Thèse n° 10 590

Hearing structure in music: An empirical inquiry into listening as representation and processing

Présentée le 17 avril 2024

Collège des humanités
Digital and Cognitive Musicology Lab
Programme doctoral en humanités digitales

pour l'obtention du grade de Docteur ès Sciences

par

Gabriele CECCHETTI

Acceptée sur proposition du jury

Prof. D. Gatica-Perez, président du jury
Prof. M. A. Rohrmeier, Dr S. A. Herff, directeurs de thèse
Prof. M. Farbood, rapporteuse
Prof. S. Koelsch, rapporteur
Prof. O. Blanke, rapporteur

École
polytechnique
fédérale
de Lausanne

2024

D’ora in avanti sarò io a descrivere le città, – aveva detto il Kan. Tu nei tuoi viaggi verificherai se esistono. Ma le cittá visitate da Marco Polo erano sempre diverse da quelle pensate dall’imperatore. Eppure io ho costruito nella mia mente un modello di città da cui dedurre tutte le città possibili – disse Kublai. [...] Anch’io ho pensato a un modello di città da cui deduco tutte le altre, – rispose Marco. È una città fatta solo d’eccezioni, preclusioni, contraddizioni, incongruenze, controsensi. [...] Dunque basta che io sottragga eccezioni al mio modello, e in qualsiasi ordine proceda arriverò a trovarmi davanti una delle città che, pur sempre in via d’eccezione, esistono. Ma non posso spingere la mia operazione oltre un certo limite: otterrei delle città troppo verosimili per essere vere.

— Italo Calvino, *Le città invisibili*

A Sofia.

Acknowledgements

I would like to thank Martin Rohrmeier for sharing his deep, wide-ranging, and uncompromising *Weltanschauung*'s intellectual inspiration, and for giving me the opportunity to pursue such an ambitious research topic while learning so much in the process. If I could forgive the sin of getting my theoretical hands dirty in empirical mud, I would also be deeply indebted to Steffen Herff for his constant support in navigating unfamiliar waters, through daily exchanges, teachings, guidance, and encouragement.

I have shared my years in Lausanne with several "generations" in and around the DCML: Christoph Finkensiep (whose insights in particular have inspired many aspects of this thesis), Fabian Moss, Johannes Hentschel, Daniel Harasim, Andrew McLeod, Petter Ericson, Ken Déguernel, Ludovica Schaerf, Ravinithesh Annapureddy, Shuxin Meng, Edward Hall, Ran Tamir. The last mile in this path has been enriched by the friendship and advice of Zeng Ren, Xinyi Guan, and Yannis Rammos. To these people I owe all the best and fun of the PhD life, countless teachings, hours of endless discussions on music, science, food, and just being a bit silly. For all these things, and those to come, thank you.

These pages enclose what is, ultimately, the product of a clash of identities: a physicist borrowed to music, a musician borrowed to theory, a theorist borrowed to empirical research. My gratitude goes to those who supported this – often uncomfortable – journey with patience and trust, and the many who planted the seeds: in particular, Lorenzo Dello Schiavo, Leonardo Zelli, Giordano Ferranti, Simone Genuini, Paolo Carini, Antonio Polosa, Ian Cross. Finally, I am grateful to Sofia for all the sweetness and brightness she shared, for dragging me so often out in the sun, and for always reminding me of what matters. Grazie mamma, papà, amici vecchi e nuovi: Zoe, Silvia, Federico, Erica, Mathias.

Lausanne, 13 march 2024 G. C.

Funding statement

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

Abstract

As a universal expression of human creativity, music is capable of conveying great subtlety and complexity. Crucially, this complexity is not encoded in the score or in the sounds, but is rather construed in the mind of the listener in the form of nuanced perceptual experiences, commonly referred to as "structural hearing". While these experiences are to some extent accessible to introspection, which is made explicit in the music-theoretical discourse, the underlying cognitive mechanisms are elusive of empirical investigation. In this thesis, we conceptualise the experience of musical structure in the context of Bayesian cognition as a form of inference: namely, the inference of representations of structure as a way of making sense of music's sensory signals. Exploiting a computational analogy with linguistic processing, we model the emergence of structural interpretations in terms of grammar-based incremental parsing.

In a series of behavioural experiments, we test some crucial implications of this modelling approach: (1) the existence of representations of structure abstracted from sensory information, which we test by adapting a structural-priming paradigm to the musical case, (2) the cognitive relevance of idiom-specific syntactic categories, exemplified by the notion of harmonic function in extended-tonal harmony, (3) the time-course of cognitive computations implementing incremental parsing in real time during listening, and (4) the existence of mechanisms of retrospective reanalysis by analogy with the linguistic garden-path effect.

Overall, these results contribute proofs of existence for some cornerstones of a computational- and algorithmic-level theory of structural hearing. They are compatible with an inference process implemented through parsing computations including the integration of newly encountered events into a pre-existing representation, the projection of expected events in the future, and the retrospective revision of the interpretation of past events. Building on the proposed framework, future work may further test implications of different fine-grained algorithmic models of parsing, in order to distinguish between accounts of processing similarly to how models of sentence comprehension are disambiguated in psycholinguistics.

Keywords: musical structure, structural hearing, grammar, parsing, music cognition, harmony, rhythm, garden-path, music and language

Riassunto

Come espressione universale della creatività umana, la musica è in grado di comunicare con grande dettaglio e complessità. Questa complessità non è codificata nella partitura o nei suoni, ma è piuttosto costruita nella mente dell'ascoltatore sotto forma di esperienze percettive articolate, denominate "ascolto strutturale". Mentre queste esperienze sono in qualche misura accessibili all'introspezione, esplicitata nella letteratura teorico-musicale, i sottostanti meccanismi cognitivi sfuggono all'indagine empirica. In questa tesi, formalizziamo l'esperienza della struttura musicale nel contesto della cognizione Bayesiana come una forma di inferenza: nello specifico, l'inferenza di rappresentazioni strutturali come risultato della necessità biologica di "dare senso" ai segnali sensoriali. Sfruttando un'analogia computazionale con il linguaggio, modellizziamo l'emergere di interpretazioni strutturali per un certo idioma musicale in termini di *parsing* incrementale sulla base di una grammatica.

In una serie di esperimenti, testiamo alcune implicazioni cruciali di questo approccio: (1) la rilevanza cognitiva delle categorie sintattiche specifiche di un particolare idioma, esemplificate dalla nozione di funzione armonica nell'armonia estesa tonale, (2) l'esistenza di rappresentazioni strutturali astratte dall'informazione meramente sensoriale, verificata adattando al dominio musicale un paradigma di *priming* strutturale, (3) l'esecuzione di computazioni cognitive che implementano il processo di *parsing* incrementale in tempo reale durante l'ascolto e (4) l'esistenza di meccanismi di rianalisi retrospettiva, in analogia con l'effetto *garden path* nel linguaggio.

Complessivamente, questi risultati contribuiscono a fornire prime evidenze a favore di alcuni pilastri di una teoria cognitiva dell'ascolto strutturale a livello computazionale e algoritmico. Sono compatibili con un processo di inferenza implementato attraverso computazioni di *parsing*, tra cui l'integrazione di nuovi eventi in una rappresentazione preesistente, la proiezione di eventi attesi in futuro, e la revisione retrospettiva dell'interpretazione di eventi passati. Sulla base del paradigma proposto, lavori futuri potranno ulteriormente testare le implicazioni di diversi modelli algoritmici di *parsing*, al fine di convergere ad un modello cognitivamente plausibile in modo simile a come in psicolinguistica vengono disambiguati i modelli di comprensione in tempo reale.

Parole chiave: struttura musicale, ascolto strutturale, grammatica, *parsing*, percezione musicale, rianalisi retrospettiva, musica e linguaggio

Contents

List of Figures

List of publications

Cecchetti, G., Herff, S. A., & Rohrmeier, M. A. (in review). Priming of abstract harmonic structure in music. *Journal of Experimental Psychology: Human Perception and Performance*
Included as **Chapter 5**

Cecchetti, G., Tomasini, C. A., Herff, S. A., & Rohrmeier, M. A. (2023). Interpreting rhythm as parsing: Syntactic-processing operations predict the migration of visual flashes as perceived during listening to musical rhythms. *Cognitive Science*, **47**(12).
https://doi.org/10.1111/cogs.13389
Included as **Chapter 7**

Herff, S. A., Bonetti, L., **Cecchetti, G.**, Vuust, P., Kringelbach, M. L., & Rohrmeier, M. A. (2023). Hierarchical syntax models of music predict theta power during music listening. *bioRxiv*.
https://doi.org/10.1101/2023.05.15.540878

Herff, S. A., **Cecchetti, G.**, Ericson, P., & Cano, E. (2023). Solitary Silence and Social Sounds: Music influences mental imagery, inducing thoughts of social interactions. *bioRxiv*.
https://doi.org/10.1101/2023.06.22.546175

Laneve, S., Schaerf, L., **Cecchetti, G.**, Hentschel, J., & Rohrmeier, M. (2023). The diachronic development of Debussy's musical style: A corpus study with Discrete Fourier Transform. *Humanities and Social Sciences Communications*, **10**(1). https://doi.org/10.1057/s41599-023-01796-7

Cecchetti, G., Herff, S. A., Finkensiep, C., Harasim, D., & Rohrmeier, M. A. (2023). Hearing functional harmony in jazz: A perceptual study on music-theoretical accounts of extended tonality. *Musicae Scientiae*, **27**(3). https://doi.org/10.1177/10298649221122245
Included as **Chapter 6**

Cecchetti, G., Herff, S. A., & Rohrmeier, M. A. (2022). Musical Garden Paths: Evidence for Syntactic Revision Beyond the Linguistic Domain. *Cognitive Science*, **46**(7).
https://doi.org/10.1111/cogs.13165
Included as **Chapter 8**

Herff, S. A., **Cecchetti, G.**, Taruffi, L., & Déguernel, K. (2021). Music influences vividness and content of imagined journeys in a directed visual imagery task. *Scientific Reports*, **11**(1), 15990.
https://doi.org/10.1038/s41598-021-95260-8

