# Hierarchical syntactic structure predicts listeners' sequence completion in music

**Steffen A. Herff**[1] **(steffen.herff@epfl.ch)** **Daniel Harasim**[1] **Gabriele Cecchetti**[1]
**Christoph Finkensiep**[1] **Martin A. Rohrmeier**[1]

[1] Digital and Cognitive Musicology Lab, École polytechnique fédérale de Lausanne
Lausanne, VD 1015 Switzerland

## Abstract

Studies in psycho-linguistics have provided compelling evidence that theoretical syntactic structures have cognitive correlates that inform and influence language perception. Generative grammar models also present a principled way to represent a plethora of hierarchical structures outside the domain of language. Hierarchical aspects of musical structure, in particular, are often described through grammar models. Whether such models carry perceptual relevance in music, however, requires further study. To address the descriptive adequacy of a grammar model in music, unfamiliar musical phrases consisting of chord progressions within the Jazz idiom were used, and zero to three chords were cut from the end of each phrase. A total of 150 participants were then presented with these stimuli and asked to provide a *Closure Response*, that is to predict how many more chords (0, 1, 2, or 3) were expected before the chord progression was complete. Simultaneously, a grammar model of hierarchical structure as well as a bigram model were trained over a corpus of 150 expert-annotated Jazz tunes. The models were then used to estimate probability distributions of *Closure Responses* in the stimuli presented to the participants. Bayesian mixed-effects models reveal that the models carry predictive value for the participants' response distributions and that the hierarchical model contains incremental predictive information over the bigram model. The present results suggest that – akin to language – hierarchical relationships between musical events have a cognitive correlate, which influences the perception and interpretation of music.

**Keywords:** Syntactic models; music perception; cognition; tree models; hierarchical structures

## Introduction

From a computational perspective, a grammar is a modelling tool to construct and infer the syntactic structure of symbolic sequences, such as sentences comprising a language (Manning & Schütze, 1999). Syntactic structures account for hierarchically organised constituents and dependencies between the symbols within a sequence (Chomsky, 1957). Symbols linked by a dependency relation in turn are expected to be perceived as forming an implication/realisation pair (Rohrmeier, 2013, 2020), thus eliciting expectations that may be observed experimentally. If a sequence can be associated with several plausible syntactic structures by the same grammar, the grammar may also specify the probability of each alternative and, consequently, the probabilities of eliciting specific expectations.

Syntactic parsing, that is the computational process of mapping from sequences to syntactic structures afforded by a grammar, is a natural candidate as a model of the cognitive processing in domains that exhibit hierarchical organisation. However, a "descriptively adequate" grammar (Chomsky, 1965) needs to characterise syntactic relationships as they are perceived by listeners. Descriptive adequacy may be tested by showing that probabilistic predictions from a grammar model carry predictive value towards behavioural or neural responses that depend on the syntactic relationships amongst the elements of a stimulus. If they do, then this suggests that a cognitive representation exists that entails the information contained within the grammar model.

In the psycho-linguistic literature, empirical evidence for a cognitive representation of the hierarchical structures in language is abundant, which contributes to our understanding of the organisational principles of language and provides insight into the information that humans use to interpret a linguistic utterance (see, e.g., Uddén, Martins, Zuidema, & Fitch, 2020). Nevertheless, other domains such as dance (Charnavel, 2019), narrative (Van den Broek, 1988), action planning (Greenfield, Nelson, & Saltzman, 1972), or music (Lerdahl & Jackendoff, 1983) are also hypothesised to exhibit hierarchical organisation that influences perception (for a review see Cohen, 2000; Fitch & Martins, 2014). In music specifically, a long tradition of music-theoretical accounts suggests that the harmonic idiom of tonal music exhibits structures that may be modelled by a grammar of context-free complexity but not by a linear or local grammar. Instances of such structures are recursively nested relationships and non-local dependencies among individual harmonies (Steedman, 1984). The requirement for such grammars have not only been theoretically described (Rohrmeier, 2020) and computationally evaluated (Harasim, Rohrmeier, & O'Donnell, 2018; Granroth-Wilding & Steedman, 2014), but have also seen some empirical support in neural and behavioural responses of music listeners that show the existence of non-local dependencies (Cheung, Meyer, Friederici, & Koelsch, 2018; Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013).

To capture these hierarchical structures in music, grammars have been proposed and formalised (e.g., Rohrmeier, 2020; Harasim et al., 2018; Steedman, 1984), and it is a topic of ongoing empirical research whether the hierarchical structures they predict are descriptively adequate, have a cognitive correlate, and predict as well as influence perception. In this study, we directly tackle the issue of testing the descriptive adequacy of a grammar model for the harmonic idiom of Jazz (Levine, 1995).

