significant differences arose in the region of the embedded noun phrase, suggesting that prosody did not produce any immediate (200 – 500ms) processing benefit following the first, active-congruent pitch boundary. In the second analysis region, taking place over the embedded verb phrase, the main effect of prosody indicated that participants made more looks to target in the active congruent pitch cues condition, than they did in the passive, all, and no cues control conditions. This suggested that in this analysis region, prior exposure (prior to the embedded noun phrase) to a pitch grouping cue consistent with the syntactic structure facilitated processing, and exposure to an incongruent cue impaired processing. The significant time-bin by prosody interaction in this analysis window indicated that the passive, no and all cues conditions showed a greater increase in target fixations over time, indicating that participants could largely compensate for the incongruent prosody. Interestingly, looks to target fell somewhat over the course of the embedded verb, perhaps indicating that the benefit of prosody was temporary. In the analysis region consistent with the final verb, the main effect of prosody indicated that passive cues produced a lower number of target fixations than did active cues, particularly at the start of the analysis region. The significant time by prosody interaction indicated that passive cues elicited a greater increase in looks to target across the final verb relative to active cues, suggesting compensation for any deficits introduced by passive prosody.

Overall, for active structures, it therefore appears that pitch similarity cues congruent with syntactic form provided a processing benefit, with a higher probability of fixating the target during the embedded verb phrase. Participants do, however, seem to compensate for incongruent prosody, or a lack of prosody over the course of the utterance, as indicated by the similar performance by the close of the main clause verb.

### *3.2.2 Passive Structures*

In these analyses, passive syntactic structures were split into 3 critical analysis windows (see Table 3), to allow us to assess whether the presence or absence of a prosodic boundary affected fixation behaviour.  Analysis region 1 occurred 200ms following the onset of the embedded, agent verb phrase. Analysis region 2 spanned 200 - 500ms following the onset of the embedded, agent noun phrase. The onset of analysis region 3 was 200ms following the onset of the patient verb phrase of the main clause, and lasted until its offset. Figure 6 illustrates the mean proportion of fixations by time for the whole passive trial, with the analysis regions marked.

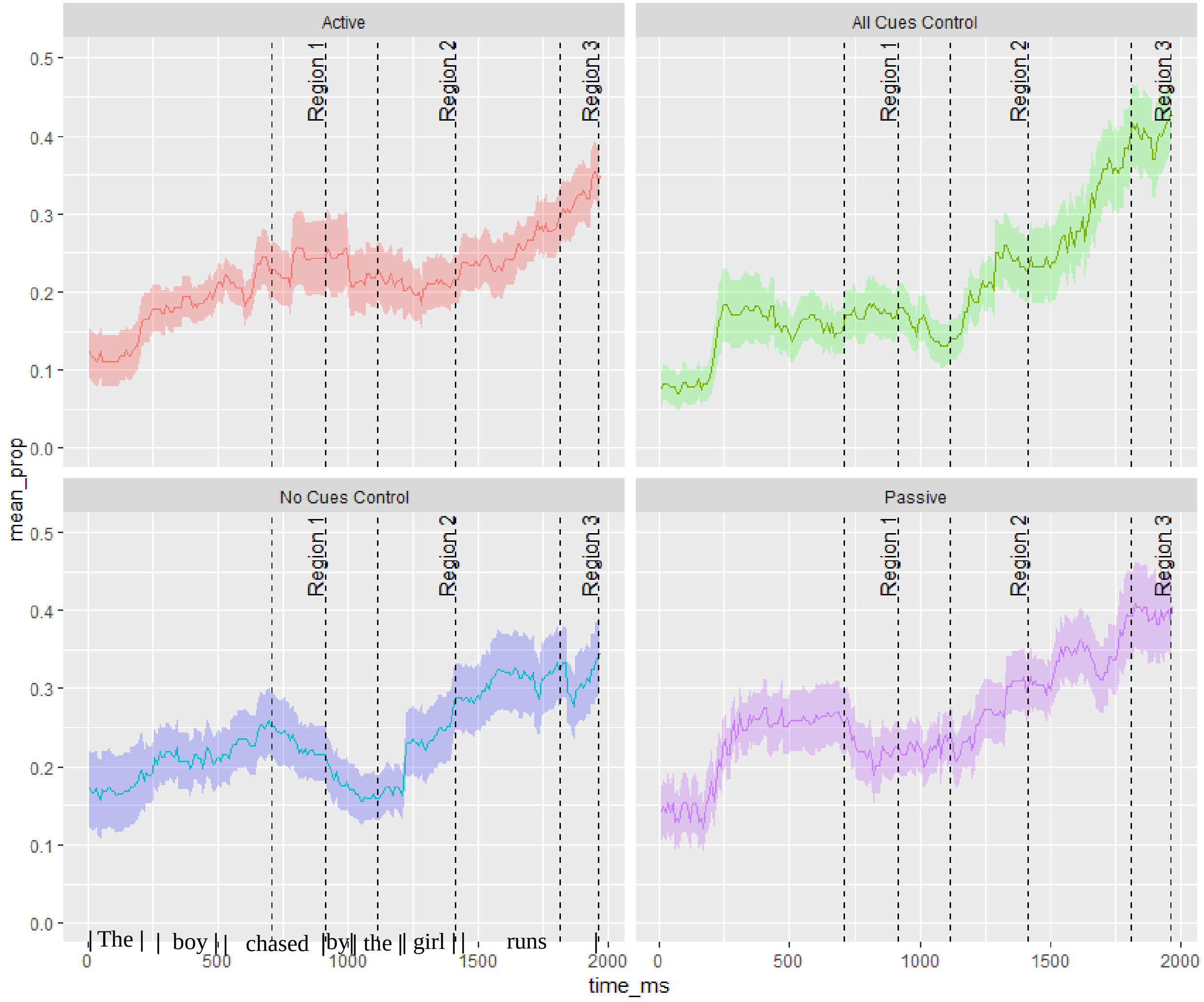


**Fig. 6**. Mean proportion of looks to target by prosodic condition, for passive structures. The coloured line illustrates the mean, and the shaded area surrounding it illustrates the standard error. Analysis regions are indicated by dashed vertical lines, where the region label is at the region offset.

### *3.2.2.1 Region 1: The embedded verb phrase*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a 4-level factor, wherein active was taken as the baseline predictor), the time-bin in which the fixation took place (1 – 24, coded as a numeric variable), and the interactions between these fixed factors.

First, we analysed the effect of time bin. Including the effect of time bin did not improve model fit ($\chi^2(1) = 0.790$, *p* = .374), indicating that participants did not make more looks to target as the verb phrase of the embedded clause unfolded.

Next, we added the main effect of prosodic condition, which significantly improved model fit ($\chi^2(3) = 10.339$, *p* = .016), reflecting the fact that, overall, participants made fewer looks to target in the all cues control, relative to the active prosodic condition (see Table 7 for the final model outcomes).

Finally, we included the interaction between prosodic condition, and time bin, which failed to improve model fit ($\chi^2(3) = 2.311$, *p* = .510).

The best-fitting model for this region included only the main effect of prosody.

**Table 7**

GLMER model outcomes predicting target fixation likelihood in analysis region 1

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -2.344 | 0.340 | -6.892 | < .00001*** |

| | | | | |
|---|---|---|---|---|
| Prosody – No Cues Control | -0.068 | 0.074 | -0.914 | .361 |
| Prosody – All Cues Control | -0.246 | 0.080 | -3.066 | .002** |
| Prosody – Passive | -0.052 | 0.075 | -0.695 | .487 |

Model syntax: Target Fixation ~ (1 + Prosody| Subject) + (1 + Prosody| Item) + Prosody, family = binomial(logit)

*3.2.2.2 Region 2: The embedded noun phrase*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, both cues control, coded as a four-level factor with active as the baseline predictor), the time-bin in which the fixation took place (1 – 36, coded as a numeric predictor), and the interactions between these fixed factors.

First, we analysed the effect of time bin. The addition of time bin to the model resulted significantly improved model fit ($\chi^2(1) = 45.058$, $p < .00001$), reflecting that overall, participants made more looks to target as the by phrase of the embedded clause unfolded. In the final model, however, the resulting estimate failed to significantly differ from 0, indicating that the variance explained by time bin was moderated by prosodic condition.

Next, we analysed the effect of prosodic condition on fixation behaviour. The addition of prosodic condition significantly improved model fit ($\chi^2(3) = 40.247$, $p < .00001$), reflecting the tendency for both control conditions to produce fewer looks to target overall during the embedded noun phrase. In contrast, passive cues tended to elicit a greater number of looks to target in this time-window, suggesting that passive congruent cues, overall tended to bias participants towards the target.

Finally, we included the interaction between prosodic condition and time bin, which significantly improved model fit ($\chi^2(3) = 23.288$, $p = .00003$). This reflected the tendency for passive, and both control prosodic conditions to elicit greater looks to target over the unfolding by phrase than the active congruent prosodic condition (see figure 7 for greater detail).

The best-fitting model therefore included the main effects of time bin, prosody, and the two-way interaction between time bin and cue condition.

**Table 8**

GLMER model outcomes predicting target fixation likelihood in analysis region 2

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -1.709 | 0.756 | -2.966 | .003** |
| Time-bin | 0.0002 | 0.0004 | -0.500 | .617 |
| Prosody – No Cues Control | -3.197 | 0.730 | -4.381 | .00001*** |
| Prosody – All Cues Control | -1.625 | 0.789 | -2.062 | .039* |
| Prosody – Passive | 2.487 | 0.745 | 3.340 | .0008** |
| Time-bin: Prosody – No Cues Control | 0.003 | 0.0006 | 4.458 | < .00001*** |
| Time-bin: Prosody – All Cues Control | 0.001 | 0.0006 | 2.078 | .037* |
| Time-bin: Prosody – Passive | 0.002 | 0.0006 | 3.763 | .0002** |

Model syntax: Target Fixation ~ (1 + Prosody* Time-bin| Subject) + (1 + Prosody* Time-bin| Item) + Prosody + Time-bin + Prosody: Time-bin, family = binomial(logit)

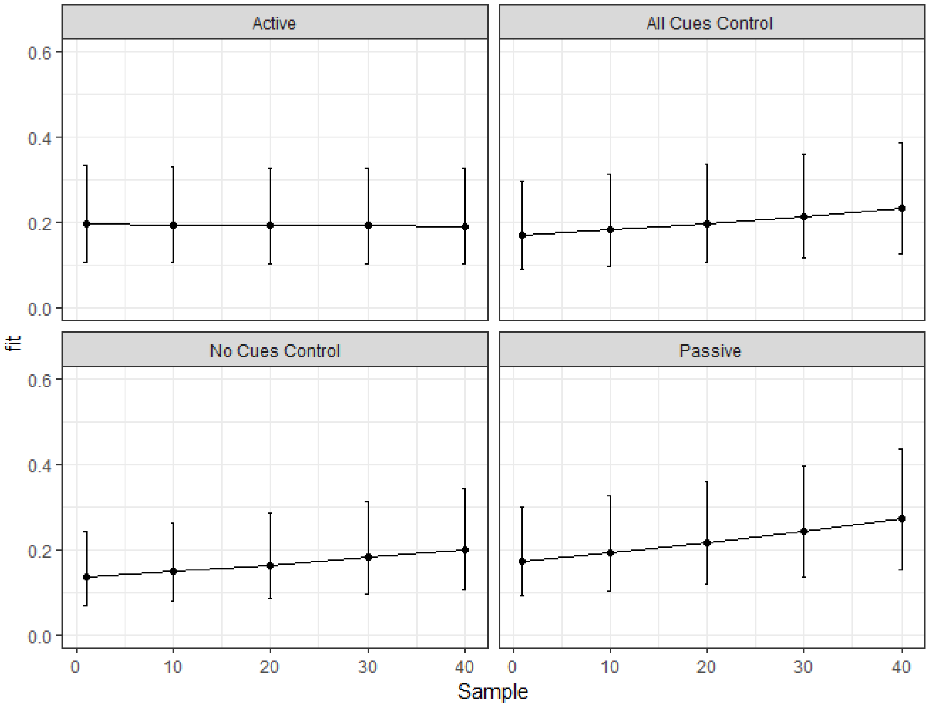


**Fig. 7**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

*3.2.2.3 Region 3: The main clause verb*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a four level factor with active as the baseline predictor), the time-bin in which the fixation took place (1 – 18, coded as a numeric predictor), and the interactions between these fixed factors.

First, we analysed the effect of time bin. The addition of time bin to the model resulted in significantly improved model fit ($\chi^2(1) = 10.815$, $p = .001$), reflecting the tendency for participants to make more looks to target over the course of the sentence final verb (see Table 9 for the final model outcomes).

Next, we analysed the effect of prosodic condition on fixation behaviour. The addition of prosodic condition significantly improved model fit ($\chi^2(3) = 24.554$, $p = .00002$).

Finally, we included the interaction between prosodic condition and time bin, which did not improve model fit ($\chi^2(3) = 2.918$, $p = .404$), indicating that the increased looks to target over time occurred globally, and was not modulated by prosodic condition.

Thus, the final model included the two simple main effects of time bin and prosody.

**Table 9**

GLMER model outcomes predicting target fixation likelihood in analysis region 3

| Fixed Effect | Estimate | Std. Error | *Z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -4.510 | 1.118 | -4.032 | .00005*** |
| Time-bin | 0.002 | 0.001 | 3.304 | .0009** |
| Prosody – No Cues Control | 0.034 | 0.073 | 0.463 | .643 |
| Prosody – All Cues Control | 0.309 | 0.079 | 3.934 | .00008*** |
| Prosody – Passive | -0.067 | 0.077 | -0.870 | .384 |

Model syntax: Target Fixation ~ (1 + Prosody + Time-bin| Subject) + (1 + Prosody + Time-bin| Item) + Prosody + Time-bin, family = binomial(logit)

### *3.2.2.4 Summary*

For passive structures, in the analysis region in the embedded verb phrase, the only cue that differed significantly from active cues was the all cues control condition, suggesting that a pitch boundary in this location reduced the likelihood of participants fixating the target. In the analysis region encompassing the agent by-phrase (the embedded noun phrase), the main effect of prosody reflected a higher number of target fixations in the passive congruent condition relative to the active cues condition, showing an immediate benefit for a pitch-prosodic boundary consistent with passive structures. In contrast, the all and no cues control conditions produced worse performance that active congruent cues, indicating that a flat prosodic contour, or including both a consistent and an inconsistent prosodic boundary disrupted processing. The time-bin by prosody interaction indicated that passive, no cues and all cues control condition elicited a greater increase in target fixations across this time window, reflected the trend in figure 6 for each of these conditions to exceed the target fixations of the active cues condition by the end of the analysis region. Finally, the main clause verb analysis region produced only a main effect of prosody, that indicated that only performance on the all cues control condition differed from the active cues condition, eliciting a higher number of target fixations.