Herff, S. A., Harasim, D., **Cecchetti, G.**, Finkensiep, C., & Rohrmeier, M. A. (2021). Hierarchical syntactic structure predicts listeners' sequence completion in music. *Proceedings of Annual Meeting of the Cognitive Science Society*, **43**. https://escholarship.org/uc/item/9w44g4x1

Cecchetti, G., Herff, S. A., & Rohrmeier, M. A. (2021). Musical syntactic structure improves memory for melody: Evidence from the processing of ambiguous melodies. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **43**.
https://escholarship.org/uc/item/985452gt
Included as **Chapter 4**

Cecchetti, G., Herff, S. A., Finkensiep, C., & Rohrmeier, M. A. (2020). The experience of musical structure as computation: What can we learn? *Rivista Di Analisi e Teoria Musicale*, **26**(2), 91–127. https://doi.org/10.53152/1032

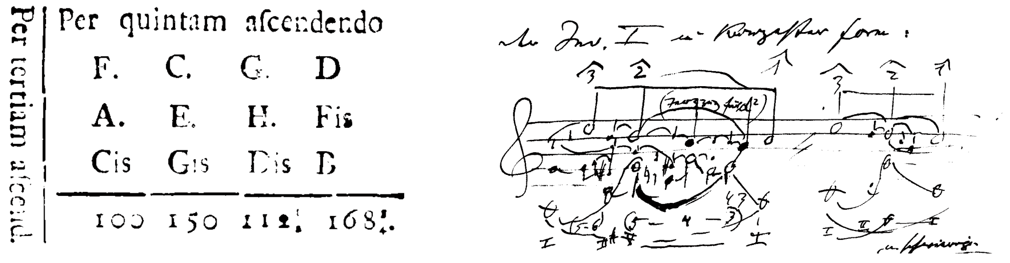
Introduction Part I

1 Structural hearing in theory and cognition

1.1 Investigating structural illusions

This thesis is concerned with the empirical investigation of a somewhat elusive topic: the mental representation of syntactic structure in music. In first approximation, by "structure", we mean here the kind of information about music that is typically discussed in music analysis and visualised in music-theoretical graphs such as the ones displayed in Figure 1.1; by "representation of structure" or "structural representation", we mean then brain or mental states that encode such information.

The main characters of the story we are going to address are three. First, the musical surface, which can be loosely understood as "what is actually presented to the listener". The musical surface forms the input of cognitive processing and perception. Although, properly, only the audio signal qualifies as "surface", the term is often used to refer to symbolic abstractions of the auditory signal (e.g., a score). In music theory, even more abstract representations of the



(**a**) Euler (1773)

(**b**) Schenker (1929)

Figure 1.1 – Two graphic representations of music-theoretical relations. In (**a**), L. Euler's *Tonnetz* displays the relatedness between types of triads in terms of their proximity in a geometric space. Fifth relations appear on the horizontal axis, while third relations are arranged on the vertical axis (major thirds) and the diagonal (minor thirds). When a chord of a certain type is employed in an actual musical surface, its relationships with other chords reflect the geometry of the *Tonnetz*. In (**b**), the analyst's introspection about the relations between actual tones in a piece of music (here, J.S. Bach's C-major invention BWV 772) is communicated by drawing slurs between notes in a score-like Schenkerian graph.

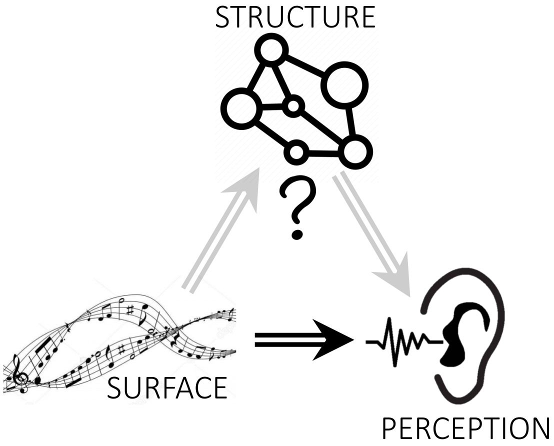


Figure 1.2 – The "music itself" (surface), the music-theoretical characterisation of the music (structure), and the listener's experience (perception). Does structure mediate the dependency of perception from the musical surface? If so, how?

musical surface as also often adopted as a starting point for analysis, such as sequences of symbols each identifying a chord. Second, we will deal with musical structures, which music theorists associate with musical surfaces as a result of their domain knowledge and introspection. Finally, the listeners' perceptual and cognitive experience resulting from exposure to the surface. In empirical settings, the listeners' experience is operationalised as an observable "response" to the musical surface, influenced by personal features of the listener such as their previous musical exposure or training. The question is whether the causal path from the surface to the listener's experience is mediated by (representations of) music theoretical structures and, if so, why and how (Figure 1.2). The working hypothesis, to be addressed empirically, is that aspects of the listening experience emerge as a byproduct of cognitive processes that produce and manipulate representations of musical structure.

The relationship between music-theoretical structure and the listener's experience of a piece of music is far from trivial. In some cases, the intentional decisions of a composer that determine how (virtually) every aspect of the musical surface is structured are fully specified, possibly to an algorithmic level of precision. In John Cage's *Music of Changes* (Figure 1.3), for example, each sounded event and each silence are determined by sampling options from pre-defined 8 × 8 charts, following a traditional Chinese divinatory text. Although the constructive principles of the piece are relatively simple and clear, it is very unlikely that a listener may experience the piece as being structured according to those principles. The problem, here, is that the piece on its own is unlikely to provide enough information for the listener to infer what the underlying structure is, either consciously or unconsciously: even knowing the compositional process explicitly does not fully remove an (intentional) sense of indeterminacy. By contrast, in tonal music, listeners can benefit from exposure to a wide repertoire of instances of the style, where the same constructive principles are applied in different ways while being reinforced through cultural transmission and explicit pedagogy. As a consequence, it is more likely that listeners may acquire explicit or implicit knowledge of

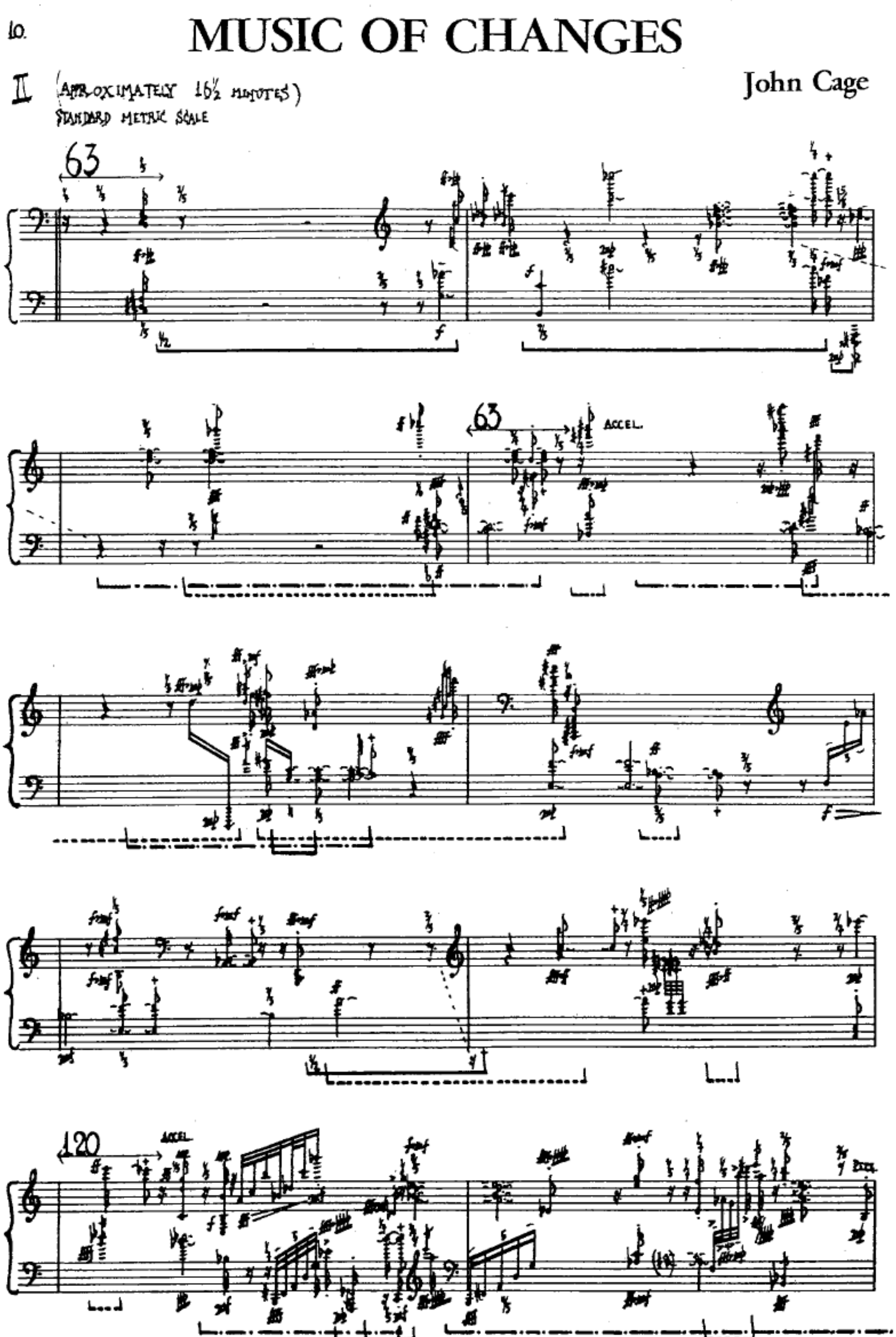


Figure 1.3 – A page of J. Cage's *Music of Changes.* The piece intentionally sounds in some sense "indeterminate", yet it is the result of an algorithmically well-specified (though stochastic) compositional process.

the structuring principles underlying the style (Rohrmeier et al., 2012; Tillmann et al., 2000). Nevertheless, even in the context of tonal music, it is an open question whether and in what sense listeners do experience music as structured accordingly to music-theoretical accounts.