### Hierarchy in tonal harmony

The idea that music exhibits hierarchical structure is linked to the understanding that individual musical events may be recursively elaborated, so that a variety of diachronic sequences are generated through the application of a limited set of generative principles (Lerdahl & Jackendoff, 1983). In this understanding, perceiving structure requires sensitivity to the relationships between events, even when such events are not adjacent in time because of an embedded elaboration. While recursive elaboration may be applied at all levels of representation within pieces, from individual notes to large-scale form, we here focus on harmony and adopt a symbolic representation of music in which a musical piece is represented as a sequence of chords. Each chord thus constitutes an event that may be elaborated recursively by other chord events, and the chords' internal structure (e.g., voicing) is not considered in this study.

### Hierarchy and expectations

Previous research suggests that hierarchical structures in music influence listeners' perception, for example by highlighting responses to the violation of non-local dependencies (Koelsch et al., 2013). Furthermore, specific hierarchical dependencies can be learnt implicitly from the exposure to an artificial musical language (Rohrmeier & Cross, 2009).

To test specific grammars, psycho-linguistic studies benefit from the availability of clear referential semantics when conducting experimental studies that explore the perception and cognition of hierarchical structures in language. For example, consider the sentence *"The dog chasing the cat is grey"*. Asking a participant which of the two animals is the grey one provides an easy approach to investigate the (non-local) structural interpretation that a participant has formed of the given stimulus. The responses can then be compared to the predictions of a grammar model. In non-linguistic stimuli, such as music, without a comparably clearly defined referential semantics, other approaches must be considered. A useful property of probabilistic grammars in such circumstances is that they afford quantitative predictions about necessary future events, which may model perceived expectations on part of the listener (Rohrmeier, 2013). For example, when asked to estimate how many more words are expected to come before a sentence is complete, the stimulus *"I like pizza"* would likely lead to fewer estimated words than the stimulus *"I like"*, whereas the stimulus *"I like and"* would likely lead to a higher number of estimated words. This is because the first stimulus is grammatical as is, whereas the second stimulus requires at least one word to form a grammatical sentence. The third stimulus, on the other hand, can not form a grammatically correct sentence with less than two additional words, because an object (e.g. *"pizza"*) is required to fulfil the grammatical implication of *"like"* and an additional verb (e.g., *"eat"*) is required to fulfil the grammatical implication of *"and"*. Since English grammar does not allow a single word to fulfil both the role of the object and of the verb at the same time, a minimum of two additional words is required. Whilst asking participants to provide referential semantic interpretations provides methodological difficulties in music, it is possible to ask participants to provide *Closure Responses*, that is how many more musical events (e.g., chords) are expected to come, before a musical piece can be complete. The same task can also be performed by a grammar and the results compared to those observed in participants. Observing predictive value of the probabilities of the continuation lengths for participants' *Closure Responses* would be consistent with the hypothesis that participants' behaviour is informed by the dependencies entailed by the grammar.

In this study two grammar models – a local (bigram) and a hierarchical (probabilistic context-free) one – are trained on expert annotations of dependency structures within the songs of a Jazz corpus. Both models are used to predict the dependency structures of previously unseen Jazz pieces and the results are compared against expert annotations. Afterwards, the trained models are used to predict perceptual responses of listeners participating in a behavioural study where participants were tasked to guess how many more chords they expected to come until the end of an interrupted musical piece.

## Hierarchical Model

The hierarchical grammar model used here is an instance of a Probabilistic Context-Free Grammar (PCFG), a standard formalism to quantitatively describe hierarchical structures in sequential data (Manning & Schütze, 1999). A PCFG analyses the structure of a sequence (here a chord progression) by re-creating it through recursive elaboration of shorter sequences. For example, Figure 1 shows a derivation tree of a stimulus from the present perceptual experiment, which represents the hierarchical relations between the chords.