Overall, passive pitch similarity cues conferred a processing benefit for passive structures. However, this benefit was not as robust as that seen for active structures, with it being present only in the region immediately following the congruent change. Intriguingly, in the region of the main clause verb, the all cues control condition elicited a greater number of looks to target, raising the possibility that either overall pitch variation improved performance, or that having a pitch boundary in locations congruent with passive syntactic structures improves performance, regardless of the preceding context.

### *3.3 Gaze behaviour: Temporal proximity cues*

### *3.3.1 Active Structures*

This analysis is similar to the analysis conducted for the pitch responses, splitting each stimulus up into three critical windows. Here, critical windows were taken as starting 200ms following the offset of lengthened pauses where appropriate. Similar to the pitch analysis, the offset of each window was placed at the offset of the following phrase. As a result, the onset and offset of analysis windows changed on the basis of prosodic condition. Table 10 denotes the onsets and offsets for each analysis window for active structures. Figure 8 displays the mean proportion of fixations by prosodic form, with the lengthened pauses removed from the time-series, to allow for cross condition comparisons.

**Table 10**

Onsets and offsets for analysis windows by prosodic form

| Prosody | Window | Onset (ms) | Offset (ms) |
|---|---|---|---|
| Active | 1 | 816 | 1161 |
| Active | 2 | 1226 | 1526 |
| Active | 3 | 1736 | 1881 |
| Passive | 1 | 710 | 1110 |
| Passive | 2 | 1326 | 1526 |
| Passive | 3 | 1731 | 1881 |
| No Cues Control | 1 | 715 | 1015 |
| No Cues Control | 2 | 1220 | 1425 |
| No Cues Control | 3 | 1625 | 1775 |
| All Cues Control | 1 | 816 | 1116 |
| All Cues Control | 2 | 1433 | 1632 |
| All Cues Control | 3 | 1837 | 1987 |

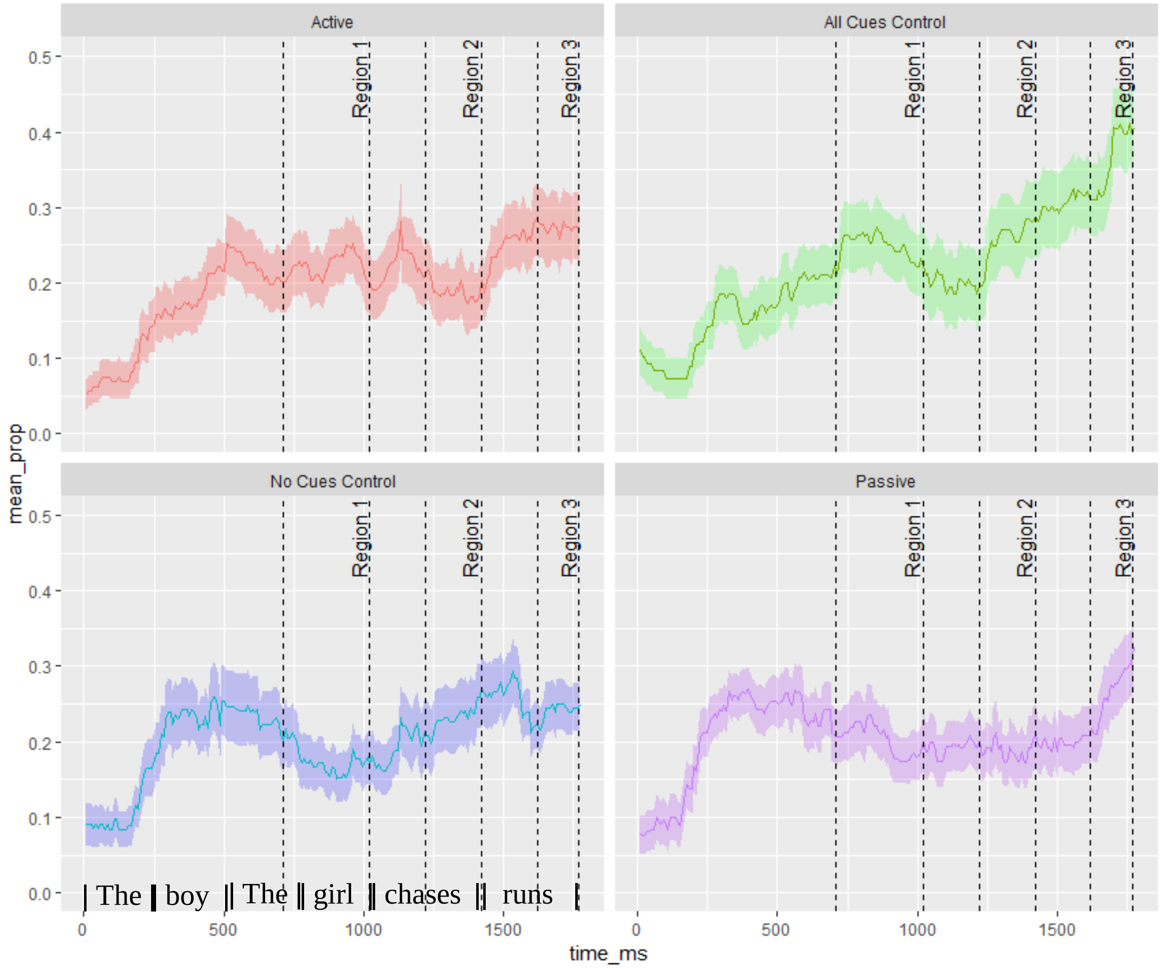


**Fig. 8**. Mean proportion of looks to target by prosodic condition, for active structures. The coloured lines illustrate the mean, and the shaded area surrounding it illustrates the standard error. Analysis regions are indicated by dashed vertical lines, where the region label is at the region offset. At the base of the bottom left panel, there is an example sentence.

### *3.3.1.1 Region 1: The embedded noun phrase*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a four level factor with active acting as the baseline condition), the time-bin in which the fixation took place (1 – 36, coded as numeric), and the interactions between these fixed factors.

The models were built up incrementally, adding in fixed effects and performing likelihood ration tests after the addition of each new fixed term (following Barr, Levy, Scheepers, & Tily, 2013).

First, we analysed the effect of time bin on fixations to target vs. elsewhere. The effect of time bin significantly improved model fit ($\chi^2(1) = 12.658$, $p$ = .0004), indicating that participants, overall made fewer looks to target as the determiner and noun phrase of the embedded clause unfolded (see Table 11 for final model outcomes). This reflects the reduction in target fixations across the passive and both control conditions in figure 8.

Next, we analysed the effect of prosodic condition. The addition of condition did not improve model fit ($\chi^2(3) = 1.831$, $p$ = .608), indicating the probability of fixating the target was not affected by the presence of a lengthened pause preceding the embedded noun phrase.

Finally, we included the interaction of prosodic condition and time bin in the analysis, which did not improve model fit ($\chi^2(3) = 5.543$, $p$ = .136), indicating that the increased probability of fixating the target over the unfolding noun phrase was not affected by prosodic condition.

The best-fitting model thus only included the main effect of time bin.

**Table 11**

GLMER model outcomes predicting target fixation likelihood in analysis region 1

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -1.143 | 0.301 | -3.803 | .0001** |
| Time-bin | -0.006 | 0.002 | -3.561 | .0004** |

Model syntax: Target Fixation ~ (1 + Time-bin| Subject) + (1 + Time-bin| Item) +Time-bin, family = binomial(logit)

*3.3.1.2 Region 2: The embedded verb phrase*

We conducted a series of generalised linear mixed-effects models (GLMER) predicting the dependent variable of fixations to target vs. other (1 vs. 0) using a binomial distribution, using a logit link function, split by analysis window. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a four level factor with active as the baseline condition), the time-bin in which the fixation took place (1 – 24, coded as a numeric predictor), and the interactions between these fixed factors. The analysis proceeded here in the same way for the first analysis. Random intercepts and slopes for participant and item were included in all reported analyses.

First, we analysed the effect of time bin on fixations to target vs. elsewhere. The effect of time bin did not improve model fit ($\chi^2(1) = 0.445$, $p = .505$), indicating that, overall, participants did not make more looks to target over the course of the unfolding utterance. The effect of time bin was retained in the final model, as it took part in a significant interaction (see Table 12 for final model outcomes). The main effect indicated that the no cues control, and all cues control elicited a higher number of target fixations over the course of the window (see figure 8).

Next, we analysed the effect of prosodic condition. The addition of condition did not improve model fit ($\chi^2(3) = 0.738$, $p = .864$), however it was retained in the model as it played a role in a significant interaction. This main effect indicated that at the onset of the analysis window, relative to active cues, each prosodic condition had a lower number of target fixations (see figures 8 and 9), which became equivalent by the middle of the sample, and never greatly exceeded the active cues condition.

Finally, we included the interaction of prosodic condition and time bin in the analysis, which resulted in significant improvements in model fit ($\chi^2(3) = 22.444$, $p = .00005$). This reflects the tendency for fixations on the target to increase more in both control conditions than in the active prosodic conditions (see figure 9 for greater detail), however the passive cues condition did not improve over time, and the active cues condition resulted in a lower number of target fixations over time.

The best-fitting model thus contained the main effects of time bin, prosody, and the interaction between time bin and prosody.

**Table 12**

GLMER model outcomes predicting target fixation likelihood in analysis region 1

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | 0.000001 | .658 | 0.000 | .999 |
| Time-bin | -.012 | .004 | -3.529 | .0004** |
| Prosody – No Cues Control | .392 | .182 | 3.320 | .0009** |
| Prosody – All Cues Control | .437 | .127 | 3.445 | .0006** |
| Prosody – Passive | -.221 | .129 | -1.710 | .087 |
| Time-bin: Prosody – No Cues Control | .026 | .007 | 3.713 | .0002** |
| Time-bin: Prosody – All Cues Control | .029 | .008 | 3.670 | .0002** |
| Time-bin: Prosody – Passive | .014 | .008 | 1.735 | .083 |

Model syntax: Target Fixation ~ (1 + Prosody* Time-bin| Subject) + (1 + Prosody*Time-bin| Item) + Prosody +Time-bin + Prosody:Time-bin, family = binomial(logit)

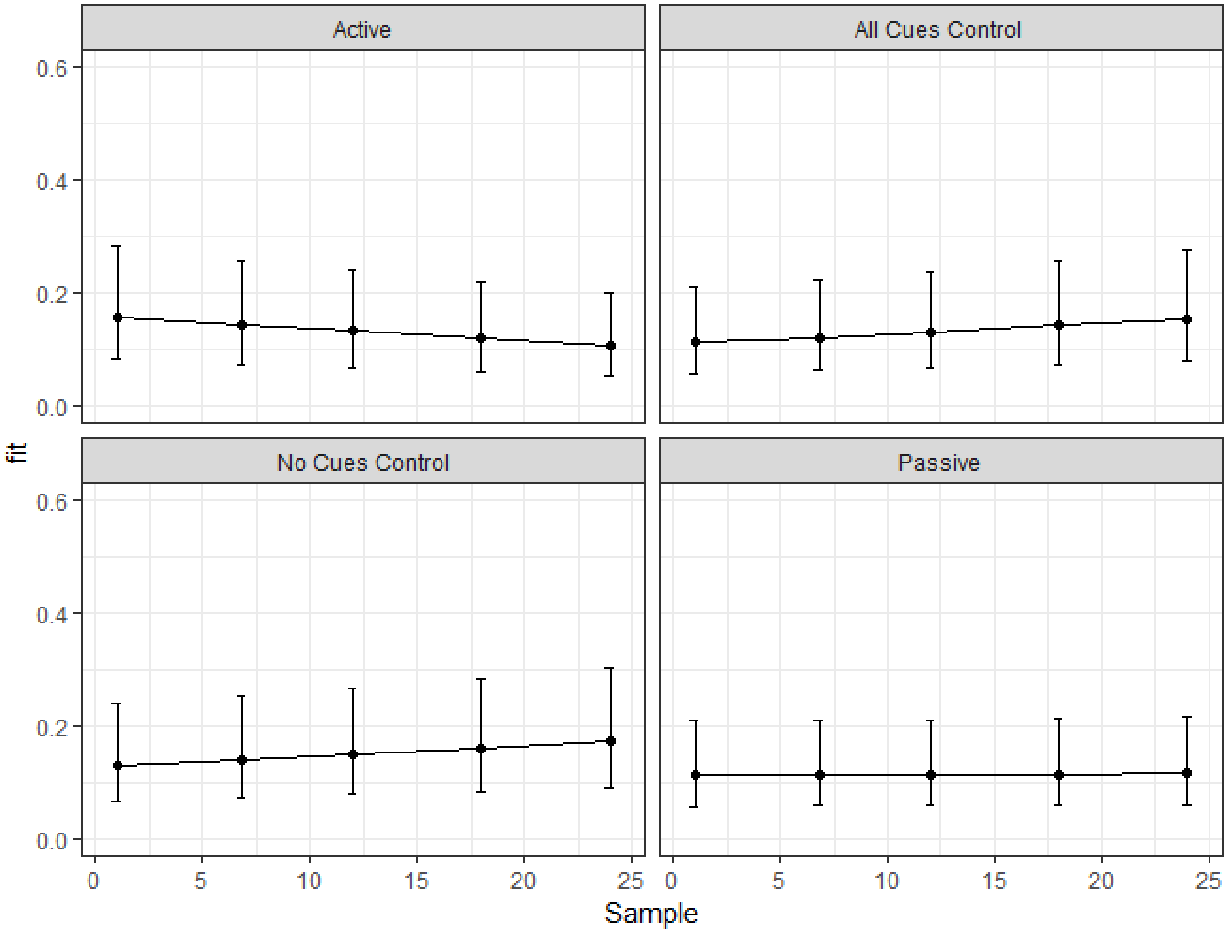


**Fig. 9**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

*3.3.1.3 Region 3: The main clause verb*

We conducted a series of generalised linear mixed-effects models (GLMER) predicting the dependent variable of fixations to target vs. other (1 vs. 0) using a binomial distribution, using a logit link function, split by analysis window. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a four level factor, with active acting as the baseline predictor), the time bin in which the fixation took place (1 – 18, coded as a numeric predictor), and the interactions between these fixed factors. The analysis proceeded here in the same way for the first analysis. Random intercepts and slopes for participant and item were included in all reported analyses.