The main problem we face in addressing this issue is that structural representations in music are, almost by definition, ephemeral. The objects of representation, like tones, chords, and structural relations in general, are *latent*: they are not observable in the score or in the auditory signal. If anything, representations of musical structure are "hallucinations" that only exist in the mind of the composer, analyst, or listener, which qualifies them as primarily cognitive phenomena. They are also largely *conventional*, in the sense that the objects that are represented are not necessarily motivated by "first principles" (e.g., causal relations, laws of physics, ...) other than their arbitrary use in the musical practices of a specific community. Altogether, latency and conventionality entail that the only ground-truth regarding the nature of structural representations is accessible through the introspective self-report of those who share a common cultural background. In this respect, musical structural representations are not unlike linguistic syntactic representations: there, too, the ultimate ground-truth lies in the introspection of the speaker (Chomsky, 1965).[1]

In language, this introspection is easy to access, even for non-explicitly-trained speakers (e.g., children). When exposed to a sentence like "The cat chased by the dog fell in the river", speakers can easily report who (the cat or the dog) fell in the river. In turn, this entails the successful inference of a structural connection between the verb "fell" and one of the two nouns ("cat" or "dog") as its subject – an inference that cannot trivially rely on superficial features such as word proximity. This supports the belief that speakers operate based on representations that encode syntactic dependencies, and that these representations are formed (to some extent) automatically based on implicit knowledge (Reber, 1989; Rebuschat, 2022; Shtyrov et al., 2010). After all, one primary function of language is the exchange of propositional content in communication, which would not be possible without the successful inference of structural relations (Jackendoff, 2002a; Tanenhaus and Trueswell, 1995).

On the contrary, the capacity of music to fulfil many of its functions does not crucially depend on the successful inference of arbitrarily complex structural features, not even when these are intended by the "composer". While music is certainly capable of expressing intentionality, it is not a requirement that all the participants in a musical interaction agree on the content of such intentionality (cf. "floating intentionality"; Cross, 2014). As a consequence, listeners can exploit music for joint movement (Clayton, 2012), social bonding (Cross, 2009; Savage et al., 2021), construing a social identity (Born, 2011), or inducing mental imagery (Herff, Cecchetti, et al., 2021; Taruffi and Küssner, 2019) with or without inferring such complex structures in full (although, complex structure may contribute to the fulfilment of these functions when present). This entails that the cognitive relevance of structural representations cannot be assumed *ipso facto*. Furthermore, even if listeners do form such structural representations,

[1] Although one might assume that linguistic syntax – more so than musical syntax – is constrained by non-conventional features of the external world linguistic representations – as opposed to musical ones – refer to.

they may lack the conceptual and terminological tools to identify and explicitly report their experience, unless they are trained to do so: musical structure is not typically associated with a standard semantics that can be exploited to report our structural understanding, as we can do in language (e.g., by answering the question "who, the cat or the dog, fell in the river?").

Since explicit introspective reports from untrained listeners are difficult to acquire, we are mainly left with two sources of information about structure as perceived in music: implicit manifestations in listeners as observed in experiments, on one hand, and explicit accounts by (trained) music theorists, on the other. The former kind of information is indicative of spontaneous effects in the general population, but such effects may not be exhaustive of structure-related phenomena: subtle effects may average out or be undetectable due to individual differences in the population. In turn, when dealing with expert knowledge, it is difficult in principle to disentangle what part of the observed behaviour is driven by implicit cognitive mechanisms, and what by expert knowledge itself (e.g., being explicitly instructed to respond in a certain way through reasoning). A third kind of evidence will not be directly addressed in the present work: namely, that acquired through computational and statistical modelling over large corpora (White, 2022). While such methodologies cannot directly demonstrate the cognitive relevance of specific structures, they can contribute to unearthing what structures may *in principle* be discovered in a certain repertoire of musical surfaces (e.g., Harasim, 2020; Laneve et al., 2023; Moss et al., 2019; White and Quinn, 2018).

As a starting point, two things are virtually uncontroversial. First, that listeners are sensitive to *some* structural features when listening to music they are (to some extent) familiar with. This is apparent, minimally, in the countless accounts of how statistical features of music are reflected in the listeners' brain and behaviour (Pearce, 2018). Second, that extremely complicated structures can be hidden and discovered among the notes of a musical piece. These are explicitly discussed in the fine-grained analyses of musicians and music theorists, who develop concepts, terms, and, ultimately, models for talking about structure in music based on skilled listening (Salzer, 1962), formal or quasi-formal models (Lerdahl and Jackendoff, 1983a; Schenker, 1935; Tymoczko, 2011; Yust, 2018), knowledge of historical and social contexts (Feld, 1984; Sanguinetti, 2012), embodied instrumental or vocal practice (Lester, 1995; MacRitchie et al., 2018; Rink, 2015), and more (Bent, 1990). However, whether and to what extent the insights coming from music theory can be related to the empirical findings of experimental psychology and neuroscience largely remains to be explored (Cross, 1998; Parncutt and Hair, 2012; Walsh, 1984).

In this work, we commit to the assumption that music-theoretical introspection provides us with reliable insights about what happens in the mind of the listener, in a very literal sense: namely, we assume that listeners form mental representations of structural features, to a greater or lesser extent, and that music theoretical introspection provides us with a characterisation of those representations. In the following, we intend to explore the putative implications of this commitment towards empirically observable cognitive processes. As a first step, we need to clarify in what way music theoretical statements, which are typically expressed in the form of music analyses, imply cognitively relevant representations. This is the object of this Chapter, where we address the topic from the perspective of Western (extended-)tonal music theory. In Chapters 2 and 3, we then introduce an approach to make such music-theoretically-inspired hypotheses concrete enough to become the object of empirical research. In the body of the thesis, Chapters 4–8, we will finally present several behavioural experiments that address the cognitive reality of structural representations and some computational-level features of the underlying processing mechanisms.

1.2 Structure and hearing in music theory

1.2.1 The musical surface

The notion of musical surface, despite its apparent simplicity, is treated somewhat idiosyncratically across and within disciplines such as music theory, music information retrieval, and music psychology (see Cambouropoulos, 2010 for a review). In music theory, the musical surface is typically understood as the "lowest level of representation that has musical significance' (Jackendoff, 1987, p. 219), comprising symbolic elements (e.g., notes) that play a similar role as phoneme do in language (Jackendoff, 1987; Lerdahl and Jackendoff, 1983a). Such symbolic representations are often visualised as scores or piano-rolls, which differ in terms of the kind of information that is encoded. For example, score notation typically assumes discretised durations expressed in metrical beat-units and encoded in the *shape* of the note symbols, whereas piano-roll notation may express durations continuously in seconds as reflected in the *spatial length* of a note symbol (Campos and Fuentes, 2016). In the music-theoretical practice, such representations of the musical surface often constitute the starting point for analysis (e.g., Figure 1.1b). Importantly, at times, music-theoretical models take a more abstract level than the note-level score as a starting point for analysis: e.g., progressions of chords symbols, or of roman-numerals.

From a psychological perspective, the musical surface has been located as "the boundary between perception and cognition" (Wiggins, 2007, p. 325), although it is unclear to what extent percpetual (as opposed to perceptual and cognitive) mechanisms are sufficient to abstract, e.g., a score-like representation from the auditory signal (Cambouropoulos, 2010). Overall, a more principled approach might be to place the only meaningful boundary – where music "as given" (from the outside as, e.g., an audio signal) becomes a latent (i.e., unobserved) mental object – at the interface between the auditory signal and the sensory system. After all, even the very early stages of auditory processing (e.g., the tonotopical filtering on the basilar membrane) introduce non-trivial levels of representation that become relevant for later stages of processing (e.g., in terms of pitch percepts, Cariani and Delgutte, 1996). Incremental stages of processing introduce a spectrum of more and more structured representations, each constituting a different abstraction of the originally observed signal.

For the purpose of the present discussion, we rather stick to the (somewhat blurry) music-theoretical notion of the musical surface as the starting point for analysis (as encoded in a symbolic, score-like notation). We do this as a necessary simplification, as our focus is on the kinds of structures that build on top of the kind of information that is encoded in a score. It remains understood that a score-like notation, as an intermediate level of abstraction from what is perceived with the senses, is to some extent already structured and is not truly "given" as part of the external signal.

1.2.2 Two analyses

Its apparent simplicity notwithstanding, the opening theme of W. A. Mozart's KV 331 piano sonata (Figure 1.4a) has been the object of many divergent analyses (cf. Allanbrook, 2008 for an overview), two of which are exemplified in Figures 1.4b,c.[2] For brevity, we will focus on the theme's consequent phrase only (mm. 5-8). In the analysis by Schenker (Figure 1.4b; Drabkin et al., 1995; Schenker, 1935), the entire phrase is understood as a descent from the *Kopfton*[3] e_5 (scale degree $\hat{5}$ in the key of A major) to the tonic a_4 through several *passing* steps (d_5, $c\sharp_5$, b_4). In particular, in this analysis, the structural $\hat{5}$ is understood as being prolonged over several measures, until scale degree $\hat{4}$ is finally encountered to start the descent at the end of m. 7. Supporting this understanding, the e_4 sounding in the left-hand part as a pedal makes the sustained *Kopfton* audible (yet skillfully concealed) until the descent begins.

In the analysis by Morgan (Figure 1.4c; Morgan, 1978), by contrast, the *Kopfton* is understood to be the $c\sharp_5$ (scale degree $\hat{3}$) in m. 1: this tone is prolonged until the descent to $\hat{2}$ and then $\hat{1}$ is finally realised in mm. 7-8. In the latter analysis, the d_5 ($\hat{4}$) in m. 7 is understood as part of a *neighbouring* motion connecting the initial $c\sharp_5$ to the $c\sharp_5$ in m. 8. This is sharply different from the interpretation of the very same tone, d_5, in Schenker's analysis: here, the d_5 is understood as part of a *passing* motion connecting scale degrees $\hat{5}$ and $\hat{1}$.