A PCFG consists of sets of *terminal symbols* $T$, *nonterminal symbols* $N$ disjoint to $T$, *grammar rules*

$$R \subseteq \{A \longrightarrow \alpha \mid A \in N, \alpha \in (T \cup N)^+\}, \tag{1}$$

and a *start symbol* $S \in N$, where $(T \cup N)^+ = \bigcup_{k \geq 1} (T \cup N)^k$ denotes the set of nonempty sequences of a mix of terminal and nonterminal symbols. A *derivation* of a sequence is defined as a list of rule applications that successively generate it from the start symbol. For each grammar rule $A \longrightarrow \alpha$, there is additionally a real number $0 \leq \pi_{A \longrightarrow \alpha} \leq 1$ such that $\sum_\alpha \pi_{A \longrightarrow \alpha} = 1$ for all nonterminals $A \in N$ (where $\pi_{A \longrightarrow \alpha} = 0$ if and only if $A \longrightarrow \alpha \notin R$). Under mild technical conditions, $\pi_{A \longrightarrow \alpha}$ can be interpreted as the probability of the rule $A \longrightarrow \alpha$ (Booth & Thompson, 1973). The probability of a derivation is then defined as the product of its rules, and the probability of a sequence of terminal symbols is the sum of the probabilities of all its derivations.

In this study, we adopt an existing grammar model (Harasim, Finkensiep, Ericson, O'Donnell, & Rohrmeier, 2020). Terminal and nonterminal symbols both are chord symbols encoded as strings that represent the chord's root note and chord form (e.g., Cmaj7, Dmin7, G7). In addition,

the nonterminal symbols contain an artificial start symbol $S$. The set of rules $R$ is partitioned into five classes that express different kinds of harmonic dependencies. Each rule is either a *start rule* $S \longrightarrow A$, a *duplication rule* $A \longrightarrow A\ A$, a *left-headed rule* $A \longrightarrow A\ B$, a *right-headed rule* $A \longrightarrow B\ A$, or a *terminal rule* $A \longrightarrow \bar{A}$, and all possible rules are included in $R$ ($A, B \in N$ such that $A \neq B$, and $\bar{A} \in T$ such that $A$ and $\bar{A}$ refer to the same chord).

The rule probabilities are inferred from a corpus of expert-created derivation trees (Harasim et al., 2020) by a standard estimation that was similarly applied to model chord sequences in previous research (Harasim, O'Donnell, & Rohrmeier, 2019). The success of the inference is then assessed by comparing the predictions of the inferred grammar to the expert analyses. For each nonterminal A, the probabilities of all rules with $A$ on their left-hand side are considered a categorical distribution over applicable rules $A \longrightarrow \alpha$, on which we put a symmetric Dirichlet prior with concentration parameter $\kappa = 0.01$. Denoting the $i$-th derivation tree in the corpus by $t_i$ and the total number of trees by $n$, we estimate $\pi_{A \longrightarrow \alpha}$ by its expected value under the posterior distribution given the derivation trees,

$$\pi^*_{A \longrightarrow \alpha} := \mathbb{E}_{p(\{\pi_{A \longrightarrow \alpha}\}_\alpha | t_1, \ldots, t_n)} [\pi_{A \longrightarrow \alpha}] \tag{2}$$

$$= \frac{\kappa + \sum_{i=1}^{n} \#(A \longrightarrow \alpha \mid t_i)}{\sum_{A \longrightarrow \beta \in R} (\kappa + \sum_{i=1}^{n} \#(A \longrightarrow \beta \mid t_i))}, \tag{3}$$

where $\#(A \longrightarrow \alpha \mid t_i)$ denotes the number of occurrence of rule $A \longrightarrow \alpha$ in the $i$-th derivation tree. The grammar model uses the inferred probabilities to predict the derivation tree of an unseen sequence of chords as its derivation with the highest probability.

To model the probability of a specific completion length, consider a chord-progression stimulus $\sigma$ and a completion $\tau$ (i.e., a chord progression of length 0, 1, 2, or 3) as shown in Figure 1, and denote the length of $\tau$ by $\text{len}(\tau)$. After the observation of a stimulus $\sigma$, the probability of a completion length $\lambda = \text{len}(\tau)$ is given by:

$$p(\lambda \mid \sigma) = \frac{p(\lambda, \sigma)}{p(\sigma)} = \frac{\sum_{\tau \in T^\lambda} p(\sigma\tau)}{\sum_{\lambda'=0}^{3} \sum_{\tau \in T^{\lambda'}} p(\sigma\tau)} \tag{4}$$

Note that all information from the model is contained in the joint distribution $p(\lambda, \sigma)$, which is given by marginalising over all completions $\tau$ with length $\lambda$. Dividing by $p(\sigma)$ makes the completion length's probability $p(\lambda \mid \sigma)$ independent from how likely (i.e., well-formed) the stimulus $\sigma$ is. Since a longer chord sequence requires the application of more grammar rules, longer completions tend to have lower probabilities. The model therefore favors shorter completions. Precisely, the trained model's prior distribution over completion lengths $p(\lambda)$ is (.33, .26, .22, .19).