First, we analysed the effect of time bin on fixations to target vs. elsewhere. The effect of time bin significantly improved model fit ($\chi^2(1) = 10.804$, $p = .001$). However, the main effect of time bin did not significantly differ in the final model indicating that the variance explained by time bin can be attributed to its interactions with other factors (see Table 12 for the final model outcomes).

Next, we added the main effect of prosodic condition, which did not improve model fit ($\chi^2(3) = 1.452$, $p = .693$), indicating that in isolation, there were no overall differences in the probability of making target fixations on the basis of prosody.

Finally, we included the interaction between prosodic condition, and time bin, which significantly improved model fit ($\chi^2(3) = 8.595$, $p = .035$), reflecting the trend towards increased likelihood of fixating the target in the passive prosodic over time, relative to the active congruent condition (see figure 10 for greater detail). Whilst the addition of the interaction term increased model fit, the estimates for the individual levels of the variables do not significantly differ from 0, and so the interaction is due to the effects of each prosodic condition diverging from one another over time.

The best-fitting model thus included the main effects of time bin and prosodic condition, and their interaction.

**Table 12**

GLMER model outcomes predicting target fixation likelihood in analysis region 3

| Fixed Effect | Estimate | Std. Error | *Z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -2.464 | 2.166 | -1.137 | .255 |
| Time-bin | 0.006 | 0.011 | 0.546 | .585 |
| Prosody – No Cues Control | 1.534 | 2.929 | 0.524 | .600 |
| Prosody – All Cues Control | -4.265 | 2.946 | -1.448 | .147 |
| Prosody – Passive | -5.184 | 2.927 | -1.771 | .077 |
| Time-bin: Prosody – No Cues Control | -0.007 | 0.014 | -0.507 | .612 |
| Time-bin: Prosody – All Cues Control | 0.023 | 0.014 | 1.574 | .115 |
| Time-bin: Prosody - Passive | 0.026 | 0.014 | 1.823 | .068 |

Model syntax: Target Fixation ~ (1 + Prosody* Time-bin| Subject) + (1 + Prosody* Time-bin| Item) + Prosody +Time-bin + Prosody:Time-bin, family = binomial(logit)

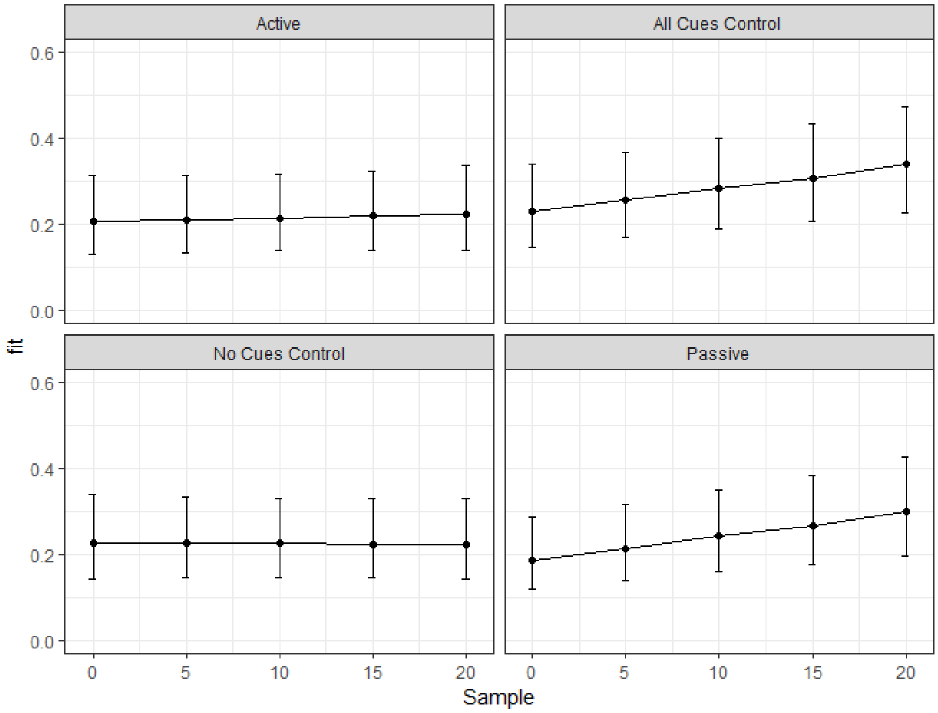


**Fig. 10**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

### *3.3.1.4 Summary*

Overall, in active structures, the temporal proximity results were similar to those for pitch similarity. In the analysis region encompassing the embedded noun phrase, active pitch cues did not produce an increased number of looks to target, relative to the other conditions, and only time-bin produced a significant effect, with a reduction of looks to target over time. In the embedded clause verb region, the no cues and all cues condition produced more looks to target than the active cues control condition. Furthermore over the course of the analysis region, looks to target were more likely in these conditions, while in the active cues condition, looks to target became less likely. During the final analysis region, there were significant main effects or interactions, however, there was a trend towards passive pause cues eliciting more looks to target, and more looks to target over time. Overall, it therefore seems that participants did not benefit from pause cues that were consistent with the syntactic structure, and seemed to benefit the most when pause boundaries were present in both locations, potentially suggesting that increased processing time may have facilitated processing.

### *3.3.2 Passive Structures*

This analysis proceeded as for the active structures with temporal grouping cues, splitting each stimulus up into three critical windows. Here, critical windows were taken as starting 200ms following the offset of lengthened pauses where appropriate. Similar to the pitch analysis, the offset of each window was placed at the offset of the following phrase. As a result, the onset and offset of analysis windows changed on the basis of prosodic condition. Table 13 denotes the onsets and offsets for each analysis window. Figure 11 displays the mean proportion of fixations by prosodic form, with the lengthened pauses removed from the time-series, to allow for cross condition comparisons.

**Table 13**

Onsets and offsets for analysis regions by prosodic form

| Prosody | Window | Onset (ms) | Offset (ms) |
|---|---|---|---|
| Active | 1 | 814 | 1016 |
| Active | 2 | 1403 | 1703 |
| Active | 3 | 1918 | 2068 |
| Passive | 1 | 710 | 910 |
| Passive | 2 | 1221 | 1521 |
| Passive | 3 | 1918 | 2068 |
| No cues control | 1 | 710 | 915 |
| No cues control | 2 | 1115 | 1415 |
| No cues control | 3 | 1812 | 1962 |
| All cues control | 1 | 814 | 1016 |
| All cues control | 2 | 1327 | 1627 |
| All cues control | 3 | 2024 | 2174 |

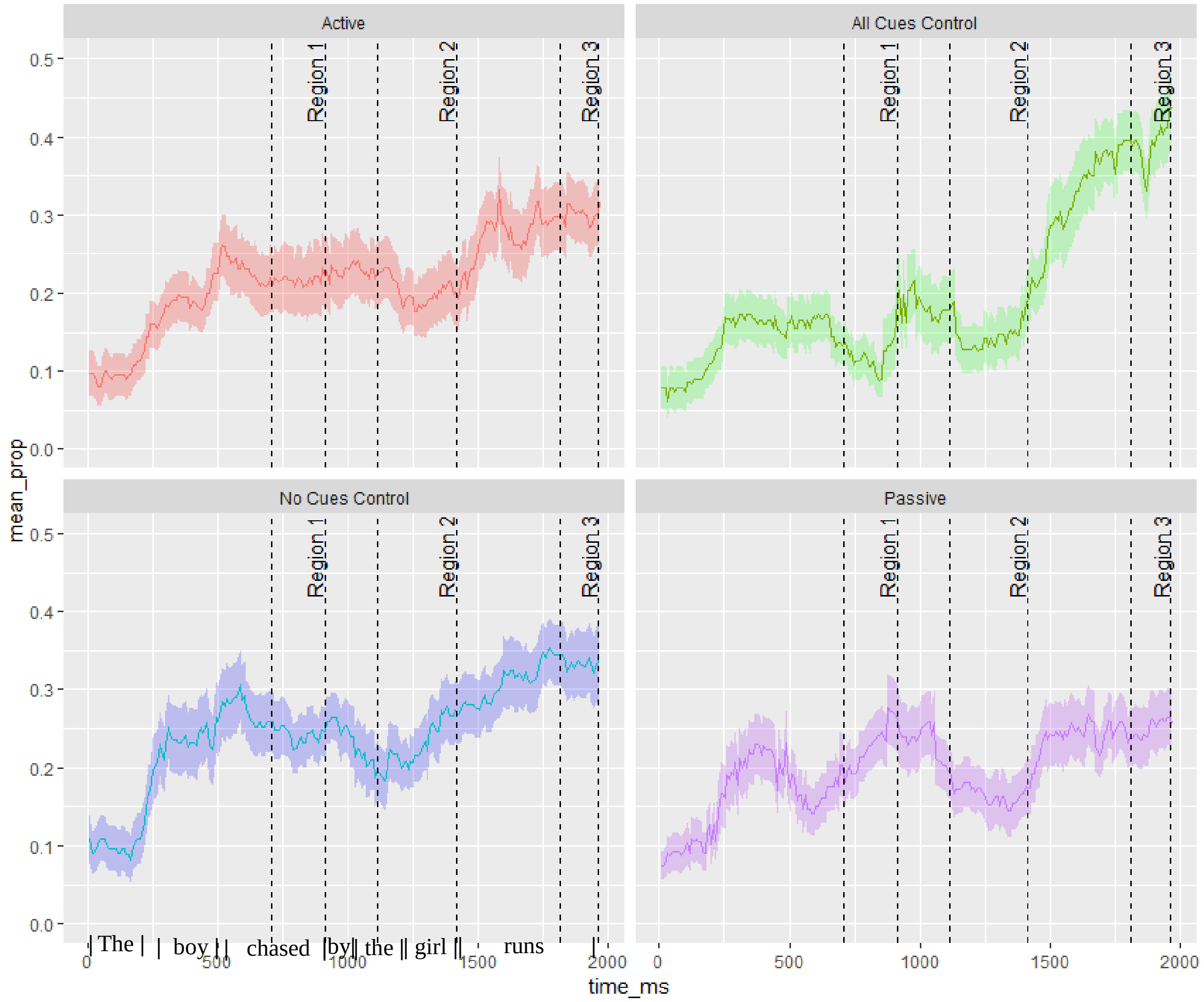


**Fig. 11**. Mean proportion of looks to target by prosodic condition, for passive structures. The coloured lines illustrate the mean, and the shaded area surrounding it illustrates the standard error. Analysis regions are indicated by dashed vertical lines, where the region label is at the region offset.

### *3.3.2.1 Region 1: The embedded verb phrase*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a four level factor with active cues as the baseline predictor), the time-bin in which the fixation took place (1 – 24, coded as numeric), and the interactions between these fixed factors.

First, we analysed the effect of time bin. Including the effect of time bin did not improve model fit ($\chi^2(1) = 2.547$, $p = .111$), indicating that participants did not make more looks to target as the verb phrase of the embedded clause unfolded.

Next, we added the main effect of prosodic condition, which did not improve model fit ($\chi^2(3) = 4.345$, $p = .227$), however it was retained in the final model (see table 14), wherein it indicated a negative effect of passive pause cues, relative to active cues. This reflects then tendency for passive pause cues to produce an initially lower number of target fixations than active pause cues at the onset of the analysis window, which only exceeded the active cues at its closure.

Finally, we included the interaction between prosodic condition, and time bin, which significantly improved model fit ($\chi^2(3) = 12.626$, $p = .006$), reflecting a greater increase in looks to target in the passive cues condition relative to active cues (see figure 12 for greater detail). Table 14 displays the final model outcomes.

The best-fitting model thus included the main effects of time bin, prosodic condition, and their interaction.

**Table 14**

GLMER model outcomes predicting target fixation likelihood in analysis region 1

| Fixed Effect | Estimate | Std. Error | *z* | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -1.762 | 0.804 | -2.190 | .029* |

| | | | | |
|---|---|---|---|---|
| Time-bin | 0.0007 | 0.006 | 0.113 | .910 |
| Prosody – No Cues Control | 0.158 | 1.130 | 0.139 | .889 |
| Prosody – All Cues Control | -0.389 | 1.226 | -0.317 | .751 |
| Prosody – Passive | -2.620 | 1.163 | -2.253 | .024* |
| Time-bin: Prosody – No Cues Control | -0.0003 | 0.009 | -0.035 | .972 |
| Time-bin: Prosody – All Cues Control | -0.008 | 0.010 | -0.770 | .442 |
| Time-bin: Prosody – Passive | 0.026 | 0.009 | 2.741 | .006** |

Model syntax: Target Fixation ~ (1 + Prosody* Time-bin| Subject) + (1 + Prosody* Time-bin| Item) + Prosody +Time-bin + Prosody:Time-bin, family = binomial(logit)

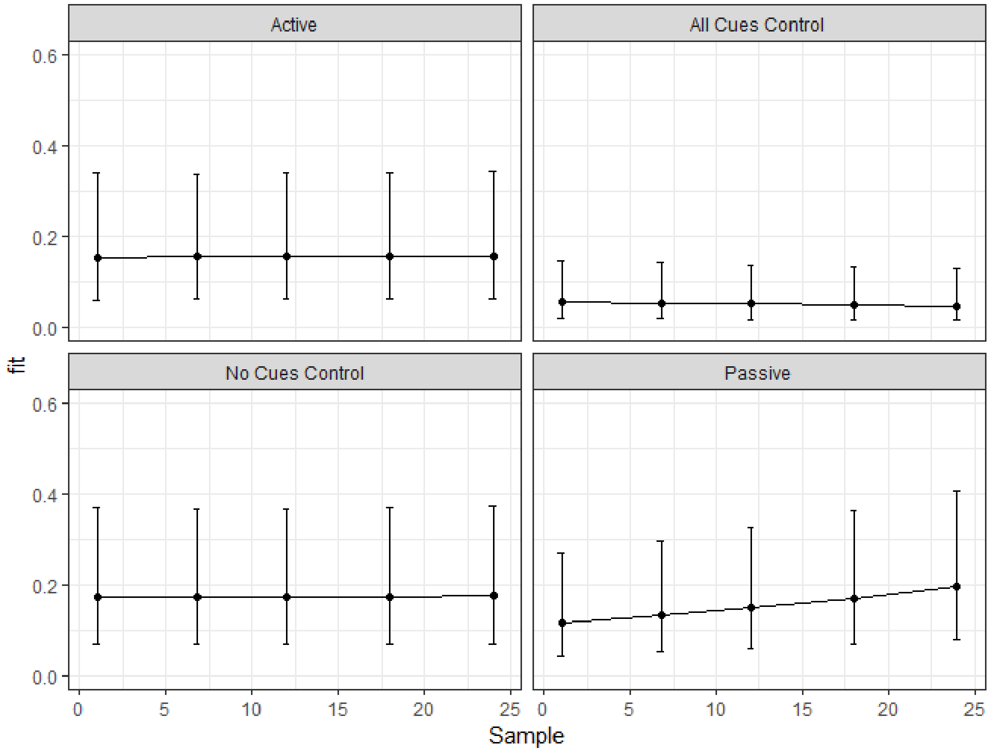


**Fig. 12**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

### *3.3.2.2 Region 2: The embedded noun phrase*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control, coded as a four level factor, with active as the baseline predictor), the time-bin in which the fixation took place (1 – 36, coded as numeric), and the interactions between these fixed factors.