The differences between the two analyses have several implications. The two instances of $c\sharp_5$ in mm. 7-8 are not to be understood as a return to the same tone in Schenker's analysis, as they belong to different voices in the free-polyphonic texture: the $c\sharp_5$ in m. 7 is part of an inner voice descending from the initial *Kopfton* through scale degree $\hat{4}$ in m. 6 while the *Kopfton* $\hat{5}$ is being prolonged in the topmost voice; the $c\sharp_5$ in m. 8 is instead part of the descending motion of the topmost voice as it traverses the *Urlinie*. Another implication of Schenker's analysis is that the *descending* motion of the topmost voice from $\hat{3}$ to $\hat{1}$ in m. 8, including the upbeat $\hat{4}$, and the *ascending* motion from $\hat{1}$ to $\hat{3}$ in m. 7 belong to two different voices: the former may be understood as a response (in *stretto*) to the latter in a two-part texture. On the contrary, Morgan's analysis may rather suggest a continuity in the arch-like gesture across mm. 7-8,

[2] Here and elsewhere in this thesis, it is not important that the reader finds any of the discussed analyses to be *the most plausible* analysis of this passage of music (e.g., as representative of the composer's intention) or the one that they can most intuitively relate to. We would rather invite readers to see these as *some* plausible (as opposed to "the most" plausible) analyses.

[3] In the Schenkerian analytical framework (Cadwallader and Gagne, 2010; Schenker, 1935), tonal pieces are understood as contrapuntal elaborations of a descending melodic line, the *Urlinie*, harmonically supported by a bass motion. The term *Kopfton* refers to the topmost tone of the *Urlinie*.

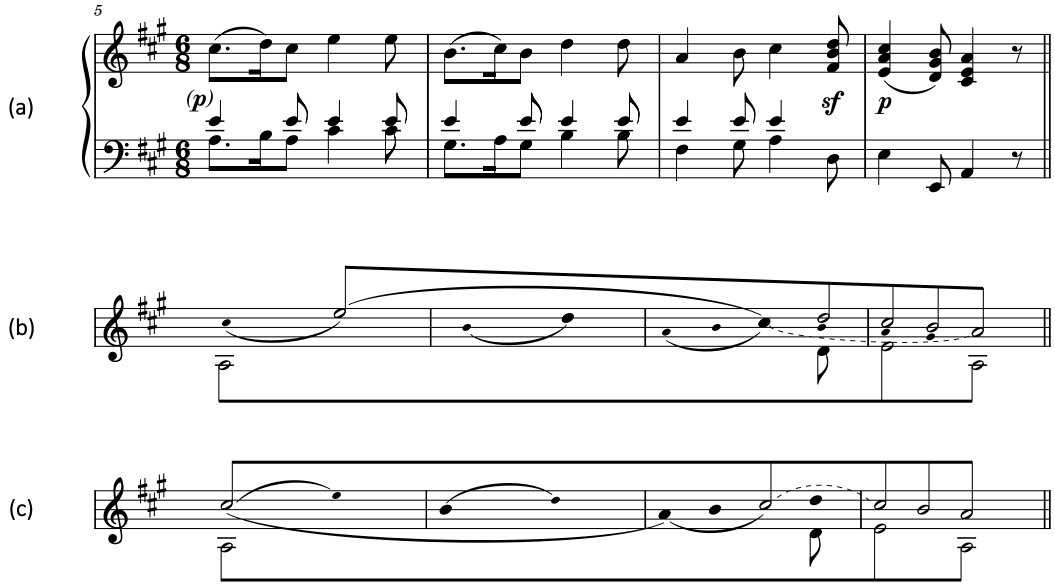


Figure 1.4 – **(a)** W.A. Mozart, *Sonata* for piano KV331, i, mm. 5ff. **(b)** Analysis after H. Schenker (adapted from Schenker, 1935). **(c)** Analysis after R. Morgan (adapted from Morgan, 1978).

emphasised by the rocking quality of the *siciliana* rhythm (♩ ♪ ♩ ♪).[4]

What do analyses like these tell us about the relationship between the musical surface, the proposed structural descriptions, and our listening experience? We argue that the structures conjured by analysts serve a twofold purpose in this respect: they are *internally* explanatory towards the musical surface, and *externally* explanatory towards the listeners' experience. Importantly, "explaining" the appearance of the musical surface and "explaining" the listener's perceptual experience are, in principle, independent goals. However, in many cases, analyses are interpreted as fulfilling both functions. In the following, we discuss the role of analyses as internal and external explanations in turn, and we propose an argument for why internally-explanatory analyses often happen to be externally explanatory, and why analyses that are externally explanatory are typically formulated as being internal explanations too.

1.2.3 Structural interpretation as internally explanatory

Minimally, the structures conjured by analysts aim at unveiling the *intentionality* that underlies the particular arrangement of events in a musical composition. Specifically, analyses typically imply an *explanatory* narrative that links a proposed structure to the surface (be it a score or the audio of one of its performances). Such a narrative should characterise the arrangement of all elements of the surface as intentional, as opposed to random or arbitrary: it does so by

[4] The slur over d_5 in Figure 1.4c is implied in Morgan's analysis and explicitly notated in Lester (1979). In Lester's analysis, which is otherwise incompatible with this understanding, the slur is likely introduced with the purpose of highlighting the intuitive relatability of the neighbour-tone interpretation.

providing an explanation for why the musical surface looks the way it does.[5] We have seen examples of such explanatory narratives in our discussion of Mozart's theme: for example, in Schenker's analysis, the d_5 in m. 7 *is there in order to* connect the descent from the *Kopfton* e_5 with a passing tone, and the e_4 pedal *is there in order to* make the prolongation of $\hat{5}$ audible. The *narrative* nature of the analysis is not just rhetoric coloring: the storyline and the choice of wordings that the analyst adopts provide meaningful and subtle characterisations of the structural relations that hold among the elements of the surface. For example, the use of the word "passing" (as opposed to "chasing" or "avoiding") ascribes to d_5 a transitory and dynamic character as part of a directed motion. In this sense, analysis is *attributive*: it attaches meaningful attributes to sounded events (Boretz, 1977; Dubiel, 1990). These attributes, in turn, may manifest themselves in aspects of listening to, and performing, music.

What kind of arguments make for a convincing explanation of the surface? Typically, an analysis points at *relations* between events of the surface, or between events in the surface and abstract concepts, to demonstrate the intentionality of the composition: events are not there randomly, they are meant to realise some relationships, or as the observable expression of something latent. Analytical narratives will typically name some entities $e \in E$, spanning across different levels of abstraction (e.g., from individual notes to chords or keys). A particular piece is then seen as the result of combining instances of those entities according to some set of relations. More precisely, a structure may be thought of as a graph $s = (N, L)$ where edges $l \in L \subseteq \{(n_1, n_2) \in N^2 | n_1 \neq n_2\}$ indicate that some of the nodes $n \in N$ are somehow *related* to one another (Figure 1.5).[6] Each node in the structure-graph is an instance of an entity, and may be labelled as such through a mapping $I_E : N \to E$. As part of their explanatory narrative, analysts justify the relatedness between nodes (i.e., the existence of an edge linking nodes in the structure) by appealing to some generic relations that may hold between entities. Such relations, that collectively comprise a set R_E of relations defined on E, can reflect properties of the entities as elements of mathematical spaces (e.g., symmetry properties), or relations with specific semantics (e.g., being contrasting or preparatory). The analyst's narrative then also defines a labelling $I_{R_E} : l \mapsto r_l$ of edges by mapping each edge $l = (n_1, n_2) \in L$ between nodes $n_1, n_2 \in N$ to the relation $r_l \in R_E$ that justifies its existence, provided that $I_E(n_1) \overset{r_l}{\sim} I_E(n_2)$.

Overall, an analysis can be seen as the attribution of an *interpretation* to the musical surface, where by interpretation we mean a structure (as reflected in a graph $s = (N, L)$) together with a narrative (as reflected by a labelling of the graph given by some I_E, I_{R_E}). Such a narrative (and, by extension, the corresponding interpretation) is *internally* explanatory towards the

[5] Note that the analyst's explanation is meant to characterise the musical surface as intentional, as opposed to random or arbitrary, but not necessarily to identify the explicit intentions of the composer: in other words, it is not necessarily the case that any analytical statement corresponds to an explicit choice on part of the composer, or one that the composer is consciously aware of. More generally, it is not the case that the listener should converge to the same interpretation that the composer may have intended (Cross, 2014).

[6] Musical structures are indeed often visualised in graphical form to make relevant entities explicit (Finkensiep and Rohrmeier, 2021; Rohrmeier, 2020b; Yust, 2006, 2018). For example, in a Schenkerian analysis such as those displayed in Figure 1.4, slurs provide a semi-formal visualisation of edges linking notes with one another (see, e.g., Ericson et al., 2023 and Kirlin, 2014; Kirlin, 2008 for a graph-theoretical encoding of Schenkerian analyses).

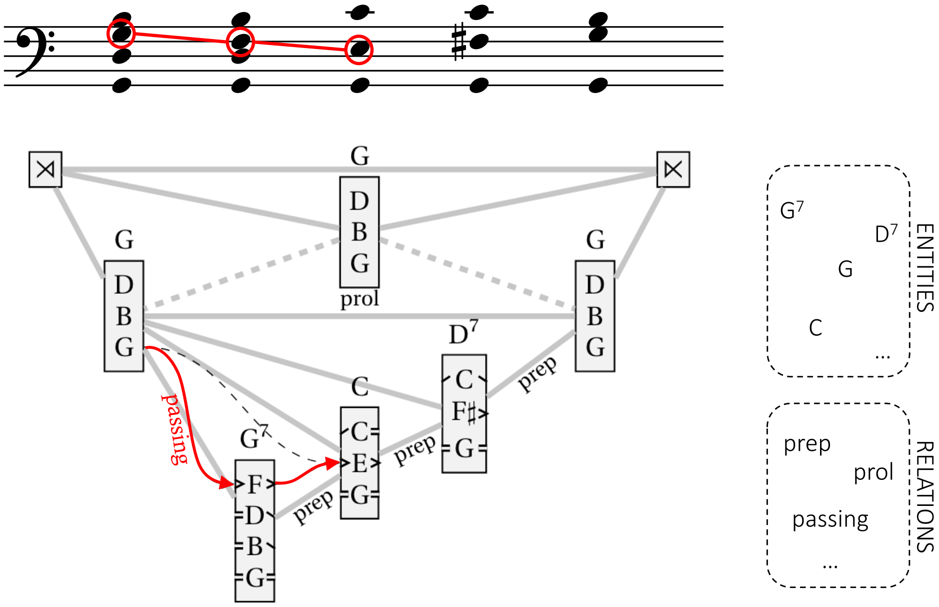


Figure 1.5 – Structural interpretation of a contrapuntal progression, expressed as a labelled proto-voice graph (after Finkensiep, 2023). Each node in the graph is labelled as an instance of one of the possible entities gathered on the right. Similarly, each edge is labelled as an instance of one of the possible relations, including harmonic relations between chords (e.g., the *prol*ongation of the same harmony, or the *prep*aration of a chord with its dominant) and contrapuntal relations between individual tones (e.g., the *passing*ness of a tone as part of a motion connecting other two tones, highlighted in red). For better visualisation, only some edges and labels are displayed.

surface in the sense that it accounts for how the presence of each element of the surface can be justified as the logical consequence of given features of the structure. By "internal", we mean here that the *explanandum* (the musical surface) is entirely made up of musical materials (pitched events, timbres, or anything a particular style considers to be the materials for building their musical surfaces): extra-musical knowledge sourced from other domains (e.g., history, performance, perception) can be invoked to conjure such explanations, but it is not part of the *explanandum* at this level.