In order to test whether the bias for shorter completions is beneficial for modelling the participants' behavioural responses, we also compute the probabilities of the completion lengths based on a uniform prior and compare the predictive performance of the original and the adjusted model. To obtain the adjusted distribution, we divide $p(\lambda \mid \sigma)$ by the original model's length bias $p(\lambda)$ and renormalise.

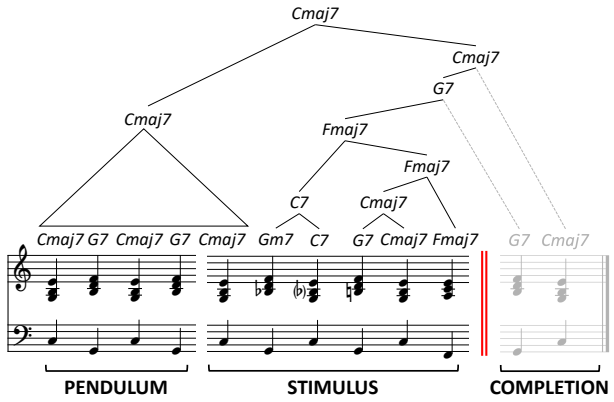


Figure 1: Example Stimulus. The musical phrase was interrupted at the red-line, and participants were asked to provide *Closure Responses*, that is to estimate the number of chords they expected to complete the phrase. A derivation tree is shown above the stimulus.

## Bigram Model

The local model used in this study is a bigram model with a weak uniform smoothing. A bigram is a pair of two chords; it represents a chord transition. The bigram model generates sequences, one chord at a time from the start to the end, where the probability of any chord only depends on its immediate predecessor. This results in a strictly right-branching tree structure, with the probability of a sequence only depending on its local chord transitions. The bigram probabilities $p(a.b)$ are inferred from the chord sequences in the corpus of Jazz standards by counting, smoothing, and normalization, so that

$$p(a.b) = \frac{s + \#(a.b)}{|T|s + \sum_{b' \in T} \#(a.b')}, \tag{5}$$

where $T$ is the set of chord symbols, $a.b$ denotes a bigram (i.e., a chord transition from $a \in T$ to $b \in T$), and the smoothing parameter $s$ is set to 0.1. An artificial end-of-sequence symbol EOS is added to the set of chords in order to model the termination of the sequence generation. The smoothing is used to account for chord transitions not seen in the corpus.

The probability of a completion length $\lambda$ is given by Equation 4 together with $p(\sigma\tau) = \prod_{k=1}^{\text{len}(\tau)+1} p(\tau_{k-1}.\tau_k)$, where $\tau_k$ denotes the $k$-th element of the sequence $\tau$ for $k \in \{1, \ldots, \text{len}(\tau)\}$, $\tau_0$ stands for the last element of $\sigma$, and $\tau_{\text{len}(\tau)+1}$ stands for the end-of-sequence symbol EOS. The assumption that a sequence can only end on the tonic chord Cmaj7 is encoded into the bigram model by setting $p(a.\text{EOS})$ to zero for all other chords $a \in T \setminus \{\text{Cmaj7}\}$. Note, that similiar to the hierchical model without a uniform prior, the bigram model has a bias towards shorter completions.

## Aims and hypotheses

This study aims to shed light on the descriptive adequacy (Chomsky, 1965) in the musical domain of hierarchical grammars that incorporate music-theoretical insights about structure. To this end, we train and evaluate an existing hierarchical grammar as well as a bigram model using a recently published data set of expert-annotated music (Harasim et al., 2019, 2020). We then compare the models' structural analyses with participants' *Closure Responses* in a new perceptual experiment. We hypothesise (i) that expert structural analyses are better captured by a hierarchical compared to a bigram model, (ii) that a bias towards shorter grammatical completions – which is intrinsic to the probabilistic hierarchical and the bigram model – also exists in participants. (iii) that the structural analyses generated by the hierarchical model carries incremental predictive power over the bigram model when it comes to predicting listeners' Closure Responses.

## Method

### Grammar Evaluation

The bigram as well as the hierarchical grammar model were used to generate tree analyses for 150 Jazz tunes. The models' predictions were compared with the expert analyses using leave-one-out-cross-validation. The *unlabeled tree accuracy* was calculated for each tune by ignoring the nonterminal labels and identifying the number of correctly predicted constituents relative to the total number of constituents, where a constituent is defined as a subtree's terminal sequence (i.e., its leafs). As a baseline, the same measure was used for a large set of randomly generated derivation trees for each piece.