First, we analysed the effect of time bin. The addition of time bin to the model resulted significantly improved model fit ($\chi^2(1) = 42.761$, $p < .00001$), reflecting that overall, the probability of fixating the target increased as the by phrase of the embedded clause unfolded (see Table 15 for the final model outcomes). Figures 13 and 11 illustrate that the increased probability of fixating the target was mainly confined to the active and no cues control conditions, and to a lesser extent, the all cues control condition.

Next, we analysed the effect of prosodic condition on fixation behaviour. The addition of prosodic condition did not improve model fit ($\chi^2(3) = 0.619$, $p = .892$).

Finally, we included the interaction between prosodic condition and time bin, which significantly improved model fit ($\chi^2(3) = 37.913$, $p < .00001$). This reflected that the increase in fixations to target across the analysis region was lesser in the passive and all cues conditions relative to the active condition, however the no cues condition saw a similar increase (see figure 13 for greater detail).

The model that best fit the data for the embedded noun-phrase thus included the main effects of time bin, prosodic condition, and their interaction.

**Table 15**

GLMER model outcomes predicting target fixation likelihood in analysis region 2

| Fixed Effect | Estimate | Std. Error | z | Pr(\|z\|) |
|---|---|---|---|---|
| Intercept | -4.776 | 0.670 | -7.128 | < .00001*** |
| Time-bin | 0.020 | 0.003 | 6.088 | < .00001*** |
| Prosody – No Cues Control | 0.120 | 0.901 | 0.134 | .894 |
| Prosody – All Cues Control | -2.361 | 0.948 | -2.490 | .013* |
| Prosody – Passive | -3.680 | 0.908 | -4.053 | .00005*** |
| Time-bin: Prosody – No Cues Control | 0.0007 | 0.005 | 0.166 | .868 |
| Time-bin: Prosody – All Cues Control | -0.015 | 0.005 | -2.920 | .004** |
| Time-bin: Prosody – Passive | -0.025 | 0.005 | -5.082 | < .00001*** |

Model Syntax: TargFix ~ (1 + Prosody*Time-bin) + (1 + Prosody*Time-bin|Item) + Prosody + Time-bin + Prosody:Time-bin

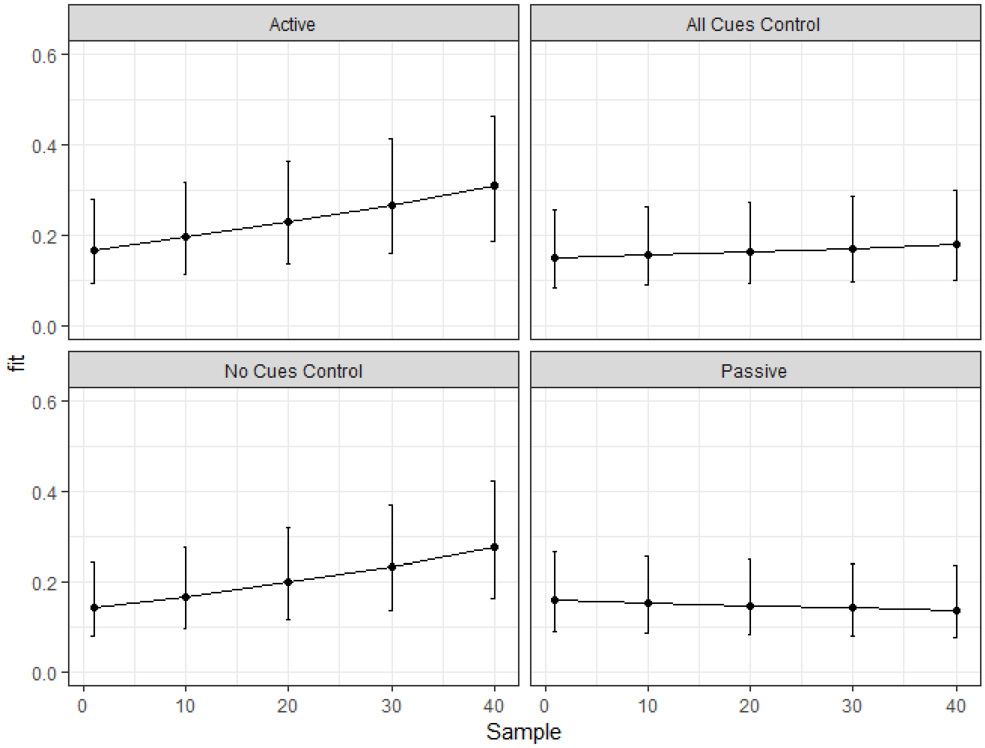


**Fig. 13**. Model estimates of the likelihood of fixating the target over the analysis region. Black points illustrate the model estimates, and back vertical bars illustrate 95% confidence intervals.

### *3.3.2.3 Region 3: The main clause verb*

To analyse participants' looking behaviour elicited by the auditory sentences, we conducted a generalised linear mixed-effects model analysis (GLMER) predicting the dependent variable of target vs. other fixation (1 vs. 0) using a binomial distribution with a logit-link function, split by analysis window. Random intercepts and slopes for participants and items were included in all reported analyses. As fixed effects, we tested the effect of the prosodic condition participants were exposed to (active, passive, no cues control, all cues control), the time-bin in which the fixation took place (1 – 18), and the interactions between these fixed factors.

First, we analysed the effect of time bin. The addition of time bin to the model did not improve model fit ($\chi^2(1) = 1.792$, $p = .181$).

Next, we analysed the effect of prosodic condition on fixation behaviour. The addition of prosodic condition did not improve model fit ($\chi^2(3) = 1.300$, $p = .729$).

Finally, we included the interaction between prosodic condition and time bin, which did not improve model fit ($\chi^2(3) = 0.877$, $p = .831$).

The best-fitting model for this region thus included only random effects; i.e. random by item and participant variance best explained the likelihood of fixating the target item.

### *3.3.2.4 Summary*

For passive structures, in the agent verb phrase, participants were initially less likely to fixate the target in the passive cues condition, however, over the course of the analysis region, participants became more likely to fixate the target, suggesting that a lack of a pause was beneficial in this location. In the analysis region encompassing the agent by-phrase, relative to the passive cues conditions, the active and no cues conditions produced a greater number of target fixations, both overall, and over time. This finding was intriguing; neither of these conditions contained a pause preceding the analysis region, suggesting that when these clauses are temporally proximate, it benefits performance. These results would seem to suggests that temporal proximity cues that suggest clausal groupings were poorly used by our participants for both active, and passive structures.

**Table 16**

*Summary of effects by Condition (Pitch Similarity vs. Temporal Proximity), Syntactic Form (active vs. passive), and Analysis Region (1, 2, 3).*

| Condition | Syntax | Analysis Region | Finding |
|---|---|---|---|
| Pitch | Active | 1 | No significant differences on the basis of prosody or time bin |
| Pitch | Active | 2 | Main effect of prosody demonstrated fewer target fixations for passive, no, and all cues conditions. The two-way interaction with prosody showed passive and both control conditions showed an increased number of fixations over the analysis region, while in actives it reduced. |
| Pitch | Active | 3 | Main effect of prosody indicated that passive cues resulted in fewer target fixations. The prosody by time-bin interaction showed that the number of target fixations increased in the passive cues condition, with similar performance to active by the end of the region. |

| | | | |
|---|---|---|---|
| Pitch | Passive | 1 | Main effect of prosody illustrated that active, passive, and no cues elicited a similar number of target fixations. Only the all cues controlled differed significantly, eliciting fewer target fixations. |
| Pitch | Passive | 2 | Main effect of prosody illustrated that passive pitch cues increased target fixations relative to active prosody. Both control conditions produced fewer target fixations than active prosody. The prosody by time bin interactions suggested that passive, no cues and all cues control conditions produced more looks to target over time than the active cues, which showed little change over time. |
| Pitch | Passive | 3 | Significant main effect of time bin reflects global trend for increased target fixations over this region. Main effect of prosody demonstrated that only the all cues control differed from active, with higher target fixations in this condition. |
| Pause | Active | 1 | Significant main effect of time-bin indicated a global reduction in target fixation over the analysis region. No main effect or interaction of prosodic structure. |
| Pause | Active | 2 | Main effect of prosody indicated a greater number of target fixations in the active condition from the onset to the middle of the window, however a significant two-way interaction indicated that the other control conditions increased over |

| | | | |
|---|---|---|---|
| | | | time, whereas the active cues condition showed a reduced number of target fixations. |
| Pause | Active | 3 | No significant main effects or interactions suggesting the performance in each condition statistically equivalent. |
| Pause | Passive | 1 | Main effect of prosody indicates that at the onset of the window, passive cues elicited a lower number of target fixations relative to active cues, however the two-way interaction indicated that the passive cues condition elicited more target fixations over time, showing more by the end of the analysis region. |
| Pause | Passive | 2 | Main effect of time-bin reflected tendency for the number of target fixations to increase across time. The two-way interaction of prosody and time bin illustrated an increase in target fixations across the window for active and no cues control conditions. Passive cues resulted in a small reduction, while the all cues control showed a slight increase. |
| Pause | Passive | 3 | No significant main effects or interactions, suggesting equivalent performance across all conditions. |

# 4. Discussion

The aim of this study was to assess whether auditory-perceptual Gestalt grouping cues consistent with speech production data of our participants' native language (Trotter, Frost, & Monaghan, Chapter 3) would facilitate the online processing of hierarchical syntactic structure. Frank, Bod and Christiansen (2012) argued that the processing of speech may be sequential, with individuals using superficial surface level cues to compute dependencies, instead of processing the incoming words in a hierarchy. Prosodic structure rapidly influences processing (Snedeker & Trueswell, 2003; Watson, Tanenhaus, & Gunlogson, 2008), suggesting that the temporal and pitch variance may provide a means through which dependencies could be computed. Trotter, Frost, and Monaghan (Chapter 3) found that in a corpus of spontaneously produced relative clauses, pitch similarity and temporal proximity provide reliable syntactic boundary information. Pitch similarity states that individuals are likely to form sequential links between sounds that occur in a similar pitch, while temporal proximity states that individuals are likely to form sequential links between sounds that occur closer in time. For active-object structures, the phrases of the embedded clause were more similar in pitch to one another than to the phrases of the main clause and were preceded by a lengthened pause. For passives, the results differed; the noun phrase of the main clause and the verb phrase of the embedded clause were more similar in pitch and followed by a lengthened pause.

To assess whether these cues support online speech processing, in the present VWP structure, we auditorily presented participants with active-object and passive relative clause structures whilst they viewed experimental scenes. We manipulated the prosodic structure of the speech to be consistent with active-object or passive relative clause production data, a no cues control (where no tonal or temporal grouping cues were present), and an all cues control (where grouping cues were added in locations consistent with both active and passive production data).

We hypothesised that participants would be more likely to fixate the scene described by the utterance when syntactic and prosodic form were congruent (e.g. active syntax and active prosody). Further, we predicted that participants would be most accurate when responding to trials with congruent syntax and prosody.

The behavioural results of the study did not fully support these predictions. In terms of response accuracy over both experimental conditions, trial number explained the greatest variance, suggesting that increased proficiency with the paradigm, or habituation to the synthesised speech, was the greatest determinant of accuracy. There were additionally trends towards increased accuracy with active-object relative clause structures, and reduced accuracy with passive relative clauses, in the pitch similarity condition, relative to the pause cues condition. These results suggest that prosodic grouping cues did not facilitate comprehension accuracy. Reflection-based tasks, e.g. comprehension accuracy, are less well-suited to assess processing-based learning, such as the online processing of speech (Christiansen, 2018; Frizell, O'Neill, & Bishop, 2017; Isbilen et al., 2018; Trotter, Monaghan, Beckers, & Christiansen, Chapter 2), thus, the role of Gestalt grouping cues should be more evident in participants' gaze behaviour.

The processing-based results of the study offered greater support for our predictions. Pitch similarity cues congruent with active structures (Trotter, Frost, & Monaghan, Chapter 3) resulted in an increased likelihood of fixating the target, that became apparent by the verb phrase of the embedded clause and endured until the response window. For passive structures, during the agent verb phrase, participants were more likely to fixate the target in the all cues control condition, which is notable, as in this time period, in terms of pitch prosody, this condition is indistinguishable from the active cues condition. In the subsequent agent verb phrase, there was an advantage for passive prosody, and the probability of making target fixations in this condition increased more over this time window than in the active condition. This suggests that pitch similarity cues consistent with passive pitch structures (Trotter, Frost, & Monaghan, Chapter 3) facilitated processing. By this stage, participants would have heard *The girl being chased*, thus, by this stage, providing participants had properly understood this structure, so increasing looks to target should be expected here. As such the pitch grouping may be beneficial in this location. However, in the main clause verb, only the high variance control condition resulted in a higher likelihood of fixating the target, suggesting that large pitch variability, and not specific grouping cues per se improved processing for passive structures.