Importantly, the association of a surface with a structural interpretation is not deterministic, as we exemplified with Mozart's theme: the same surface can be interpreted in different ways – possibly, with different likelihoods. However, interpretations are also not arbitrary: they are constrained by criteria of specificity, parsimony, compositionality, and generalisability.

Specificity: structural cues

In order for an explanation to be useful, as many aspects of the musical surface as possible need to find a place in the explanatory narrative, so that the latter is convincingly telling us something that is *specific* to that particular surface. Aspects of the surface that are particularly well explained by one analysis compared to a different one can be considered to be (structural) *cues* that favour that analysis. The musical surface in Mozart's theme contains plenty of such cues favouring either of the two analyses we presented. The persistent presence throughout mm. 5–7 of e_4 as a pedal in the left-hand part finds an elegant explanation in Schenker's analysis: effectively, the pedal makes the prolonged *Kopfton* $\hat{5}$ audible, yet concealed in an lower register. Similarly, the sudden *sforzando* on $\hat{4}$ in m. 7 seems rather compatible with emphasising the break between the ascending voice $a - b - c\sharp$ and the descending voice $(d-)c\sharp - b - a$ implied by Schenker's analysis, as opposed to the arch-like line in Morgan's. In turn, the metrical prominence of the initial $c\sharp_5$ compared to the weaker e_5 is commonly invoked as a cue favouring an analysis like Morgan's where $c\sharp_5$ is the *Kopfton* (Lerdahl and Jackendoff, 1983a).

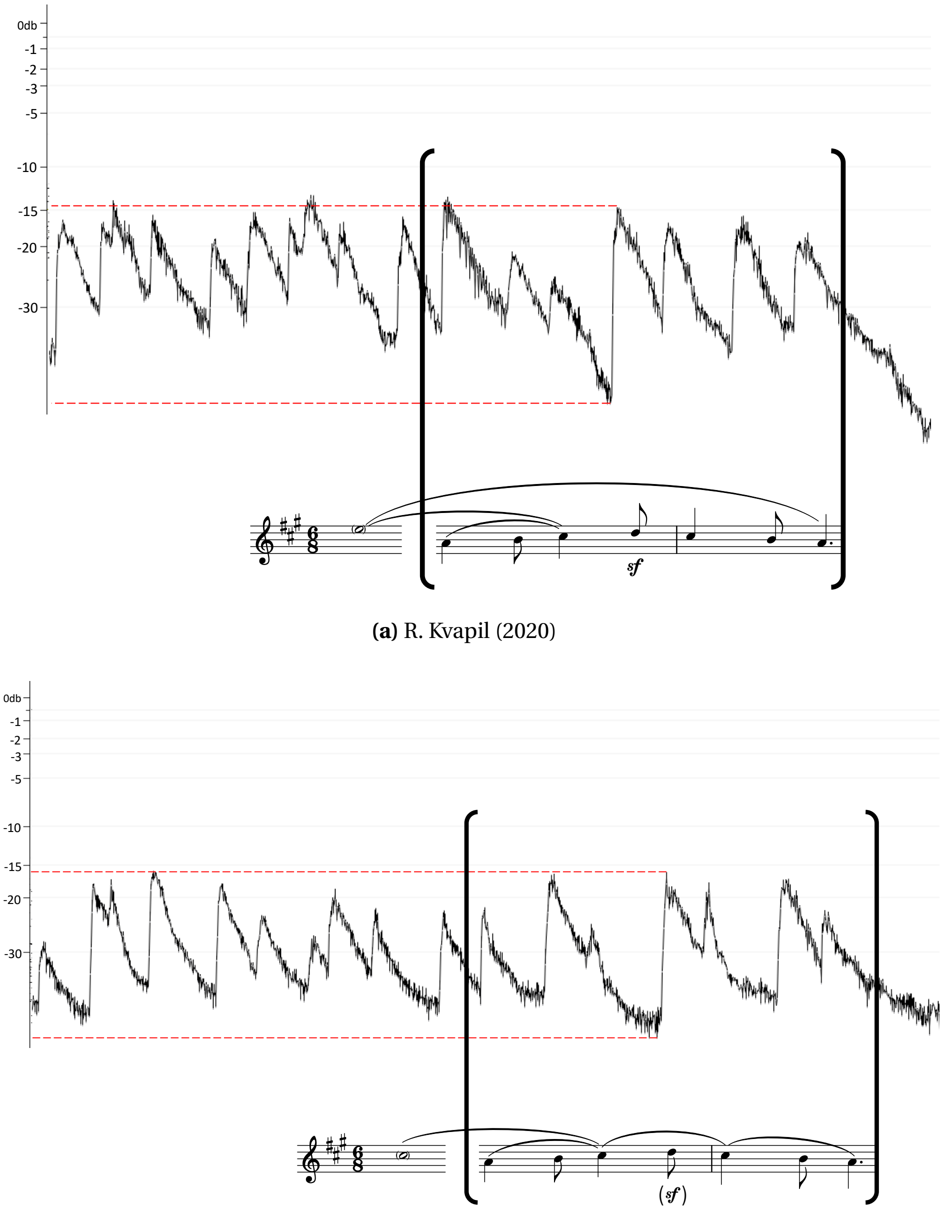
Structural cues may also be provided by specific performances (Cook, 1999). Figure 1.6 shows the comparison between two recordings of Mozart's theme, one by Radoslav Kvapil on the fortepiano (Musical Concepts, 2020), and one by Aldo Ciccolini on the piano (Musique Classique, 1985). In the former recording, a pronounced, closural diminuendo on the ascending $a_4 - b_5 - c\sharp_5$ in m. 7, and a marked *sforzato* on the following d_5, might encourage us to consider the ascending motion as separated from the following descending motion: the resulting "question-answer" responsorial dynamics is particularly consistent with Schenker's analysis. Vice versa, in Ciccolini's recording, the rhythmic cell ♩ ♪ ♩ is repeated almost identically across mm. 7–8, with a mellow *sforzando* on d_5: this rather highlights the continuity of the arch-like gesture $c\sharp_5 - d_5 - c\sharp_5$ across mm. 7-8 – a trait that is more in line with Morgan's analysis. Overall, structural cues can bias the plausibility of one interpretation over another by providing audible features of the musical surface that impose additional constraints to be explained by the analysis.

Parsimony and compositionality: structural primitives

The intentionality underlying musical composition lies in the commitment, on part of the composer, to some limitations to the arbitrariness of their creative agency. That is, intentionality is often recognised (1) in the creative use of finite means to produce potentially infinite outcomes (cf. Chomsky, 2002; Hofstadter, 1979), as well as (2) in the intentional choices[7] when using those means to conjure a particular instance among those that are possible in principle (i.e., what makes the piece "what it is" as opposed to "something significantly different", Dubiel, 1990; cf. Boretz, 1977).

In order to constitute a useful explanation, then, an interpretation should appeal *parsimoniously* to a limited set of primitive entities E and relations R_E, from which "digital infinity" can emerge through compositionality (Chomsky, 1957; Hofstadter, 1979). Interpretations of individual pieces can then be seen as instances of a "music theory", which specifies such sets E and R_E of primitives as well as criteria for combining them to form individual struc-

[7] Cf. note 5 above: recognising *that* the composer made intentional choices in arranging the musical surface as observed does not necessarily require the analyst to identify *what* those choices are, nor that those choices were made consciously by the composer.



(a) R. Kvapil (2020)

(b) A. Ciccolini (1985)

Figure 1.6 – Waveform amplitude (dB) of two performances of Mozart's theme, mm. 5-8. In R. Kvapil's recording on the fortepiano **(a)**, a marked diminuendo can be heard throughout m. 7, interrupted by the *sforzato* on d_5 on the upbeat of m. 8. On the contrary, A. Ciccolini's performance on the piano **(b)** is more uniform across mm. 7-8, with a mellow, barely hinted-at *sforzato*. The difference between the two renditions of *sforzato* is reflected in the dynamic contrast introduced by d_5: in comparison to other attack transients, this is much larger in (a) than (b) as highlighted by the red dashed lines. Overall, Kvapil's performance in (a) seems to provide cues that are consistent with Schenker's analysis (displayed under the waveform), whereas performance cues by Ciccolini in (b) are rather consistent with Morgan's analysis.

tures. Schenkerian theory is an example of such a music theory: its primitive entities are tones, and its primitive relations comprise contrapuntal concepts pertaining to pitch – such as passingness, neighbourness, or consonance – and to temporality – such as simultaneity and sequentiality – (Cadwallader and Gagne, 2010; Drabkin et al., 1995; Finkensiep, 2023; Schenker, 1987). Schenker's and Morgan's analyses are then two instances of this music theory. In Schenker's analysis, the d_5 in m. 7 of Mozart's theme is interpreted as a *passing* tone in the descent from e_5 to a_4, and the e_5 that starts the passing motion is related to the a_4 that ends it: the relation that links them is the horizontalisation (i.e., the arrangement in sequential order as opposed to simultaneously) of two tones that are consonant expressions of the same harmony (A major). In turn, Morgan's analysis appeals to the same entities and relations, yet combined in a different manner: the d_5 is understood as a *neighbour* tone of $c\sharp$, and the e_5 in m. 5 is not directly related to the final a_4 but rather to the initial $c\sharp$ as a horizontalised expression of an A major harmony.