### Perceptual Experiment

**Participants.** A total of 150 members of the general public took part in the experiment during the École polytechnique fédérale de Lausanne Open Days. Participation was voluntary, took place in an experimental booth located at a public site, and was fully self-administered and unsupervised. The study received local ethics board's approval (HREC 009-2019/21.02.2019). Due to the public location of the experiment, the sample is representative of the broader community, but no sensitive, personal, or demographic data was collected. After reading the instruction and providing informed consent, the study took approximately 30 minutes to complete. The entire experiment was completed by 99 participants. However, as the results between the participants that completed the entire experiment and those that did not were not significantly different from one another, the final analysis includes the data of all 150 volunteers (3585 observations).

**Stimuli.** Isochronous stimuli in the harmonic idiom of the common Jazz-standard repertoire (Levine, 1995) were generated, which consisted of 6 to 9 chords. The last 0, 1, 2, or 3 chords of each stimulus were deleted, and the remaining stimulus was merged with a preceding $I^{maj7}$-$V^7$ pendulum. A fade-in combined with the pendulum functioned to establish the global key of the current stimulus and to blur its metrical organisation, which could otherwise potentially influence participants' responses. The result was a total of 32 stimuli of equal length. Figure 1 shows an example stimulus.

**Perceptual task.** Participants were instructed that they would hear the last few chords of a song which may or may not be interrupted early. It was their task to provide *Closure Responses*, that is to estimate how many more chords they expected to come before the song was completed. Each participant performed 32 trials in which they were presented with one stimulus in a random transposition (max. six semitones up or down). After each presentation, participants were prompted to estimate the number of missing chords (0, 1, 2, or 3), where 0 indicated that they perceived this stimulus to be complete as presented. Feedback was provided implicitly by playing the ending of each stimulus after a response was given. This was done to reinforce the musical idiom used in the present study. Analogous to the *"Pizza"* example in the introduction, the task utilises the fact that the representation of a grammar also affords structural orientation which would allow predictions for stimuli generated by that grammar. The task tests whether the ability to navigate the hierarchical structure of a stimulus, which can be demonstrated in language, is also available in music. This would provide evidence that hierarchical relations between chords have a cognitive correlate that influences the interpretation of music.

### Statistical approach

We use Bayesian mixed effects models to account for cross-random effects of *Response Category* and *Stimulus* effects. Each model was provided with weakly informative priors $t(3,0,1)$, and we report coefficient estimates ($\beta$), Estimated Errors in the coefficients (EE), and evidence ratios (Odds) for the individual hypotheses. For convenience, we indicate effects that can be considered significant at an $\alpha = 0.05$ level with * (i.e., odds ratio greater 19; Milne & Herff, 2020). All continuous variables are scaled ($M = 0$, $SD = 1$).

To further assess the ability of the grammar models to predict perceptual responses on unseen data, we use Pareto-Smoothed Importance Sampling with approximate Leave-One-Out cross-validation (PSIS-LOO; Vehtari, Gelman, & Gabry, 2017), and report Expected Log point-wise Predictive Densities (ELPDs). An ELPD difference of $\Delta$ means that the likelihood ratio between the two models, given the priors and out-of-sample data, is $e^{|\Delta|}$ in favour of the model with the higher ELPD.

## Results

All data, code, stimuli, and fitted models can be obtained from https://osf.io/9wjyg/.

### Grammar Evaluation Results

A Bayesian mixed effects model was used to predict the accuracy values for each piece using the *Model* (Random vs. Bigram *B* vs. Hierarchical *H*) as predictor, whilst control-

ling for a random effect of musical piece. Strong evidence was observed in favour of higher accuracy in the hierarchical model compared to the random baseline ($\beta_H$ = 1.61, $EE_H$ = .07, *Odds* ($\beta_H > 0$) > 9999*). The bigram model, however, did not show better accuracy than the baseline ($\beta_B$ = -.11, $EE_B$ = .07, *Odds* ($\beta_B > 0$) = .07). This is visualised in Figure 2. This result supports hypothesis (i), and suggests that the hierarchical grammar is a better model of the abstract *syntactic competence* (Chomsky, 1965) of the Jazz harmonic idiom, as reflected in expert annotations.