The gaze behaviour in the temporal grouping cues conditions suggested some difficulty with using temporal grouping cues to process phrase structure. For active structures, during the embedded noun phrase, prosodic cues did not affect processing. Over the embedded verb phrase, no differences were present between active and passive cues, however both control conditions resulted in a lower probability of fixating the target, and the likelihood of fixating the target increased in these conditions, whereas it reduced in the active condition. Over the main clause verb, there was a trend towards a reduced likelihood of fixating the target in the passive cues condition, and no difference with either control. Therefore, participants were unable to use temporal proximity cues to process the active structures, however the presence of two lengthened pauses (all cues control), and a lack of pauses (no cues control), impaired processing. For passive structures, active pause cues produced a higher overall number of target fixations in the embedded verb-phrase region than passives, though it was equivalent to the no and all cues controls. However, as the analysis region unfolded participants became more likely to fixate the target in the passive cues condition when compared to active cues, suggesting that the lack of a pause was beneficial. In the agent by-phrase analysis region, the two-way interaction between time-bin and prosody reflected a larger increase in the active and no cues control conditions, suggesting that a lack of a temporal boundary in this location (i.e. when these phrases are temporally proximate) is beneficial to processing.

In terms of active syntax, the results for the pitch similarity condition confirmed our predictions; when pitch similarity was highest between the phrases of the embedded clause, participants were more likely to fixate the target image during both the embedded verb phrase and the verb phrase of the main clause. For passive syntax, the results were more mixed; the number of fixations were similar between active and passive conditions following the first boundary, suggesting that the presence, or absence of a pitch grouping did not greatly affect performance. A pitch boundary preceding the agent by-phrase, however, facilitated performance in for both the passive, and all cues condition, suggesting that the pitch boundary from Trotter, Frost, and Monaghan (Chapter 3), is useful for processing passive syntax. The authors explained this finding as an example of Ferreira's (2003) good enough processing account; if the noun and agent-verb phrase are grouped together ("the boy being chased"), then participants have produced enough information to disambiguate the scene, making this the most salient grouping to highlight. In the present study, the structure differs slightly. After the agent verb phrase, participants would only have heard "the boy chased". However, again, this grouping may have been the most salient. With chased being past tense, it confers an agent and patient role. Given the display composition, with this information, participants would be provided with enough information to eliminate the agent-patient role reversal, and the different agent-verb condition, leaving only direct competition between two scenes. As such, this grouping serves to eliminate the highest number of competitors, making it the most salient. Hence, reinforcing it with a pitch boundary should facilitate performance.

Temporal grouping cues provided less consistent results for both syntactic structures, however, with active congruent cues not consistently producing better performance for active structures, suggesting a temporary benefit in the region of the embedded verb phrase, and for passive structures, the benefit of temporal grouping cues seemed to suggest that increased processing time overall was beneficial. Overall, this suggests that, at least for actives, that participants can use temporal proximity cues, though they are less salient than pitch cues for processing. This may relate to our choice of temporal cue, namely, unfilled pauses. In previous studies of speech timing, it has generally been shown that in addition to lengthened unfilled pauses at syntactic boundaries, individuals also employ final-syllable lengthening (Snedeker & Trueswell, 2003). Final-syllable lengthening, further, has been shown in production studies to have a more consistent duration that unfilled pauses (e.g. Cooper, Paccia, & Lapointe, 1978). Ferreira (2002) argued that the particular durational cue that appears after a word will depend on its phonetic qualities, and notably, on these ground, many studies prefer to utilise relative durations measures. For example, Kraljic and Brennan (2005) used the combination of the previous noun phrase combined with an unfilled pause to assess the length of prosodic boundaries. Trotter, Frost, and Monaghan (Chapter 3) also assessed pause duration as a percentage of the entire utterance. Notably, in the present study, we employed simply the mean duration of critical inter-clause pauses, suggesting that the measure here may have been too course. Together, these observations suggest that durational cues may consist of a hierarchy, requiring more than simple, unfilled pauses to elicit consistent grouping preferences, and that durational cues here may have been too coarse to elicit natural temporal grouping behaviours.

The time course of the prosodic effects replicates Snedeker and Trueswell (2003). In their study, the effects of final lengthening and a lengthened pause were present 200ms following the onset of the subsequent noun phrase. Similarly, in the present study, the effects of prosody were generally evident 200ms following the lengthened pause, or pitch reduction. However, for active structures, the effect emerged later; for pitch cues, there were trends towards an effect in the embedded noun phrase, and for pause cues, no prosodic effects were evident in the same time window. This suggests participants minimally needed the verb phrase to disambiguate the scene, and prosodic information supports this grouping. In the present study, this seems likely; the agent and patient were consistent across all scenes, and the relative order in which they are presented in active sentences should provide only agent and patient assignment. As there was a scene with the agent and patient roles swapped, it suggests looks should be split between these scenes until the verb information was provided.

This study extends the findings of Trotter, Monaghan, and Frost (Chapter 5). In this AGL study participants were first trained on an artificial grammar modelling hierarchical centre-embedded structures ("The boy the girl chases runs", $A_1A_2B_2B_1$), and then performed a classification task on novel stimuli. Participants were trained with pitch similarity and temporal proximity cues modelled off an English-speaking corpus (Trotter, Frost, and Monaghan, Chapter 3). Participants exposed to pitch similarity cues were more accurate at classifying grammatical structures in which the violation was made salient by pitch grouping cues, which was supported by an analysis of sensitivity and response bias. In contrast, participants trained with temporal cues did not have increased sensitivity, reduced response bias, or increased accuracy for dependencies highlighted by temporal groupings. In the present study, temporal proximity cues were not as useful for the processing of hierarchically centre-embedded structures, as were pitch similarity cues. Reflection-based measures, such as classification tasks, may not be well suited for measuring processing-based tasks, such as the online comprehension of speech (Christiansen, 2018; Frizell, O'Neill, & Bishop, 2017; Isbilen et al., 2018). This raised the question of whether processing benefits present in Trotter, Monaghan, and Frost (Chapter 5) may have gone undetected, as the measure may not have been sufficiently sensitive to detect it. The temporal proximity, and pitch similarity results however, seem to support these conclusions; the benefits of temporal proximity cues were unreliable in comparison to pitch similarity cues. Thus, the overall findings across these two studies is pitch cues are salient, and support the grouping of dependent elements in complex, hierarchical structures.

Finally, it is sensible to raise the point that pitch-prosodic and temporal cues in speech may not be purely syntactically driven, which may explain why the effect of prosodic conditions uniform across the analysis windows, and not evident in the accuracy data. Syntactic factors are unlikely to be the only factor influencing prosody (Kraljic & Brennan, 2005); Ferreira (1993) argues that semantics can mediate the relationship between a syntactic representation and its articulation. Kraljic and Brennan (2005) suggest that pragmatic information could also influence prosodic lengthening if that information is available before articulation. This raises the question of whether conceptualising the role of prosody as solely intended to support syntactic processing is too narrow.

Specifying a robust account of prosody will require integrating contextual factors that will influence prosodic cues in everyday interaction. Simulation accounts, such as Pickering and Garrod's (2004) Interactive Alignment Theory, interlocuters converge across all levels of linguistic communication, from semantics, to syntax, phonetics and gesture in order to reduce the complexity of online speech processing, allowing dialogue partners to predict upcoming speech. Here, I will restrict the observations to phonetics for the sake of space. There is substantial evidence that interlocuters automatically imitate several aspects of one another's speech, including; accent, speech rate, intonation and speech style (Delvaux & Soquet, 2007; Webb, 1969; Goldinger, 1998; Shockley et al., 2004; Pardo et al., 2010). Any of these will correlate with prosodic boundaries and reflect one's audience. For example, if your interlocuter does not understand you, lowers their speech rate, increases pitch accenting on critical words, and pauses more often, you will likely imitate them. Alignment through imitation will likely affect many studies on temporal and prosodic cues, and presumably, pitch cues as well. Given these observations, it is likely that by attempting to disentangle prosody with tightly controlled task, that several, vital factors have been missed, or their contribution underestimated. To develop a mechanistic account of prosody, and how it affects listeners, it will be necessary to integrate these factors into our experimental designs, and asses prosody in more interactive settings.

In summary, in the current study, we aimed to assess whether temporal proximity and pitch similarity cues based on the participants' native language experience would facilitate the online processing of reduced active-object and passive relative clauses. This was based on the idea that if hierarchical sentences can be processed sequentially, then individuals may compute dependencies based on auditory perceptual Gestalts. The results of this study partially confirmed these claims; for active structures, participants showed a higher likelihood of fixating the target when exposed to pitch cues consistent with active structures, however, participants were less able to use temporal proximity cues. For passive structures, in the pitch condition, participant performance was better when they were exposed to cues consistent with active structures. Overall, it thus appears that for hierarchical structures, it is beneficial for processing when pitch similarity reinforces the dependency between the phrases of the embedded clause ("The boy [pitch reduction] the girl chases runs", "The boy [pitch reduction] chased by the girl runs"). We therefore conclude that participants can effectively use low-level Gestalt grouping strategies to assist online, sentence processing.

# 7. General Discussion

This thesis assessed whether auditory Gestalt processing may obviate the need to process hierarchically structured speech hierarchically. Hierarchical syntactic structure has long been a central focus of psycholinguistics, due to its theoretical importance. Hierarchical structures can be challenging to process, even for native speakers (Bach, Brown, & Marslen-Wilson, 1986; Gibson & Thomas, 1999), though they remain present in natural language (e.g. Karlsson, 2007). As a result, probing the manner in which they are correctly (or incorrectly) processed offers insights into the cognitive mechanisms underpinning language processing. As a way of classifying these mechanisms, Chomsky (1957; 1959) proposed a generative hierarchy of rule systems capable of producing an infinite set of sequences by defining increasing constraints on possible linguistic structures. Finite state grammars occupy the lowest level of this hierarchy, and can be fully specified by transitional probabilities between a finite number of states (Hauser & Fitch, 2004). Processing finite state sequences necessitates a large enough memory stack to hold sequential states, and the transitions between them, in order to concatenate them into longer sequences. Phrase structure grammars can similarly concatenate items, but can additionally embed strings within other strings through the recursive application of the *merge* operation (Chomsky, 1995), resulting in phrase structures and long-distance dependencies. The presence of long-distance dependencies thus requires an open-ended memory system to maintain dependent elements whilst processing intervening material (e.g. Karlson, 2010; Gibson & Thomas, 1999), and the perceptual mechanism necessary to recognise that distant elements are related (Hauser & Fitch, 2004).

Phrase structure grammars (Hauser, Chomsky, & Fitch, 2002), and more recently, the *merge* operation (Berwick & Chomsky, 2016) have been stated to be the primary, essential feature of human language, allowing for an infinite set of meaning to be expressed from a finite number of words. The *merge* operation combines linguistic items, e.g., *the* and *girl*, to create composite terms (*the*, *girl*), that can be combined with another term, *runs*, to form ((*the*, *girl*), *runs*); the recursive application of *merge* thus results in hierarchical structures (Yang, Crain, Berwick, Chomsky, & Bolhuis, 2017).

Crucially, these theoretical advances all highlight hierarchical structures generated by phrase structure grammars as evidence that language is not a finite state system, and thus present a minimal set of mechanisms that individuals must possess for their production. However, this thesis deals not with the production of hierarchical structure, but rather its comprehension. Whilst phrase structure may be necessary for describing production, it is less clear whether individuals actually process hierarchical structures hierarchically when parsing incoming speech.

Sequential processing accounts (Frank, Bod, & Christiansen, 2012) suggest that individuals rely on surface level grouping cues to rapidly determine a sentence’s dependencies, instead of processing the incoming words as part of a hierarchy. For example, in the hierarchically centre-embedded sentence, “The ligament the surgeon repaired was torn”, the listener can use world knowledge to identify the dependencies; surgeons operate on injured people, and unlike the surgeon, ligaments can tear.

In this thesis, I examined whether pitch and temporal variance in speech may be sufficient to trigger grouping according to the pitch similarity and temporal proximity Gestalts. The principle of pitch similarity states that individuals will form sequential links between sounds that occur at a similar pitch, while the principle of temporal proximity states that sequential groupings will be formed between sounds that occur closer together in time (Deutsch, 2013). The pitch and temporal structure of speech are plausible candidates for cues that support sequential grouping of dependencies due to the speed at which durational (Snedeker & Trueswell, 2003) and pitch (Watson, Tanenhaus, & Gunlogson, 2008) structure influence processing, as well as their domain-general role in grouping auditory sequences (e.g. Hamaoui & Deutsch, 2010). To investigate whether the pitch and temporal structure of speech facilitates the processing of hierarchical structure, we adopted a multi-methods approach. First, we assessed the various cues and structures that are learned in artificial grammar learning experiments using meta-analytic techniques. Next, we assessed whether spontaneously produced speech contains pitch similarity and temporal proximity cues consistent with syntactic structure. To assess whether these cues were useful in acquisition and comprehension, we conducted two artificial grammar learning studies, and one visual world paradigm study to assess the utility of these pitch and temporal cues in comprehension

## 7.1 Methodologies for studying speech processing

In Chapter 2 (Trotter, Monaghan, Beckers, & Christiansen), I sought to assess what participants actually acquire in artificial grammar learning studies conducted across species. The results indicated evidence of learning artificial grammars, though effect varied by species. Adult humans had the largest effect, with human children and non-human mammals having significant effects, though not birds. Human adults were found to perform similarly between reflection- and processing-based tasks, though this likely reflected the far larger number of reflection-based tasks in the sample. This effect was surprising in light of compelling evidence suggesting that processing-based measures are better suited for processing-based tasks, such as the online processing of speech (e.g. Christiansen, 2018; Isbilen et al., 2018). For birds, the presence of training items at test produced large effects, though a larger amount of training produced lower effects, and further, a larger vocabulary produced larger effects, reflecting the large effects seen in studies with larger vocabulary sizes. These results demonstrated that surface level features of the language that increase its complexity can have different effects across species.

## 7.2 Speech corpora: Elicitation, and reading aloud paradigms

Chapter 3 (Trotter, Frost, & Monaghan) detailed a speech corpus study on spontaneously produced active and passive relative clauses. Here, we found evidence to suggest that for active-object structures (e.g. “[The boy] [the girl] [chases] [runs]”), pitch similarity and temporal proximity cues provide grouping information consistent with syntactic boundaries. Relative to the phrases of the external clause, the phrases of the internal clause were similar in pitch, and closer together in time (e.g. “[The boy] pause/pitch reduction [the girl] [chases]”). This was not the case for passives, where pitch similarity and temporal proximity suggested groupings that were inconsistent with syntactic structure. Here similarity and proximity were highest between the first phrase of the main and embedded clauses (e.g. “[The boy] [being chased] pause/pitch reduction [by the girl]”). We suggested that for passive structures, this may reflect a least effort principle, or “good enough” processing (Ferreira, 2003), in that by the time individuals had provided the patient noun-phrase (“the boy”) and agent verb phrase (“being chased”) speakers had provided sufficient information to disambiguate the scene (in contrast to actives), making it efficient to group these phrases. The results for actives, however, suggested that the acoustic structure of speech contains cues sufficient to group its dependencies using Gestalt mechanisms. These findings are in line with previous speech production studies, in which we believe the results are consistent with the view that Gestalt mechanisms could facilitate the grouping of dependencies.