Generalisability: musical idioms

Parsimony and compositionality entail that interpretations, even when they are developed "from scratch" to account for one piece in isolation,[8] should be seen as the (possibly only) instance of a music theory that – in principle – can express interpretations for other (possibly non-existent) pieces too. However, musical pieces are hardly ever to be considered in isolation: individual musical surfaces are typically understood as instances of a particular musical idiom, which is represented (or, if it is not, may *in principle* be represented) by several such instances. One form of music-theoretical validity for an interpretation, then, is given by its *generalisability*: namely, the assumption that a given interpretation for a given musical surface should be an instance of a music theory that also provides internally-explanatory interpretations for all other surfaces that are considered instances of the same musical idiom.

Generalisability poses a significant constraint on what interpretations may be validly internally explanatory towards given surfaces: it is certainly not the case that any arbitrary explanation we may find for a given musical surface would generalise in the above sense. On the other hand, there is no guarantee that a given surface only affords one explanatory interpretation under a given music theory (we have seen how Mozart's theme affords at least two interpretations under Schenkerian principles), nor that a given surface should only be explainable through one set of general principles, i.e., as being part of the repertoire of just one musical idiom.[9] Overall, music theories identify general principles, in the form of sets of entities and possible relations that may in principle hold between them, for forming generalisable internally-explanatory interpretations of musical surfaces seen as expressions of a given repertoire.

Importantly, generalisability is defined with respect to musical surfaces: the generalisability of

[8] As is often the case in *avantgarde* music, cf. Figure 1.3.

[9] For example, many pieces in late common-practice tonality can be internally explained through both Schenkerian and Tonfeld-theoretic principles, though possibly with different degrees of plausibility; Haas, 2004; Polth, 2006.

the internally-explanatory interpretation of a given surface is a consequence of that surface being part of a specific set of musical surfaces, a repertoire, against which generalisability is assessed. This makes it possible, in principle and under some assumptions about the general shape such interpretations can take, to infer internally-explanatory interpretations for individual surfaces, as well as the music theory that makes them generalisable to the entire repertoire they belong to, just by looking at the set of surfaces comprising that repertoire (see, e.g., Clark, 2013a for a formal account in terms of unsupervised grammar induction, and Harasim, 2020 for an application to the musical domain; cf. Jacobs and Kruschke, 2011; Tenenbaum et al., 2006 for analogous Bayesian models of learning in the context of Bayesian cognition).

In summary, generalisable internal explanatory power constitutes the relationship between structural interpretations, as they appear in the music-theoretical discourse, and the musical surface: the former capture the intentionality that underlies the appearance of the latter. However, in principle, multiple music theories are possible for each repertoire, and multiple interpretations for each surface, that satisfy the above criteria: how does an analyst decide which music theory and which interpretation to commit to when approaching a given musical surface? While this question admits various answers,[10] it is certainly a widespread assumption when approaching tonal music that "understanding tonal organisms is a matter of hearing" (Salzer, 1962). In other words, internally-explanatory interpretations are expected to say something about "hearing" too. After all, the intentionality of the compositional process is not only directed at shaping the musical surface *per se*, but also (and possibly primarily) at engendering specific "effects" of auditory experience in the listener (Dubiel, 1990; Polth, 2006). This turns music theories from theories *about* the musical surface to theories *about* listening.

1.2.4 "Hearing as": external explanatory power

By choosing certain (internally) explanatory narratives over others, music theories identify classes of possible perceptual effects in a given musical idiom, and encourage the listener to adopt listening strategies that make it possible to experience such effects in a given piece or repertoire. The validity of a music theory and, at the same time, the reality of the effects it suggests as phenomena of perceptual experience are "proven" introspectively by (1) the acknowledgement of a difference in our own perceptual experience when the proposed listening strategy is adopted, as well as (2) the "failure" of other pieces or repertoires to engender those particular differences when the same listening attitude is adopted (Polth, 2006, p. 170).

[10] What external criteria to invoke in support of analytical claims is the object of debate in music theory (M. Brown and Dempster, 1989; Dubiel, 2000; Tymoczko, 2020). In particular, for some theories and musical idioms, the listener's perception does not seem to be an appropriate criterion of analytical validity: for example, pitch-class set theory (Forte, 1973) is successfully internally explanatory towards twelve-tone serialist music, irrespectively of the expectation that set-theoretic transformations are audible in any meaningful sense (Bruner, 1984). In this case, a criterion of external validity may rather be the adherence of the theory to biographical and historiographical evidence about the composer's artistic intentions.

What does it mean to "hear" an analytical interpretation? First, in a neutral sense, hearing refers to the musical surface as an auditory signal that is presented through performance (or simulated through audiation; Gordon, 1985). Structural cues, in particular, are physically part of the musical surface: they are directly audible. Since structural cues correlate with a particular analysis (or class of analyses), listeners that (consciously or unconsciously) identify structural cues exhibit an experience of music that has a clear correspondence to the given analysis (or class of analyses) (Lester, 1979). For example, listeners can typically detect boundaries between musical phrases even in unfamiliar musical idioms, possibly thanks to the presence of clear auditory cues that mark phrase endings and the following gap (Lartillot and Ayari, 2009; Popescu et al., 2021). Nevertheless, hearing the structural cues, and even responding to them as such (for example, by marking with a nod every phrase boundary), is not the same as "hearing" the analytical interpretation that events preceding the boundary and those following the boundary are related to one another to form groups.

The difference between "plain" hearing and structural hearing proper lies in the kind of information that is attached to events in the musical surface. In plain hearing, only information pertaining to the intrinsic acoustic features of each event, such as its pitch and temporal location, are encoded: this is the *surface identity* of events. Structural hearing, instead, consists of *hearing* events in the surface *as* having additional features that are not part of the auditory signal (Dubiel, 2017). Specifically, structural hearing entails attributing to events a *structural identity*, i.e., an encoding of the attributes that each event inherits from its being part of a particular internally-explanatory interpretation (Dubiel, 1990, 2017).

For example, the sets of (internally-explanatory) primitives assumed by Schenkerian theory such as "passingness", "neighbourness", and "consonance" are proposed by Schenker as reflecting elementary "auditory" effects that are made possible by contrapuntal writing (Schenker, 1987). Hearing Schenker's analysis of Mozart's theme entails *hearing* the d_5 in m. 7 *as* a tone with a transitory, "passing" quality that is related to a preceding e_5 (occurring several bars earlier) and a following $c\sharp_5$ (which, by the time the d_5 is heard, has not even occurred yet). It also entails hearing the descending $d_5 - c\sharp_5 - b_4 - a_4$ motion across mm. 7-8 *as* belonging to a different voice than the ascending $a_4 - b_4 - c\sharp_5$ motion in m. 7. In this sense, as discussed in Section 1.2.3, the diminuendo and the *sforzato* in Kvapil's performance (Figure 1.6a) may be seen as correlates of this hearing, insofar as they can influence the likelihood that we may attach these particular additional features – as opposed to those implied by Morgan's analysis – to the events comprising the musical surface. Yet, hearing the diminuendo itself, as a feature of the auditory signal, is not *per se* constitutive of structural hearing – it is part of the *explanandum* of an interpretation, not of the explanation. Also note that, in principle, listeners may learn to recognise and respond to structural cues simply based on plain-hearing their surface identity, without ever recognising their structural identity. This observation highlights the core difficulty in investigating structural hearing empirically: since the presence of specific audible structural cues correlates with a certain *hearing as*, it is difficult to disentangle the empirically observable effects that are due to the perception of the cues themselves from those that may be due to the attribution of structural features to surface events, i.e., to structural hearing proper.

In summary, from an analysis, we do not only expect *internal* explanatory power towards the musical surface, i.e., the capacity to explain the appearance of the musical surface. We also expect the analysis to be *externally* explanatory towards some possible experience we may have of the musical surface. More importantly, it appears to be the case that theorists find it useful to appeal to internally explanatory interpretations to explain their experience: that is, the claim is made that the listener "feels" in a certain way "because" the musical surface affords a certain (internally-explanatory) interpretation, i.e., because it can be seen as being arranged intentionally according to a particular narrative (Temperley, 2001b). Interpretations, then, are not only theories whose *explanandum* is the musical surface, but they are at the same time theories whose *explanandum* is (some aspect of) the listener's perceptual and cognitive experience.

Having characterised structural hearing in terms of the relationship between musical surface, musical structure, and listening experience as they are discussed in the music-theoretical discourse, we move now to the core issue we intend to address: namely, investigating whether and how this notion of structural hearing plays a role in human cognition. In doing this, we face a methodological and epistemological challenge as we shift concepts that are developed within the music-theoretical discourse into the domain of cognitive science. The problem is that there is no natural translations of concepts that are defined as (generalisable) internal explanations or as a "folk psychology" based on introspection (Cross, 1998) into empirically well-defined phenomena. As a consequence, it is unclear (1) whether the empirical observations we make in experiments are really *about* those music-theoretical concepts, and (2) how can we make empirical observations that *are* indeed about those music-theoretical concepts. In the following, we seek to integrate the music-theoretical notion of structural hearing into a cognitive-scientific framework, with the ultimate goal of making it amenable to empirical investigation.

As a starting point, we propose to understand structural hearing as an instance of Bayesian inference in the context of Bayesian cognition (Griffiths et al., 2008). In the following section, we introduce the Bayesian perspective as a general framework for understanding structural hearing as a cognitive phenomenon. The empirical challenge is then to investigate how this inference may be implemented in terms of cognitive processes. In Chapter 3, we will further characterise the inference process as a form of incremental parsing in the context of grammar-based generative modelling, and we will discuss the empirical implications of such a modelling assumption. Based on this perspective, the work of the thesis will focus on examining some specific aspects of (1) the representations putatively produced as an output of the inference (Part II), and (2) the cognitive mechanisms that may implement the parsing process during listening (Part III).