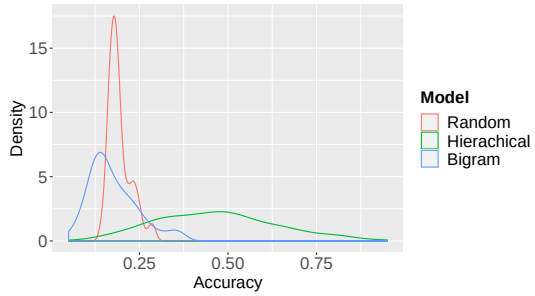


Figure 2: Model accuracy in predicting expert structural annotations of 150 previously unseen musical pieces. The hierarchical model (green) performed significantly better than chance (red), whereas the bigram model (blue) did not.

However, to what extent such syntactic competence is also reflected in non-expert listeners' *performance* when listening to music is yet unclear. To address this point, we compare now the effectiveness of the hierarchical and the bigram model in utilising information contained in expert annotations to predict non-expert listeners' perception of closure

## Perceptual Results

For each stimulus and each of the four possible responses (0, 1, 2, 3), the proportion of the participants that gave that response was predicted using Bayesian mixed effects models. Each model was provided with Random Effects (*RE*) for stimulus and response category.

**A bias towards shorter completions.** To assess hypothesis (ii), two models were compared. The first model was provided with the predictions of the hierarchical model that favours shorter completions in addition to the random effect structure (*H + RE*). In addition to the random effects, the second model was provided with the predictions of the hierarchical model that uses a uniform length prior, to remove the intrinsic bias towards shorter completions ($H_{(uniform\ length\ prior)}$ + *RE*). Both models were compared with each other and a baseline model that only contained the random effects (*RE*). Table 1 shows that both hierarchical models outperformed the baseline model. However, the model favouring shorter completions slightly (|ΔELPD| = 1.7) outperformed the model with a uniform length prior, supporting hypothesis (ii).

Table 1: Model comparison between a hierarchical model favouring shorter perceived completions and one that does not. The difference in ELPD between each model and the best models (ΔELPD), as well as the standard error of this differences are reported ( $SE_\Delta$).

| Predictors | ΔELPD | $SE_\Delta$ |
|---|---|---|
| Hierarchical + RE | 0 | 0 |
| Hierarchical $_{(\text{uniform length prior})}$+ RE | -1.7 | 0.6 |
| RE | -8.7 | 4.6 |

**Hierarchical model contains incremental predictive power.** To assess hypothesis (iii), the hierarchical model favouring shorter completions (*H + RE*) was compared with a model containing the predictions of the bigram model as well as the random effect structure (*B + RE*). As seen in Table 2, both the hierarchical as well as the bigram model outperformed the baseline model (*REF*). However, the hierarchical model also showed slightly (|ΔELPD| = 1.7) better predictions than the bigram model.

We further assessed the incremental predictive value of the hierarchical model by implementing a model that contains the predictions of both the bigram and the hierarchical model, as well as the interaction between the two (*H + B + H:B + RE*). Within this model, the hierarchical predictor ($\beta_H$ = .18, $EE_H$ = .09, *Odds* ($\beta_H > 0$) = 42.68*), as well as the Hierarchical:Bigram interaction ($\beta_{H:B}$ = .17, $EE_{H:B}$ = .07, *Odds* ($\beta_{H:B} > 0$) = 127.44*) carried predictive value, but the bigram predictor did not ($\beta_B$ = .07, $EE_B$ = .09, *Odds* ($\beta_B > 0$) = 3.25). This is likely because the information contained in the bigram predictor was already captured by the other predictors. As a result, we implemented another model, identical to the previous one, but dropping the individual bigram predictor (*H + H:B + RE*). Table 2 shows that of all the models tested, this model descriptively performed best and, specifically, better than the bigram model alone.

Note that predictions of the hierarchical and the bigram model are correlated (*r* = .72). It is possible that the information contained in the hierarchical:bigram interaction was the result of a quadratic link between the hierarchical prediction and perceptual data. We tested for this possibility by including a $H+H^2+RE$ model. Table 2 shows that this model was outperformed by the model containing the hierarchical:bigram interaction, suggesting that the bigram predictor did provide the model with an independent contribution, but only in its interaction with the hierarchical predictor. Taken together, these results support hypothesis (iii).