Notably, Fery and Schubö (2010) found that German native speakers, produce the constituent phrases of embedded clauses of hierarchical centre embedded structures in a similar pitch, reducing pitch between levels of embedding. This was the case even after the offset of the deepest level of embedding, where participants would increase their pitch to the level of previous phrase of the second level of embedding; pitch variance within a level of embedding was low, but high between levels.  Thus, it appears that in German, Gestalt processes could also be used for dependency detection. Further corpus studies should be conducted in future to assess whether these findings generalise across languages. Similarly, in English, there are syntactically driven pitch reductions (Cooper & Sorensen, 1977), lengthened pauses, and syllable lengthening (Cooper, Paccia, & Lapointe, 1978) at clausal boundaries, which I suggest indicate that Gestalt processes may assist in the grouping of dependencies more broadly. Whilst these production studies are important, their use of reading aloud methods questions their ecological validity.

Chapter 3, in contrast, represents an important, more ecologically valid extension of these studies; it assessed the presence of these cues in spontaneously produced speech. Whilst the current results are generally consistent with Fery and Schubö (2010), Cooper and Sorensen (1977), and Cooper, Paccia, and Lapointe (1978), in future studies seeking to assess prosodic features of speech, we recommend adopting the approach used here; generate corpora of spontaneously produced speech with elicitation paradigms (e.g. Montag & MacDonald, 2014), and conduct acoustic analyses on this data. Whilst finding the presence of prosodic cues suggests they may have utility in comprehension, to prove this requires systematically manipulating their presence in subsequent experimental work. Chapter 4 (Trotter, Frost, & Monaghan), 5 (Trotter, Monaghan, & Frost), and 6 (Trotter & Monaghan) therefore assessed these cues experimentally.

### 7.3 Pitch similarity and phrasal groupings

Across these experiments, there was evidence for the utility of pitch similarity cues, but weaker involvement of the use of temporal proximity cues. In two AGL studies, we assessed the utility of temporal proximity and pitch similarity cues for acquiring hierarchical centre embedded structures in an artificial language by adjusting the salience of each respective cue. In one version, Trotter, Frost, and Monaghan (Chapter 4) utilised stronger durational cues, whilst in Trotter, Monaghan, and Frost (Chapter 5), pitch cues were made relatively more salient, by reducing the duration of all pauses, and increasing the distinction between $F_0$Hz values for each level of embedding, based on the speech corpus analysis results of Trotter, Frost, and Monaghan (Chapter 3). In Chapter 4, participants exposed to pitch cues became more accurate at classifying grammatical structures correctly relative to baseline, whilst remaining below chance performance for ungrammatical structure. This suggested that the presence of pitch cues globally increased the plausibility of all test structures (as the additional acoustic cues were present over training and testing), creating a bias to classify any structure as grammatical. Thus, whilst pitch cues were salient to participants, they served to mask local, grammatical violations. This theory was not confirmed by an analysis of participants' response bias, which was high across all conditions, and highest for the combined cues conditions. In Chapter 5, this was not the case. Participants exposed to pitch similarity cues were overall more accurate at detecting grammatical violations in the fifth sequence position ($A_1A_2A_3B_3\underline{B_5}B_1$). Critically, these violations are highlighted by pitch similarity, unlike violations in the sixth position, where there was no increase in accuracy with increased exposure. An analysis assessing participants sensitivity to the grammatical structure and response bias reinforced these findings; participants were more sensitive towards the grammar in LoE 2 sequences and were less biased towards classifying sequences as grammatical responses. Taken together, these studies suggested that pitch similarity cues were salient to our participants, even though they did not uniformly increase performance across the whole sequence.

To verify these findings, Trotter and Monaghan (Chapter 6) conducted a follow-up visual world paradigm study. In this study, participants were tasked with identifying the target image of an active or passive relative clause in the presence of three distractors, each of which portrayed an agent and patient interacting. The prosodic structure of each sentence was manipulated to have a tonal or temporal boundary following the first (active congruent), second (passive congruent), following both (high variance control), or no boundaries (low variance control). The pattern of fixation reinforced the findings of Trotter, Frost, and Monaghan (Chapter 3), with pitch similarity congruent with the syntactic structure biasing looks towards the target during processing. Thus, in active-object relatives pitch similarity cues that increase the salience of the dependency between the phrases of the embedded clause, i.e. active congruent pitch similarity cues (e.g. "[The boy] pitch reduction [the girl] [chases] [runs]"). Thus, it appears that for hierarchically organised structures with an embedded phrase, if pitch cues are sufficient to support grouping according to Gestalt principles, then participant performance is improved. Thus, auditory-Gestalt processing of pitch may provide a mechanism through which non-hierarchical processing of hierarchical structures could be achieved.

### 7.4 Pitch similarity: Limitations and future directions

The benefits of pitch similarity are consistent with sequential processing accounts (Frank, Bod, & Christiansen, 2012), in that dependent phrases are more similar in terms of pitch, which supported processing of both syntactic forms (Trotter & Monaghan, Chapter 6). However, there are limitations to the studies reported here. Notably, we focussed on only two syntactic structures, active-object and passive relative clauses, and only used natural language in one of the experimental studies. Future work should seek to assess the presence of pitch similarity cues in corpora of speech including other syntactic structures, such as high- and low-attachment relative clauses. For example, in the sentence “Don mentioned the servant of the actress who was on the balcony”, a high-attachment interpretation would be that the servant is on the balcony, whereas a low-attachment interpretation is that the actress is on the balcony (Scheepers, 2003). In the high-attachment example, a Gestalt processing account would predict pitch similarity to be highest between “the servant”, and “who was on the balcony”. In the low-attachment example, however, pitch should be dissimilar between “the servant” and “the actress who was on the balcony”. Contrasting the pitch cues in these studies would offer insights into whether pitch similarity is useful for processing dependencies in different syntactic structures.

A similarly informative contrast would be between object- (“The lawyer that the banker irritated filed a hefty lawsuit”) and subject-relative (“The lawyer that irritated the banker filed a hefty lawsuit”) clauses. In the object-relative clause, pitch similarity should be highest between “the banker” and “irritated”, consistent with the results of Chapter 3 (Trotter, Frost, & Monaghan). However, in the subject-relative clause, “the lawyer that irritated the banker” functions as a complex noun-phrase, and “filed a hefty lawsuit” as the verb-phrase. Hence, under a Gestalt processing account, it would be expected that pitch similarity would be high across the entire sentence, with no large pitch reductions. While, currently lacking the ability to determine whether this is the case, this thesis generates testable hypotheses for future studies assessing whether prosodic cues in speech may facilitate grouping dependencies using Gestalt principles.

Assessing the presence of pitch similarity cues in alternative structures will be important for judging whether Gestalt speech cues assist processing broadly, and critically, allow for insights into the degree that speech processing may be driven by general, cognitive mechanisms. Finally, we also recommend repeating these experiments again with real speech; we cannot entirely rule out that our pitch cues may have provided a unique benefit to processing synthesised speech. This seems unlikely given that the results of Trotter and Monaghan (Chapter 6) did not show any overall accuracy differences between the low-variance (equal pitch over the sentence) and the active and passive prosody for either syntactic form. This potential issue, however, can only be resolved with further empirical work.

## 7.5 Temporal proximity and phrasal groupings

Overall, there was less evidence supporting participants' ability to use temporal proximity to group syntactically dependent phrases. In Trotter, Frost and Monaghan (Chapter 4), there was a null effect of pause cues, despite their emphasis in this AGL study. Similarly, in Trotter, Monaghan, and Frost (Chapter 5), where temporal proximity cues were reduced in salience, participants did not become more accurate, sensitive, or have reduced response bias at any point in testing. Participants were less able to effectively use temporal cues in Trotter and Monaghan (Chapter 6), in the context of an eye-tracking study. For active structures, participants were not more likely to fixate the target whether they were provided with active or passive temporal structure, though active congruent temporal structure did elicit more looks to target than both control conditions. For passive structures, gaze behaviour was mixed, but overall results suggested that increased processing time, regardless of the location of pause, improved performance. Taken together, these studies thus suggest that whilst participants do receive some benefit from pause cues, it is to a lesser degree than pitch similarity, and we have little evidence to claim there is benefit when temporal boundaries are at syntactic boundaries, or elsewhere in the sentence.

## 7.6 Temporal proximity: Unfilled pauses as an insufficient grouping cue

Is it the case that participants are not as sensitive to temporal proximity as pitch similarity? There are a few empirical reasons to assume this may be the case for English speakers. Fernald and McRoberts (1996) analysed durational cues at the ends of sentences and clauses, and found that 50% of all lengthened pauses in their sample occurred at non-syntactic boundaries, such as between two words that are not separated by a boundary. When pauses at syntactic boundaries do occur, their duration is highly variable; relative to final syllable lengthening, Cooper, Paccia, and Lapointe (1978) found that unfilled pauses following the syllables widely varied, and did tend to differ on the basis of the level of the syntactic hierarchy they lay at. Indeed, Trotter, Frost and Monaghan (Chapter 3) found the duration of pauses as highly variable, and that their likelihood of occurrence was not predicted by syntactic form or location within the sentence. This has led some to view pauses as reflecting cognitive load – and not syntactic features – in English (Goldman-Eisler, 1972). As a result, it has been suggested that pauses do not reliably correlate with syntactic boundaries, unlike pre-boundary lengthening (Martin, 1970).

Literature on the prosodic bootstrapping hypothesis (Gleitman & Wanner, 1982; Morgan, 1986; Peters, 1983) suggests that the results for temporal proximity may be explained by the unreliability of pause cues in English leading to a language-specific, low cue weighting. In a series of experiments, Seidl (2007) determined that 6-month-old infants were sensitive to prosodic boundaries in the absence of pause cues (experiment 2), however, they were insensitive to pauses when pitch cues were removed (experiment 3). In contrast, Männel and Friederici (2009) found that for German acquiring infants, pauses at prosodic boundaries were necessary to elicit the closure positive shift – an event-related potential that is reliably evoked at the close of a prosodic phrase. The varying sensitivity of pause cues may therefore reflect language-specific factors: German has a larger number of inflections and a flexible word order, suggesting that the functional demands on pitch may be greater for English speakers in highlighting phrase structure (Männel & Friederici, 2009).

This view is potentially consistent with usage-based accounts (e.g. Christiansen & Chater, 2016) of language processing. In terms of syntax, English native speakers typically rate hierarchical centre-embedded utterances with a missing a verb-phrase (e.g. "The patient who the nurse who the clinic hired met jack") as grammatically acceptable (Gibson & Thomas, 1999). Dutch (Frank et al., 2015) and German (Vasishth et al., 2010) speakers, in contrast, find their grammatical counterparts easier to process. In Dutch and German, Verb-final constructions are common, and require the listener to track dependency relations over long distances, suggesting that experience results in language-specific processing improvements (Christiansen & Chater, 2015). Thus, if language-specific experience suggests pitch cues are functionally important, and unfilled pauses are not, usage-based accounts suggest that cue weighting of pauses will reduce, in turn, reducing their salience.

Under a Gestalt processing account, the variable duration of pauses is not problematic – the source of the grouping cue is irrelevant, only that it is present. In contrast, their reliability and cue weighting are. If unfilled pauses are unreliable, and have a low cue weighting as a result, participants may simply fail to attend to them. This thesis reinforces the suggestion that, in isolation, pause cues are insufficient to elicit grouping behaviour.

## 7.7 Temporal Proximity: Future Directions

In future work, to further examine this problem in greater detail, I would recommend utilising an experimental paradigm which explicitly manipulates the degree of both final syllable lengthening, and unfilled pauses at syntactic boundaries in the context of comprehension. For example, if a study required participants to choose between two interpretations of syntactically ambiguous sentences such as those in Cooper, Paccia, and Lapointe (1978), (e.g. "Pam asked the cop who Jake confronted", (a) "Who did Jake confront?", (b) "Which cop? The cop that Jake confronted?"), where the length of (/ka/) in "cop" (syllable lengthening) and the following pause were manipulated, it might allow greater insights into the role of both cues. It would be wise to incrementally increase both variables, allowing insights into whether either are necessary, sufficient, or neither. However, we have insufficient evidence to disambiguate either possibility, due to only manipulating the length of unfilled pauses in the present work.

## 7.8 Domain-General Vs. Domain-Specific Processing

One aspect of our results which is difficult to reconcile is the increased utility of pitch Gestalt cues compared to temporal proximity Gestalt cues, when compared to the music processing literature. For example, Lerdahl and Jackendoff (1983) proposed that musical grouping boundaries are placed at longer intervals between note onsets (pauses) and at changes in values of attributes including the pitch range, which Deliège (1987) verified for Western Classical music, where participants were most likely to place groupings following long notes. Similarly, Hamuoui and Deutsch (2010) found that pauses become a stronger grouping cue the longer they are, overpowering hierarchically structured pitch similarity cues. In the results presented in Chapters 4, 5, and 6, however, participants did not seem to elicit as much of a processing benefit from temporal proximity as they did for pitch similarity cues. Why should there be this apparent disconnect between the results across domains?

At the level of acoustic processing, we should assume that these cues should be readily available and useful; the tonotopic organisation of the auditory cortex (Pantev, Hoke, Lehnertz, Lutkenhoner, Anogianakis, & Wittkowski, 1988; Elberling, Bak, Kofoed, Lebech, & Saermark, 1982; Tiitinen, Alho, Huotilainen, Ilmoniemi, Simola, & Naatanen, 1993; Yamamoto, Uemaura, Llinas, 1992; Yamamoto, Williamsen, Kaufman, Nicholson, Llinas, 1988; Bertrand, Perrin, Pernier, 1991), and hemispheric specialisation of temporal and spectral processing (Flinker, Doyle, Mehta, Devinsky, & Poeppel, 2019) indicate that bottom-up projections from primary auditory areas should bias processing early and effectively regardless of domain. Whilst this can explain the cross-domain applicability of auditory grouping cues, it does not take into account top-down processing.