1.3 The structural listener

1.3.1 Generative modelling in Bayesian cognition

A central issue in general cognition is how brains infer what unobserved configurations of entities in the environment are the cause of observed sensory signals, and how such inferences manifest themselves in the form of percepts (von Helmholtz, 1867). First, the observer cannot know veridically what the space of environmental states that *really* cause the sensory signals is: only the signals themselves are known. Furthermore, the sensory signals are typically underspecified and sparse, so that even successfully inferring the causes of all the *actually* observed signals would not trivially generalise to new signals that have not been observed before. In vision, for example, the outcomes of perception are representations of objects in the environment. Since the objects themselves are not accessible to observation, other than through the visual signals they physically cause, these representations should be understood as the result of an inference (cf. Knill and Richards, 1996). Note that this inference is not deterministic, since it has to rely on underspecified data due to the retinal image being typically ambiguous (e.g., different objects in different orientations may produce the same retinal image; Kersten et al., 2004). Furthermore, in many cases, the brain's "best guess" as to which the external states may be ends up not reflecting the "real" entities at all, resulting in illusory phenomena (Weiss et al., 2002). Given the ambiguity of the external signals, how is this inference even possible in the first place?

Since the inference is not deterministic, it should be dealt with in probabilistic terms. The external environment is defined in terms of a space of possible unobserved external states $\hat{S}$. Generative processes in the environment give rise to a space of possible observed signals O, such that the state $\hat{s} \in \hat{S}$ causes signal $o \in O$ with probability $p(o|\hat{s})$. In order to know what state the environment is in, the observer would need to compute the probability distribution $p(\hat{s}|o)$: this *posterior* distribution of the signals' causes models the observer's beliefs about the plausibility of such unobserved states *after* observing sensory data. Under Bayes theorem, this probability is proportional to the product of the *likelihood* $p(o|\hat{s})$ of the observations, times the *prior* probability distribution over latent states $p(\hat{s})$, which models the observer's beliefs about the plausibility of such states irrespectively of any data (i.e., *before* observing data):

$$p(\hat{s}|o) \propto p(o|\hat{s})\,p(\hat{s}) \tag{1.1}$$

Exploiting this mathematical relation between the posterior (the desired outcome of the inference) and the likelihood, Bayesian-cognition models reverse the inference problem on its head. While the space of *real* latent states $\hat{S}$ is unknown, the observer is free to (1) "invent" a set of inner *representational* states $S \ni s$, (2) specify prior beliefs $p(s)$ about how plausible each such state is independently of new observed data, and (3) simulate how representational states cause observed signals by specifying the likelihood $p(o|s)$. A model that formalises these three mathematical objects is termed a generative model in the sense that it simulates generative processes that may cause signals in the environment. Having internalised such a

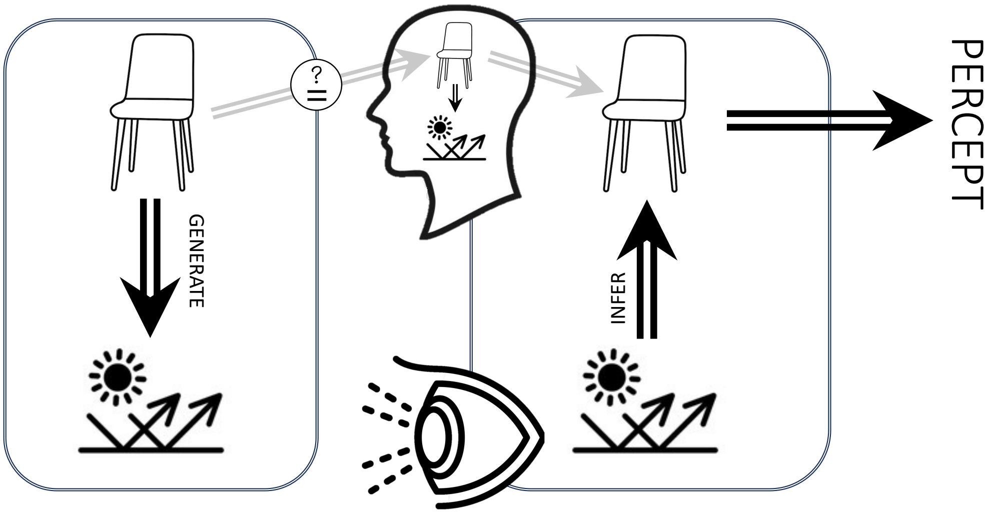


Figure 1.7 – A schematic visualisation of Bayesian cognition. Objects in the external environment (left) produce sensory signals as a result of generative processes (e.g., reflection and refraction of light). Organisms acquire inner generative models of such generative processes. Perception, consisting of mental representations of objects in the external environment, reflects the outcome of Bayesian inference based on the acquired generative models (right).

model, when a specific signal o is actually observed inference can be performed as per Eq. 1.1 by replacing the real states with the representational ones.[11] The outcome of the inference is what is reflected in perception (Figure 1.7). The problem of inference is then turned into one of optimisation: what is the optimal set S of representational states that should be chosen to specify the inner generative model?

In the ideal case, inner generative models simulate exactly the real-world generative processes, and the inner states in S correspond to, or *represent*, the very states $\hat{S}$ of the external world that implement the real generative processes. However, internal generative models need not match external generative processes identically – and, in fact, they never really do (Pezzulo et al., 2021; R. Smith et al., 2022). The role of inner representational states is to provide "good enough" explanations of the causes of sensory data, rather than to correctly encode states of the external environment. How "good" is "good enough"? The optimisation problem can be seen as a form of unsupervised learning (Hinton, 2007) subject to evolutionary constraints. From an evolutionary perspective, organisms benefit from inferences that positively impact performance in tasks that are relevant to their fitness, for example by favouring effective sensorimotor coupling (Baltieri and Buckley, 2019; Friston, 2009, 2010), prediction (Clark, 2013b; Pezzulo et al., 2021), parsimonious encoding (Benjamin et al., 2023; Sablé-Meyer et al., 2022), or abstraction and generalisation (Tenenbaum and Griffiths, 2001; Tenenbaum et al., 2011). The representational states, some of which give rise to phenomenal experience in the form of percepts, may or may not correspond trivially to "real" states in the external world,

[11] Since exact Bayesian inference is typically intractable, approximate methods are assumed to be implemented in practice; Abbott et al., 2013; Shi and Griffiths, 2009; R. Smith et al., 2022).

as long as they are useful towards these goals. From this perspective, illusory percepts are no more illusory than any other percept.

As a concrete example of how such optimisation problem may be implemented, let us consider the case of predictive coding (Clark, 2013b). In this framework, evolutionary pressure is hypothesised to impinge on the capacity of brains to formulate predictions about sensory signals (Pezzulo et al., 2021). Specifically, predictions are formulated by computing the probability

$$p(o|o_0,\ldots,o_t) \tag{1.2}$$

of future observations based on past observations, and observers aim at minimising the discrepancy between these predictions and the actually observed signals. Crucially, in a Bayesian framework, the computational tool that allows predictions is the internal generative model: based on past observations $o_0,\ldots,o_t$ observers infer the posterior distribution $p(e|o_0,\ldots,o_t)$ over plausible current representational states, while the generative model specifies the likelihood $p(o|e)$ that signal o is caused by state e, so that Eq. 1.2 can be computed as

$$p(o|o_0,\ldots,o_t)| = \sum_E p(o|e)p(e|o_0,\ldots,o_t). \tag{1.3}$$

Since, in this view, predictions are mediated by the generative model and its representational states, prediction error can be improved by updating the generative model itself.[12] Overall, the observers' beliefs about unobservable states that may cause observable sensory signals, encoded in the form of internal representational states, are those that provide a *useful* explanation of the signals themselves, in the sense that they, e.g., afford effective predictions. Such optimal generative models are partially hard-coded genetically through evolution, and partially learnable in a flexible way through exposure to new sensory signals (Clark, 2013b; Friston et al., 2016; Pezzulo et al., 2021).

1.3.2 The role of representations

What is the advantage of computing predictions based on representational states, after Eq. 1.3, as opposed to computing predictions directly from the previous observations, after Eq. 1.2? First, since sensory signals are ambiguous and underspecified, it is likely that an alternative representation may be more meaningful than the raw signal itself for making useful and correct predictions. For example, for the purpose of predicting the trajectory of a golf ball, a *symbolic* representation of the ball as a unitary entity together with a sequence of spatial coordinates is plausibly more useful than the sequence of arrays of coloured pixels comprising the visual signal: from the latter, it is even difficult to determine whether the cluster of dark pixels impressed on the retina is a single entity or a cluster of individual (and potentially dynamically independent) entities. Furthermore, representations that optimise predictions are often parsimonious in terms of encoding compression, which may also be a desirable

[12] Observers can also influence prediction error by *acting* on the external world directly, in order to manipulate the sampling of sensory signals (REF Friston, 2010; Friston et al., 2012).

goal for a biological being: in this example, encoding the ball as a categorical entity in a space of possible categorical entities is likely more parsimonious than an encoding of the array of colored pixels as part of a high dimensional continuous space. While these arguments are not sufficient conditions to establish the cognitive relevance of specific representational states (which should be investigated on a domain-specific basis), they do outline the usefulness *in principle* of adopting representational states as mediators of cognitive capacities such as prediction.

Whether representations are relevant in cognition and, if so, what kinds of representations are, has been historically debated in the cognitive sciences. Representational states in connectionist models are configurations of activation of units, or of connection weights between units (Rogers and McClelland, 2008). Connectionist models allow for *distributed* representations emerging by the co-activation of multiple units, each encoding an *atomic* representation (e.g., the state representing the semantics of the word "Odyssey" may comprise the co-activation of four units representing the atomic concepts "book", "journey", "Ulysses", and "Homer"; cf. J. L. McClelland and Cleeremans, 2009). However, in connectionist models, only one type of relation is conceived between representational states: namely, the causal "transfer" of activations and weights from one state to another. Accordingly, atomic representations can only be combined to form distributed representations by the one relation that is admitted, namely, their being co-activated (J. A. Fodor and Pylyshyn, 1988). Furthermore, representational states in connectionist models are not "stored" as memories, as long-term knowledge is only encoded in terms of residual connection weights (J. L. McClelland and Cleeremans, 2009). In other words, information stored in connectionist representational states cannot be "retrieved" or "manipulated", but only reconstructed.