## Discussion and conclusion

We used a hierarchical model to generate analyses that incorporate the music-theoretical understanding of harmonic structure (Harasim, 2020), as encoded by expert analyses in a corpus of Jazz chord progressions (Harasim et al., 2020). As hypothesised, the hierarchical model captured expert annotations of musical pieces, whereas a bigram model did

Table 2: Model comparison between bigram and hierarchical predictions.The difference in ELPD between each model and the best models (ΔELPD), as well as the standard error ( $SE_\Delta$) of this differences are reported.

| Predictors | ΔELPD | $SE_\Delta$ |
|---|---|---|
| Hierarchical+Hierarchical:Bigram+ RE | 0 | 0 |
| Hierarchical+Bigram+Hierarchical:Bigram+ RE | -0.5 | .5 |
| Hierarchical+RE | -2.7 | 3.2 |
| Hierarchical+Hierarchical$^2$+ RE | -3.2 | 2.8 |
| Bigram+RE | -4.3 | 3.7 |
| RE | -11.4 | 5.3 |

not. Nevertheless, both the bigram as well as the hierarchical model carried predictive information for the distribution of listeners' responses in a behavioural task, in which the participants were asked to estimate how many more chords they expected to come before a musical piece could be complete. Model comparison revealed a small bias in listeners to prefer shorter completions. Furthermore, a predictive model containing the predictions of the bigram model improved significantly when additionally provided with the predictions of the hierarchical model as well as an interaction term, mediating the link between the hierarchical and the bigram models' predictions, and the participants' responses.

The incremental predictive value of the hierarchical model for predicting the proportion of participants that respond with a certain completion length for a specific stimulus supports the descriptive adequacy (Chomsky, 1965) of the grammar. This does not necessarily mean that the cognitive representations of syntactic structure formed by the listeners are identical to those generated by the present hierarchical grammar model. However, this study supports the hypothesis that, at the computational level of description, the task performed by the hierarchical grammar model in predicting the completion lengths based on the grammar's rules and their inferred probabilities does model cognitive processes that are relevant for music perception. This includes two aspects.

First, the model's rule probabilities are inferred from a corpus of ecological musical material annotated by experts. The predictive performance of the model then suggests that some of the music-theoretical relationships captured by the annotators and modelled through the grammar are also available to the listeners and influence their perception. Note that the specific dependencies found in the Jazz corpus require a grammar that is able to express long-term dependencies or center embedding, such as a context-free grammar (Harasim et al., 2020).

Second, the predictions of the grammar model were tested against a behavioural response, the *Closure Response*, which quantifies participants' expectations towards future events. Since the hierarchical grammar model bases its predictions on dependency relations being open at the moment the terminal sequence is interrupted, our results support the understanding that open dependency relations manifest themselves perceptually and behaviourally in terms of expectations (Rohrmeier, 2013). In this respect, it is important to consider that the hierarchical grammar model, differently from the local bigram model, affords multiple dependency relations to be nested into one another, which would translate into multiple expectations to coexist in perception. While the bigram model was also able to predict participants' responses, the hierarchical model carried incremental predictive information, supporting that nested dependencies may capture an aspect of music perception. Note that it was not the objective of the present study to find the best way of predicting participants' responses in the perceptual task. Instead, this study aimed at finding a suitable grammar that models the link between expert annotations and perceptual predictions.

The predictions of the hierarchical as well as the bigram model used here are biased towards shorter completions. This is because the models estimate the probability of a completion as the product of the probabilities of the rules used to generate the completion. As the grammars do not contain substitution rules, and each generative rule can not produce more than two children, longer sequences are indicative of more applied rules, thus more multiplicative steps. A model comparison showed that this bias also carries a small incremental predictive value for participants' responses. This bias towards shorter completion is a shared feature of rule-based models and reflects the cost of additional rule applications on the likelihood of an entire derivation. As participants also favour shorter completions, a similarly incremental application cost for generating longer sequences could apply to the cognitive representation of musical structure.

In a direct comparison, the hierarchical and bigram grammar models used in the present study are much simpler compared to state-of-the-art models in language (Brown et al., 2020). This is worth noting, as the models carry great predictive value for perceptual responses despite their simplicity, and further refinement may even further increase the models' performance. However, it is also important to note that we did not directly test syntactic structure as perceived by participants. In contrast, the present study can be seen as a musical implementation of the "pizza" example presented in the introduction. It is possible, albeit unlikely, that the participants formed structural representations that are different from those predicted by the grammar model and share closure-rating distributions only coincidentally. For example, the response bias for shorter sequences may also have been the result of a specific response strategy, such as a search strategy that begins with shorter completions to save resources. Other effects may also influence participants' predictions, such as aesthetic preference. Future studies could investigate the precise syntactic structure formed by the participants using novel paradigms (Cecchetti, Herff, & Rohrmeier, 2021) to assess whether the cognitive representation of musical structure is not only quantitatively, but also qualitatively, best captured by a hierarchical grammar model.