Considering the differences between language music may therefore be readily explained. Zhiang, Jiang, Zhou, and Yang (2016) note that musical structure confirms to hierarchical structural rules (musical idioms), and Patel (2003) notes that musical phrases are marked by pauses, differences in tone height, and the durations of beats; both music and language are thus reliant on similar grouping cues (Patel & Iverson, 2007). However, music differs in a few regards; as music is purely a system of sound relationships, music is ultimately reliant upon them, whereas in language, formal syntactic and semantic relationships are critical to understanding, so acoustic cues can be unreliable, without preventing communicative success. We are able to detect and correct errors in speech (Nozari, Dell, & Schwartz, 2011), and still successfully communicate. For example, the ill-formed prosodic utterances in Nazzi et al. (2000), such as "*…leafy vegetables… Taste so good*" may be recognised by adult speakers as a disfluency and repaired on the basis using context. In music, however, these cues have a higher weighting in processing; a missed note, or a note from a different key will presumably be more disruptive. Thus, we hypothsise that cues will receive different weightings, resulting in the competition in Hamuoui and Deutsch (2010) resolving in favour of pause cues as pause length increases. It would be interesting to assess in further work whether competing temporal proximity and pitch similarity would have produce similar performance in linguistic stimuli.

Cross-linguistic differences in cue weighting are a more difficult question, though I would hypothesise that they are likely to reflect the cultural transmission processes of language. Christiansen and Chater (2016) describe language evolution as language change over an iterative chain of language acquisition and language use. Language change refers to processes such as reduction, where frequently used items tend to become reduced (e.g. *god be with ye*, to *goodbye*) and syntacticisation, whereby loose discourse sequences such as sequences such as *He pulled the window and it opened* become reduced to rigid syntactic constructions, such as *He pulled the window open*. These processes are believed to result from incremental or chunk-based processing; *He pulled the window and it opened* describes a single event, described by a relatively complex, two-event structure. Due to this, the result becomes syntactically reduced into a single, syntactic construction. On the other hand, reduction is constrained; reduction decreases effort for the speaker, but increases effort for the listener, therefore reduction only occurs to the degree that it does not damage communication. The other side of language change is that sequences that are difficult to produce or understand will disappear from language use. Language change is the result of multiple competing factors, deriving from factors affecting processing and acquisition, leading to linguistic diversity (Christiansen & Chater, 2016). Domain-general processing constraints, such as the neural architecture of the auditory system, will likely constrain the set of possible languages, but not necessarily determine how those languages use those domain-general processes.

Building on this, I believe usage-based constraints language change reflects why different populations may have different cue weightings for various perceptual grouping cues. Previously, I have noted the findings of Seidl (2007), who found that English acquiring infants are sensitive to pitch boundaries in the absence of pauses but cannot detect prosodic boundaries when pitch cues are absent. In contrast, Männel and Friederici's (2009) results indicated German acquiring infants require pauses to detect prosodic boundaries. The authors explained these findings as reflecting language-specific factors; German has a large number and a flexible word order, in contrast to English, hence the functional demands for pitch prosody may be of greater importance for highlighting phrasal structure in English. Based on these arguments, I believe therefore that it is likely that factors driving language change could explain cross-linguistic differences in the use of durational and pitch cues. For example, in tonal languages such as Thai, we might expect temporal cues to be of greater importance for prosodic groupings. Phonetic differences affect the use of final-lengthening or unfilled pausing (Ferreira, 2002), hence cross-linguistic variation across languages could be expected to produce different reliance on durational prosodic cues. The use of cues within each language are also likely to reflect what cues are useful during acquisition; in German, durational cues that are easily detected without top-down, language-specific knowledge will be retained in language, whereas in contrast, in English, easily detected pitch cues will be retained, as infants will be able to detect them without change. Whilst it remains beyond the scope of this thesis to make conclusive arguments regarding this issue, it remains a fruitful topic of investigation for future work.

### 7.9 Stimulus Limitations

A notable limitation Chapters 4, 5, and 6 must be addressed here; the use of artificial stimuli. While synthesised speech offers a unique degree of control over the speech stimulus, it cannot be claimed that it is natural, in most cases, participants describe it as sounding "robotic". In the present thesis, we do not view this as particularly problematic; to establish the baseline utility of these cues, stripping away natural variation from the stimuli was useful. After all, if the cues are useful with highly artificial, robotic speech, it stands to reason that they will be effective in real speech. There is, however, reason to question this assumption.

In Chapters 4 and 5, we noted that including pitch similarity and temporal proximity cues may have produced greater response bias. Similarly, in Chapter 6, we raised the issue of whether improved performance by experimental block may have reflected increased familiarity with the synthesised speech. It could well be the case that any performance does not reflect learning or processing per se, only that including natural speech cues are more efficient at retaining participants' attention. In this case, comparing a no cues (Chapter 6) control, or "baseline" prosody (Chapters 4, 5) may not provide a pure measure of the cue utility, but rather how attention may interact with processing. Indeed, baseline prosodic conditions may not be an adequate control to compare prosodic manipulations against. Using MEG, Herrmann, Friederici, Oertel, Maess, Hahne, and Alter (2003) found right-lateralised activation consistent with pitch-prosodic processing while they processed stimuli that had their pitch-prosodic cues removed by flattening the pitch contour. The authors interpreted this as suggesting that the brain generates its own prosody when it is absent during speech processing. Provided this argument holds true, this suggests that a-prosodic conditions do not provide an informative control and may to some extent explain why performance was similar in some cases to the prosodic manipulations in Chapter 6. In future work, therefore, it may be wise to implement controls wherein the prosody is unstructured (e.g. random rise-fall patterns within phrases) or designed to directly contradict syntactic structures. Given these observations, we cannot claim to have purely be measuring the utility of these prosodic cues in processing and acquisition, but also attention, and to some degree, implicit prosody generated by the brain.

In the present thesis, we cannot distinguish whether either the artificial speech or a different baseline would have been affected the pattern of results, as we have no studies where we implemented natural speech. In future work, however, I would strongly recommend that natural speech is employed wherever possible. Furthermore, it is clear that we should approach a more nuanced approach for developing control conditions; it may be more effective to implement uninformative, or scrambled prosodic contours, or utterances that straddle syntactic boundaries (e.g. "Leafy vegetables taste so good" vs. "…leafy vegetables. Taste so good"; Nazzi et al., 2000). Without doing so, it makes any conclusions that we draw from comparisons with control conditions unclear. Given that this thesis set out to establish the utility of Gestalt cues in processing of speech, however, we believed that employing artificial speech was methodologically justified and allows us to generate hypotheses for future work. However this caveat, combined with our use of a-prosodic controls suggests that the results should be interpreted with a degree of caution.

## 7.10 Conclusions

Thus, in short, this thesis sought to examine whether the structure of speech could be processed with auditory Gestalt mechanisms, and whether this would facilitate hierarchical dependency detection. Three experimental studies assessed the role of pitch similarity and temporal proximity cues taken from a corpus of spontaneously produced relative clause structures. Taken together, the results suggested that participants were unable to effectively use temporal proximity to group syntactically dependent elements, and that any benefits of temporal structure were unreliable. The temporal proximity results suggest that the application of Gestalt processing of speech is nuanced, potentially requiring the overlap of several, overlapping durational cues (e.g. unfilled pauses and final syllable lengthening). This is potentially troubling for a Gestalt processing account, and thus future work should probe the role of a combination of syllable lengthening and unfilled pauses, as opposed to only examining latter in isolation. On the other hand, participants found pitch cues salient, facilitating learning of artificial grammars, and the disambiguation of complex scenes. We therefore suggest that superficial pitch cues generated by the speaker during production can be processed effectively using the pitch similarity Gestalt, facilitating the rapid grouping of dependencies that does not rely on processing the incoming words in a hierarchy. Speech processing must be fast, online and robust to noise and variation (Christiansen & Chater, 2016). Applying low-level auditory processes to support this process of comprehending speech is one way by which this may be accomplished.

## 8. Compiled Bibliography

Abe, K., & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience, 14(8),* 1067-1076.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419 – 439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73,* 247 – 264.

Arlinger, S., Elberling, C., Bak, C., Kofoed, B., Lebech, J., & Saermark, K. (1982). Cortical magnetic fields evoked by frequency glides of a continuous tone. Electroencephalography and Clinical Neurophysiology, *54(6),* 642-653.

Audactiy [Computer Software]. (2016). Retrieved from https://www.audacityteam.org/download/

Baayen, R. H. (2008). *Analyzing Linguistic Data. A practical introduction to Statistics Using R.* Cambridge, UK: Cambridge University Press.

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*, 390-412.

Bach, E., Brown, C., & Marslen-Wilson (1986). Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes, 1(4),* 249 – 262.

Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial grammar processing engages Broca's area. *Neuroimage, 42*(2), 525-534.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13,* 99 – 102.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68(3),* 255 – 278.

Bates, D. M., Maechler, M., & Bolker, B. (2011). Lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-40.

Beckers, J. L., Berwick, B. C., Okanoya, K., & Bolhuis, J. J. (2012). Birdsong neurolinguistics: Songbird context-free grammar claim is prememature. *Neuroreport, 23(3),* 139-145.

Beckers, J. L., Berwick, R. C., Okanoya, K., & Bolhuis, J. J. (2017). What do animals learn in artificial grammar studies? *Neuroscience and Biobehavioral Reviews, 81(Part B),* 238-246.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3,* 15-70.

Berwick, R. C., & Chomsky, N. (2016). Why Only Us: Language and Evolution. Cambridge, MA: MIT Press.

Bertrand, O., Perrin, F., & Pernier, J. (1991). Evidence for a tonotopic organization of the auditory cortex observed with auditory evoked potentials. *Acta Oto-laryngologica, 491*, 116-122.

Bilecen, D., Scheffler, K., Schmid, N., Tschopp, K., & Seelig. J. (1998). Tonotopic organization of the human auditory cortex as detected by BOLD-FMRI. *Hearing Research, 126(1 -2),* 19-27.

Black, A. W., Taylor, P., & Caley, R. (1990). The festival speech synthesis system. Edinburgh, UK: Centre for Speech Technology Research (CSTR), University of Edinburgh. http://www.cstr.ed.ac.uk/projects/festival.html

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.24, retrived 23 January 2017 from http://www.praat.org.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*, 244-249.

Carcea, I., Insanally, M. N., & Froemke, R. C. (2017). Dynamics of auditory cortical activity during behaviour engagement and auditory perception. Nature Communication, 8:14412.

Chen, J., & ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes, 117*, 29 – 34.

Chen, J., van Rossum, D., & ten Cate, C. (2015). Artificial grammar learning in zebra finches and human adults: XYX versus XXY. *Animal Cognition, 18(1),* 151-164.

Chomsky, N. (1957). Syntactic Structures. The Hague, NL: Mouton.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control, 2(2),* 137 – 167.

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

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry, 36*, 1-22.

Christiansen, M. H. (2018). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science.*

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*, 489-509.

Christiansen, M. H., & Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology, 6:1182*, 1 – 18.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioural and Brain Sciences, 39(e62).*

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114(3),* 356-371.

Cooper, G., & Meyer, L. (1960). *The rhythmic structure of music.* Chicago: University of Chicago Press.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6,* 84 – 107.

Cooper, W. E., & Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *The Journal of the Acoustical Society of America, 62,* 683 – 692.

Cooper, W. E., Paccia, J. M., & Lapointe, S. G. (1978). Hierarchical coding in speech timing. *Cognitive Psychology, 10,* 154 – 177.

Costa-Faidella, J., Sussman, E. S., & Escera, C. (2017). Selective entrainment of brain oscillations drives auditory perceptual organization. Neuroimage, 159, 195-206.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47,* 292 – 314.

Dannenbring, G. L., & Bregman, A. S. (1976). Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 544-555.

de Diego-Balaguer, R., Martinez-Alvarez, A. & Pons, F. (2016). Temporal attention as a scaffold for language development. *Frontiers in Psychology, 7:44.*

de Vries, M. H., Christiansen, M. H., & Petersson, K. M. (2011). Learning recursion: Multiple nested and crossed dependencies. *Biolinguistics, 5(1/2),* 010 – 035.

de Vries, M. H., Monaghan, P., Knecht, S., Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: the learnability of embedded hierarchical structures. *Cognition, 107(2),* 763 – 774.

Deliège, I. (1987). Grouping conditions in listening to music: an approach to Lerhdahl & Jackendoff's grouping preference rules. *Music Perception, 4,* 325-360.

Delvaux, V. & Soquet, A. (2007). The influence of ambient speech on adult speech productions trough unintentional imitation. *Phonetica, 64,* 145-173.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics, 28,* 381-389.

Deutsch, D. (2013). Grouping Mechanisms in Music. In Deutsch, D. (Eds.), *The Psychology of Music* (184-238). San Diego, USA: Elsevier.

Deutsch, D. (2013). Grouping Mechanisms in Music. In Deutsch, D. (Eds.), *The Psychology of Music* (184-238). San Diego, USA: Elsevier.

Deutsch, D. & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review, 88,* 503-522.

Dowling, W. J. (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception & Psychophysics, 4,* 37-40.

Dowling, W. J., Lung, K. M., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics, 41,* 642-656.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*(2), 247–299.

Endress, A. D., Carden, S., Versace, E. & Hauser, M. D. (2010). The apes' edge: positional learning in chimpanzees and humans. *Animal Cognition, 13(3)*, 483-495.

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., & Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in Neuroscience, 9:157.*

Farmer, T.A., Christiansen, M.H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences, 103,* 12203-12208.

Fernald, A., & McRobers, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 365 – 388). Mahwah, NJ: Lawrence Earlbaum Associates.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review, 100,* 233-253.

Ferreira, F. (1992). Prosody. *Encyclopedia of cognitive science.* London, UK: Macmillan Reference Ltd.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47(2),* 164 – 203.

Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory: A complex systems approach. *Language Dynamics and Change*, *5(1),* 114 -137.

Fery, C., & Schubö, F. (2010). Hierarchical prosodic structures in the intonation of center-embedded relative clauses. *The Linguistic Review, 27*, 293-317.

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

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B*, *367*(1598), 1933-1955.

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. Nature Human Behaviour*, 3*, 393-405.

Frank, S. L. & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*, 829 – 834.

Frank, S. L., Bod, R., & Christianson, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B*, *279,* 4522-4531.

Frank, S. L., Trompenaars, T., & Vasishth, S. (2015). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science, 40(3),* 554 – 578.

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 of the United States of America, 103(7),* 2458 – 2463.

Frizelle, P., O'Neill, C., & Bishop, D. V. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language, 44*(6), 1435-1457.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences, 19*(3), 117-125.

Frost, R. L. A., Monaghan, P., & Tatsumi, T. (2017). Domain-general mechanisms for speech segmentation: The role of duration information in language learning. *Journal of Experimental Psychology: Human Perception and Performance.* Advance online publication. http://dx.doi.org/10.1037/xhp0000325

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes, 14(3),* 225-248.

Gleitmen, L. R., & Wanner, E. (1982). Language acquisition: The state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3-48). Cambridge: University Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105(2)*, 1166-1183.

Goldman, J.-Ph. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. *Proceedings of Interspeech*.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*, 103-113.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70(2)*, 109-135.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language, 51*, 548-567.

Haegeman, L. (1991). *Introduction to government & binding theory.* Oxford, UK: Blackwell.

Hamaoi, K., & Deutsch, D. (2013). The perceptual grouping of musical sequences: Pitch and timing as competing cues. In S. M. Demorest, S. J. Morrison, & P. S. Campbell (Eds.), *Proceedings of the 11th International Conference on Music Perception and Cognition (ICMPC11)* (pp. 81 – 87). Seattle, Washington, USA.

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

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,* 1569 – 1579.

Hawthorne, K., & Gerken, L. (2014). From pauses to clauses: Prosody facilitates learning of syntactic constituency. *Cognition, 133,* 420-428.

Hawthorne, K., Mazuka, R., & Gerken, L. (2015). The acoustic salience of prosody trumps infants' acquired knowledge of language-specific prosodic patters. *Journal of Memory and Language, 82,* 105-117.

Hay, J., & Diehl, R. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception and Psychophysics, 69,* 113-122.

Heimbauer, L. A., Conway, C. M., Christiansen, M. H., Beran, M. J., Owren, M. J. (2018). Visual artificial grammar learning by rhesus macaques (Macaca mulatta): exploring the role of grammar complexity and sequence length. *Animal Cognition, 21(2),* 267-284.

Hermann, C. S., Friederici, A. D., Oertel, U., Maess, B., Hahne, A., & Alter, k. (2003). The brain generates its own sentence melody: A Gestalt phenomenon in speech perception. *Brain and Language*, *85,* 396-401.

Howard, M. A. 3rd, Volkov, I. O., Abbas, P. J., Damasio, H., Ollendieck, M. C., & Granner, M. A. (1996). A Chronic microelectrode investigation of the tonotopic organization of human auditory cortex. *Brain Research, 724(2),* 260-264.

Hudson, R. A. (1996). *Sociolinguistics*. Cambridge: Cambridge University Press.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137,* 151 – 171.

Isbilen, E S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018), Bridging artificial and natural language learning: Chunk-based computations in learning and generalizing statistical structure. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1856-1861). Austin, TX: Cognitive Science Society.

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: An analysis using information theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31(1)*, 9-23.

Jusczyk, P. W., Hohne, E., & Mandel, D. (1995). Picking up regularities in the sound structure of the native language. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in a cross-language speech research* (pp. 91-119). Timonium, MD: York Press.

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition, 124,* 66 – 71.

Karlsson, F. (2007). Constraints on multiple centre-embedding of clauses. *Journal of Linguistics*, *43(2)*, 365 – 392.

Knoeferle, P. & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye-movements. *Journal of Memory and Language, 57,* 519 – 543.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22(1)*, 169-181.

Kraljic, T. & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, *50,* 194-231.

Kuchibhotla, K. V., Gill, J. V., Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. Miller, K. D., & Froemke, R. C. (2017). Parallel processing by cortical inhibition enables context-dependent behavior. *Nature Neuroscience*, *20(1)*, 62-71.

Kuchibhotla, K. & Batheliier, B. (2018). Neural encoding of sensory and behavioral complexity in the auditory cortex. *Current Opinion in Neurobiology, 52*, 65-71.

Kuhl, P. K., & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics, 31,* 279-292.

Lai, J. & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition, 118(2)*, 265-273.

Lai, J. & Poletiek, F. H. (2013). How "small" is "starting small" for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology, 25(4)*, 423 – 435.

Langus, A., Marchetto, E., Hoffmann Bion, R. A., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language, 66*, 285-306.

Lantos, G., Liu, G., Shafer, V., Knuth, K., & Vaughan, H. (1997). Tonotopic organisation of primary auditory cortex: An fMRI study. *Neuroimage, 5(4): S177.*

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* Cambridge, MA: MIT Press.

Locurto, C., Fox, M., & Mazzella, A. (2015). Implicit learning in cotton-top tamarins (*Saguinus Oedipus*) and pigeons (*Columba livia*). *Learning & Behavior, 43(2),* 129-142.

Männel, C., & Friederici, A. D. (2009). Pauses and intonational phrasing: ERP studies in 5-month-old German infants and adults. *Journal of Cognitive Neuroscience, 21(10),* 1988 – 2006.

Martin, E. (1970). Toward an analysis of subjective phrase structure. *Psychological Bulletin, 74,* 153 – 166.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: information-processing time with and without saccades. *Perception & Psychophysics, 53(4)*, 372 – 380.

McMullen, E. & Saffran, J. R. (2004). Music and Language: A developmental comparison. *Music Perception, 21(3),* 289-311.

Mertens, P. (2004). Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitment Automatique des Langues, 45*(2), 109-130.

Miller, G.A. (1958). Free recall of redundant strings of letters. *Journal of Experimental Psychology*. *56,* 485–491.

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *Journal of the Acoustical Society of America, 22*, 637-638.

Moho Studio 12 [computer software]. (2016). Retrieved from https://my.smithmicro.com/anime-studio-debut.html

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science, 9*, 21 – 34.

Monaghan, P. Brand, J., Frost, R. L. A., Taylor, G. (2017). Multiple variable cues in the environment promote accurate and robust word learning. In G. Gunzelman, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017) (pp. 817-822). Retrieved from https://mindmodeling.org/cogsci2017/papers/0164/index.html

Monaghan, P., Christiansen, M.H., & Chater, N. (2007). The Phonological Distributional Coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259-305.

Montag, J. L., & MacDonald, M., C. (2014). Visual salience modulates structure choice in relative clause production. *Language and Speech, 57*(2), 163-180.

Morgan, J. L. (1986). *From simple input to complex grammar.* Cambridge, MA: MIT Press.

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2010). Learnability of embedded syntactic structures. *Cognitive Science, 34(2),* 338 – 349.

Nazzi, T., Bertocini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 756-766.

Neiworth, J. J., London, J. M., Flynn, M. J., Rupert, D. D., Alldritt, O., & Hyde, C. (2017). Artificial grammar learning in Tamarins (*Saguinus Oedipus*) in varying stimulus contexts. *Journal of Comparative Psychology, 131(2),* 128-138.

Nespor, M., & Vogel I. (1986). *Prosodic phonology.* Dordrecht, NL: Foris Publications

Nespor, M., Shukla, M., van de Vijver, R., Avesani, C., Schraudolf, H., & Donati, C. (2008). Different phrasal prominence realizations in VO and OV languages. *Lingue e Liinguaggio, 2,* 1-29.

Newmeyer, F. J. (Eds.) (1988). *Linguistics: The Cambridge Survey* (Vols. 1-3). Cambridge: Cambridge University Press.

Newmeyer, F. J. (2017). Form and function in the evolution of grammar. *Cognitive Science, 41(S2),* 259 - 276.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Jornal of Memory and Language, 53*, 225 – 237.

Palmer, C., & Krumhansl, C. L. (1987). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Attention, Perception, & Psychophysics*, *41*(6), 505-518.

Pantev, C., Hoke, M., Lehnertz, K., Lütkenhöner, B., Anogianakis, G., & Wittkowski, W. (1988). Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields. *Electroencephalography and Clinical Neurophysiology, 69(2)*, 160-170.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics, 72(8),* 2254-2264.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6,* 674-681.

Patel, A. D. & Iversen, J. R. (2007). The linguistic benefits of musical abilities. *Trends in Cognitive Neuroscience, 11(9),* 369-372.

Peña, M., Bonatti., L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298,* 604 – 607.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of experimental psychology: General, 119*(3), 264-275.

Perruchet, P. & Rey, A. (2005) Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review, 12,* 307–313.

Perruchet, P., Tyler, M. D., Galland, N. & Peereman, R. (2004). Learning nonadjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General, 133(4)*, 573 – 583.

Peters, A. M. (1983). *The units of language acquisition.* Cambridge: Cambridge University Press.

Pickering, M. J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27(2),* 169-190.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *Journal of Acoustic Society of America, 66,* 363-369.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271 – 311). Cambridge, MA: MIT Press.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination in human newborns and by cotton-top tamarin monkeys. *Science, 288,* 349-351.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Verbal Learning and Verbal Behavior*, *5*, 855–863.

Rodgers, C. C. & DeWeese, M. R. (2014). Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents. *Neuron, 82(5),* 1157-1170.

Runyan, C. A., Piasini, E., Panzeri, S., & Harvey, C. D. (2017). Distinct timescales of population coding across cortex. *Nature, 548(7665)*, 92-96.

Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarind monkeys. *Cognition, 107(2),* 479-500.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition, 89(3),* 179 – 205.

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57,* 24 – 48.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: The MIT Press.

Schockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66(3),* 422-429.

Snedeker, J, & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language, 48,* 103 – 130.

Spierings, M., de Weger, A., & ten Cate, C. (2015). Pauses enhance chunk recognition in song element strings by zebra finches. *Animal Cognition, 18(4),* 867-874.

Spierings, M., Hubert, J., & ten Cate, C. (2017). Selective auditory grouping by zebra finches: testing the iambic-trochaic law. *Animal Cognition, 20(4),* 665-675.

Strainer, J. C., Ulmer, J. L., Yetkin, F. Z., Haughton, V. M., Daniels, D. L., & Millen, S. J. (1997). Functional MR of the primary auditory cortex: an analysis of pure tone activation and tone discrimination. *American Journal of Neuroradiology, 18,* 601-610.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers, 31(1),* 137-149.

Talavage, R. M., Benson, R. R., Galaburda, A. M., & Rosen, B. R. (1996). Evidence of multiple tonotopic fields in human auditory cortex [Abstract]. *Proc ISMRM, 4*, 1842.

Talavage, R. M., Ledden, P. J., Sereno, M. I., Rosen, B. R., & Dale, A. M. (1997). Multiple phase-encoded tonotopic maps in human auditory cortex. *Neuroimage, 5:S8.*

Tiitinen, H., Alho, K., Huotilainen, M., Ilmoniemi, R. J., Simola, J., & Näätänen, R. (1993). Tonotopic auditory cortex and the magnetoencephalographic (MEG) equivalent of the mismatch negativity (1993). *Psychophysiology, 30(5)*, 537-540.

Toro, J. M., Trobalon, J. B., & Sebastián-Gallés, N. (2003). The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition, 6,* 131-136.

Townsend, D. J., & Bever, T. G. (2001). *Sentence Comprehension: The integration of habits and rules.* Cambridge, MA: MIT Press.

Trotter, A. S., Monaghan, P., Beckers, G., & Christiansen, M. H. (Chapter 2). What do humans and animals learn in artificial grammar learning experiments? A focused meta-analysis. *Manuscript Accepted.*

Trotter, A. S., Frost, R. L. A., & Monaghan, P. (Chapter 3). Chained Melody: Low-level acoustic cues as a guide to phrase structure in comprehension. *Manuscript in preparation.*

Trotter, A. S., Frost, R. L. A., & Monaghan, P. (Chapter 4). Multiple natural language cues assist the processing of hierarchical structure. *Manuscript in preparation.*

Trotter, A. S., Monaghan, P., & Frost, R. L. A. (Chapter 5). Auditory-perceptual Gestalts affect the acquisition of hierarchical structure. *Manuscript in preparation.*

Trotter, A. S., & Monaghan, P. (Chapter 6). Gaze behaviour in the visual world suggests auditory-perceptual Gestalts facilitate the comprehension of hierarchical structure. *Manuscript in preparation.*

van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal Cognition, 16(2),* 165-175.

van Heijningen, C. A. A., de Visser, J., Zuidema, W., & ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences of the Unites States of America, 106(48)*, 20538-20543.

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.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor package. *Journal of Statistical Software, 36*(3).

Watson, D. G. Tanenhaus, M., & Gunlogson, C. (2008). Interpreting pitch accents in on-line comprehension: H* vs LH*. *Cognitive Science, 32,* 1232 – 1244.

Webb, J. T. (1969). Subject speech rates as a function of interviewer behaviour. *Language and Speech, 12(1),* 54-67.

Wessinger, C. M., Buonocore, M. H., Kussmaul, C. L., & Mangun, G. R. (1997). Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. *Human Brain Mapping, 5(1),* 18-25.

Wilson, B., Smith, K., & Petkov, C. I. (2015). Mixed-complexity artificial grammar learning in humans and macque monkeys: evaluating learning strategies. *European Journal of Neuroscience, 41(5),* 568-578.

Wilson, B., Spierings, M, Ravignan, A., Mueller, J.L., Mintz, T.H., Wijnen, F., van der Kant, A.., Smith, K., & Rey, A. (in press). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science, in press*.

Yamamoto, T., Uemura, T., & Lilnás, R. (1992). Tonotopic organization of human auditory cortex revealed by multi-channel SQUID system. *Acta Oto-Laryngalogica, 112(2),* 201-204.

Yamamoto, T., Williamson, S. J., Kaufman, L., Nicholson, C., & Lilnás, R. (1988). Magnetic localization of neuronal activity in the human brain. *Proceedings of the National Academy of Sciences USA, 85(22),* 8732-8736.

Yang, Y., Engelien, W., Zonana, J., Stern, E., Silbersweig, D. A., Physiological mapping of human auditory cortices with a silent event-related fMRI technique. *Neuroimage, 16,* 944-53.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N. & Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews, 81,* 103- 119.

Yu, C. & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70,* 2149 - 2165.

Zaccarella, E., & Friederici, A. D. (2017). The neurobiological nature of syntactic hierarchies. *Neuroscience and Behavioral Reviews, 81,* 205-212.

Zhang, J., Jiang, C., Zhou, L., & Yang, Y. (2016). Perception of hierarchical boundaries in music and its modulation by expertise. *Neuropsychologia, 91,* 490-498.