The present study supports the view that the underlying organisational principles of language and music are comparable, and that hierarchical structures identified by music theorists indeed carry perceptual relevance in addition to local structures. However, this does not mean that implementation of hierarchical syntactic processing is identical in the two domains. Indeed, recent neuroscientific insights suggest that the neural structures involved in the processing of hierarchical structure in language and music rely on domain-selective neural populations that are inversely lateralised (Friederici, 2020). By establishing the perceptual relevance of hierarchical structures in other domains beyond music and language, future research could describe how the processing of seemingly unrelated stimuli – that are structured according to common syntactic principles – shows commonalities and differences in order to reveal fundamental principles of both cognitive as well as neural architecture.

## Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No 760081-PMSB.

## References

Booth, T., & Thompson, R. (1973, May). Applying Probability Measures to Abstract Languages. *IEEE Transactions on Computers*, *C-22*, 442–450.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Cecchetti, G., Herff, S. A., & Rohrmeier, M. A. (2021). Musical syntactic structure improves memory for melody: evidence from the processing of ambiguous melodies. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Charnavel, I. (2019). Steps toward a universal grammar of dance: Local grouping structure in basic human movement perception. *Frontiers in psychology*, *10*, 1364.

Cheung, V. K., Meyer, L., Friederici, A. D., & Koelsch, S. (2018). The right inferior frontal gyrus processes nested non-local dependencies in music. *Scientific reports*, *8*, 1.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Walter de Gruyter.

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

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of cognitive psychology*, *12*, 1–36.

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Ann. N. Y. Acad. Sci*, *1316*, 87–104.

Friederici, A. D. (2020). Hierarchy processing in human neurobiology: how specific is it? *Philosophical Transactions of the Royal Society B*, *375*.

Granroth-Wilding, M., & Steedman, M. (2014). A Robust Parser-Interpreter for Jazz Chord Sequences. *Journal of New Music Research*, *43*(4), 355–374.

Greenfield, P. M., Nelson, K., & Saltzman, E. (1972). The development of rulebound strategies for manipulating seriated cups: A parallel between action and grammar. *Cognitive psychology*, *3*, 291–310.

Harasim, D. (2020). *The Learnability of the Grammar of Jazz: Bayesian Inference of Hierarchical Structures in Harmony*. PhD Thesis, EPFL, Lausanne.

Harasim, D., Finkensiep, C., Ericson, P., O'Donnell, T. J., & Rohrmeier, M. (2020). The Jazz Harmony Treebank. In *Ismir*. Montréal, Canada.

Harasim, D., O'Donnell, T. J., & Rohrmeier, M. (2019). Harmonic Syntax in Time: Rhythm Improves Grammatical Models of Harmony. In *Ismir* (pp. 335–342). Delft.

Harasim, D., Rohrmeier, M., & O'Donnell, T. J. (2018). A generalized parsing framework for generative models of harmonic syntax. In *Ismir* (pp. 152–159).

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proc. Natl. Acad. Sci. U.S.A*, *110*(38), 15443.

Lerdahl, F., & Jackendoff, R. S. (1983). *A Generative Theory of Tonal Music*. MIT Press.

Levine, M. (1995). *The Jazz Theory Book*. Petaluma: Sher Music.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Milne, A. J., & Herff, S. A. (2020). The perceptual relevance of balance, evenness, and entropy in musical rhythms. *Cognition*, *203*, 104233.

Rohrmeier, M. (2013). Musical Expectancy: Bridging Music Theory, Cognitive and Computational Approaches. *Zeitschrift der Gesellschaft für Musiktheorie*, *10*(2), 343.

Rohrmeier, M. (2020). The syntax of jazz harmony: Diatonic tonality, phrase structure, and form. *Music Theory and Analysis*, *7*, 1–63.

Rohrmeier, M., & Cross, I. (2009). Tacit tonality: Implicit learning of context-free harmonic structure. In *Escom*.

Steedman, M. J. (1984). A generative grammar for jazz chord sequences. *Music Perception*, *2*(1), 52–77.

Uddén, J., Martins, M., Zuidema, W., & Fitch, W. T. (2020). Hierarchical structure in sequence processing: How to measure it and determine its neural implementation. *Topics in cognitive science*, *12*, 910–924.

Van den Broek, P. (1988). The effects of causal relations and hierarchical position on the importance of story statements. *Journal of Memory and Language*, *27*, 1–22.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413.