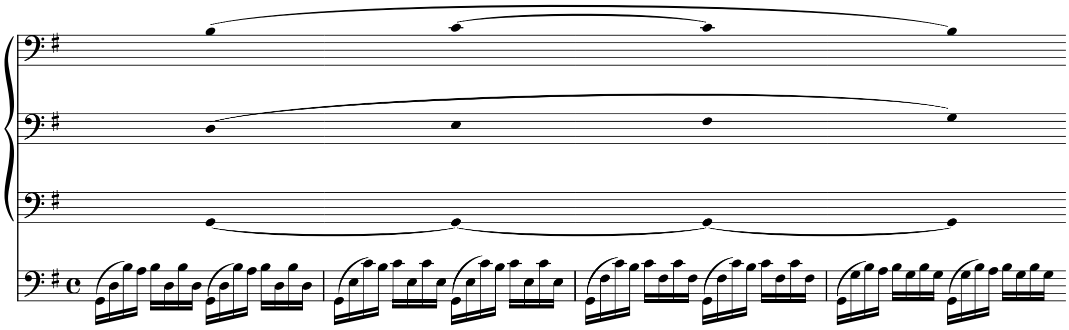
In classical symbolic models, instead, representational states have inner structure, in the sense that they result from the composition of "more elementary" (ultimately, atomic) representational states that are combined in terms of a variety of possible relations (J. A. Fodor and Pylyshyn, 1988). The representing system's inner states are representational in the sense that they *resemble* the system that is being represented (e.g., in the form of a homomorphism or isomprphism; O'Brien and Opie, 2004; Shea, 2014). Crucially, the behaviour of classical symbolic systems is defined in terms of operations performed on the structural features of its representational states. In other words, the information encoded in the representational states is required to be *causally relevant* (Gładziejewski and Miłkowski, 2017) and *exploitable* (Shea, 2014): state transitions of the representing system are causally influenced by the information *about* the represented system that is encoded in the representations. The empirical burden of proof, for a classical representationalist account, is to show that aspects of behaviour are only possible as a result of *exploiting* such information.

Connectionist and classical symbolic models seem naturally apt to capture different aspects of cognition: the former reflect a plausible computational architecture of the brain (in terms of interconnected units or neurons), whereas the latter faithfully capture the seemingly symbolic nature of thought as it appears to introspection. Recent developments in connectionist models of cognition have challenged the cognitive relevance of structured representations as assumed in classical symbolic models of cognition (J. L. McClelland et al., 2010). For example, the linguistic performance of Large Language Models (LLM), which are based on connectionist architectures, seems to render symbolic representations superfluous and at best epiphenomenal to cognitive processing even in a traditionally representationalist domain such as language (Piantadosi, 2023). However, it is possible for such models to learn to explicitly encode structured representations, and to exploit their explanatory power. For example, Manning et al. (2020) recently showed that distances between words in the embedding layer of a Large Language Model (LLM) is related to geodesic distances between words in classical syntactic trees, thus effectively encoding syntactic structure. In this case, it would seem that a symbolic model is not alternative to a connectionist one, and may rather be seen as modelling the convergence point of connectionist learning.

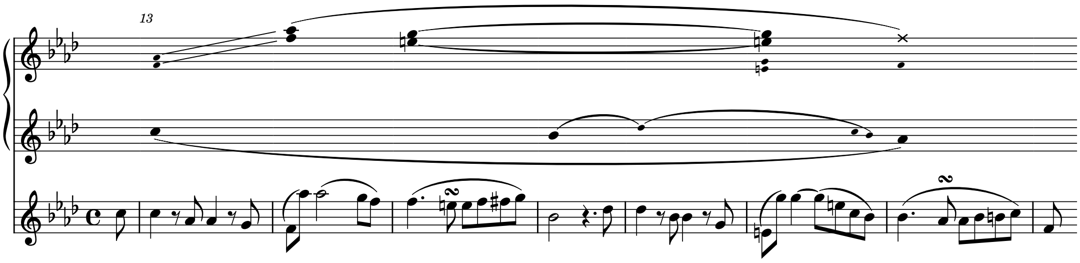
Overall, it is unclear whether and to what extent the two approaches are mutually exclusive: cognition may well be captured by a symbolic model at Marr's 1982 computational level while resembling a connectionist system at the implementational level (Griffiths et al., 2010). Addressing this issue in greater detail exceeds the scope of the present work, but the upcoming presentation should be seen as rooted in the classical symbolist perspective. While this epistemological choice is, ultimately, arbitrary, we commit to this view for one main reason. Namely, that the phenomenon we intend to capture is the introspection of the listener, and this introspection (based on the verbal accounts we have access to) has little to do with connectionist activation patterns, and everything to do with structured representations of symbolic entities and relations between them. As a result, we are not directly concerned with establishing whether the computational architecture at the implementational level is classical or connectionist. Nevertheless, any neurologically plausible, possibly connectionist model of (music) cognition will need to account for, hence be constrained by, the phenomena we identify in the process.

1.3.3 Structural interpretations as generative models

If Bayesian brains aim at inferring explanations of the observed sensory signals in terms of their latent causes, what generative models would they learn in order to make sense of musical signals? Minimally, signals in the auditory domain are proximally caused by their physical sources: the acoustic properties of the musical signal can be explained by inferring the location, identity, and physical properties of its sources as part of auditory scene analysis (Bregman, 1994). From a Bayesian perspective, inferring the unobservable causes that explain the observed signals would then require a generative model that calls upon representations of (physical or virtual) auditory sources to simulate the production of sound (Cusimano et al., 2018). Auditory streaming, in particular, refers to the attribution of different sound events to different physical or virtual sources, resulting in the perception of several distinct streams as opposed to a single one (B. C. J. Moore and Gockel, 2012).



(a) J.S. Bach, Suite for unaccompanied cello n. 1, *Prelude*, mm. 1ff



(b) F. Ries, Piano Quartet op. 13, i, mm. 13ff, right-hand part

Figure 1.8 – **(a)** Implied polyphony in strict counterpoint. The three implied voices (top three staves) are clearly segregated in different registers and different strings, each with its own timbral quality. **(b)** Implied polyphony in a free-polyphonic texture. In the top staff, the same two (proto-)voices appear shifted across different registers, making it difficult to identify them as continuous streams based on auditory-scene-analysis principles alone.

Crucially, the perception of distinct streams often does not correspond to distinct sources actually being present in the external world: for instance, distinct streams can be perceived in music produced by a single instrument (Deutsch, 1999). This (illusory) perceptual phenomenon allows musicians to introduce polyphony in apparently monophonic contexts (Figure 1.8; Davis, 2006, 2011). Indeed, in the most trivial cases, the basic principles of auditory streaming can account for the perception of distinct voices in music *as if* they were produced by distinct sources: for instance, when transitions across voices are marked by clear registral or timbral differences (Figure 1.8a). However, tonal music exploits subtler forms of implied polyphony in the context of so-called "free" polyphony, and the phenomena pertaining to free polyphony cannot be reduced to auditory streaming (Finkensiep, 2023; Schenker, 1935). In particular, the notion of (proto-)voice in free polyphony is independent of the attribution of tones to physical or virtual sources (Figure 1.8b): while the intuitive notion associated with the term "voice" is an auditory stream produced by a single source, what holds individual proto-voices together is not (necessarily) being part of a perceptual stream, but rather the existence of structural relations between tones (Finkensiep, 2023). In Schenker's analysis of Mozart's theme (Figure 1.4b), the ascending $a_4 - b_4 - c\sharp_5$ motion in m. 7 belongs to a different voice than the descending $d_5 - c\sharp_5 - b_4 - a_4$ motion across mm. 7-8. If a listener were to *hear* these tones *as* per Schenker's analysis, an auditory-streaming account would not only have to explain based on what auditory criteria the descending motion is segregated from the ascending one (breaking the quite intuitive gestural *Gestalt* formed by the arch-like contour across mm. 7-8), but also how the d_5 connects back to the e_5 in m. 1. Furthermore, even identifying the two proto-voices as distinct streams (which falls within the scope of streaming) would not ascribe structural relations to the tones: for instance, identifying $e_5 - d_5 - c\sharp_5 - b_4 - a_4$ as a stream does not imply that d^5 relates to e_5 and $c\sharp_5$ as a *passing* event connecting the two.[13] Crucially, though, what *explains* the presence of a d_5 *specifically* in the surface is exactly its being part of a passing connection between e_5 and a_4, not merely its being part of a certain auditory stream.

In summary, the external *causes* that determine the appearance of the musical surface do not only comprise physical sound sources, but also the idiom-specific structuring principles that reflect the intentionality underlying composition. From this perspective, the structural relations that cause the musical surface to look the way it does do not need to conform to auditory streaming principles, and generative models of musical surface based on latent physical sources as proximal causes can only partially explain the appearance of the musical surface. Here we are, back to the question that opened this section: if Bayesian brains aim at inferring explanations of the observed sensory signals, what generative models would they learn in order to make sense of musical signals? Music theory provides us with a manifold of possible answers to this question. As discussed in Section 1.2.3, music-theoretical interpretations are rational explanatory models of the musical surface – they are internally explanatory. Exactly because music theoretical interpretations are internally explanatory, brains may converge to such interpretations as illusory yet useful mediators of cognitive capacities including memory and prediction.

Specifically, the (internally) explanatory relation between musical structure and musical surface can be formalised by thinking of the musical surface as the result of a *generative process*. Metaphorically speaking, this entails to explicitly model a virtual "environment" populated by latent entities that may influence the presence of observable surface events, and by specifying the transformations that map the entities and relations comprising this unobservbale "environment" into observable surface events. A narrative or formal description of such a generative process can be understood as an (internal) explanation of the musical surface, of the kind we find in music analytical discourse. Note that the generative process does not need to correspond to anything that happens in reality (e.g., the composer's creative process as reflected in their sketches that progressively build up to the entire composition). A specific instance of the process, which generates a specific surface, should be understood as a way to represent the structural relations that justify the appearance of the musical surface. In turn, the generative model as a whole captures the pairing of surfaces and interpretations, providing a characterisation of this mapping at the computational level.

To give a concrete example, the analytical claim expressed by labelling a segment of the musical

[13] The structural status of a $\hat{3}$ supported by a cadential 64 suspension is debated in the Schenkerian literature (cf. Beach, 1990; Cadwallader, 1992), but this debate does not qualitatively impact the argument made here.

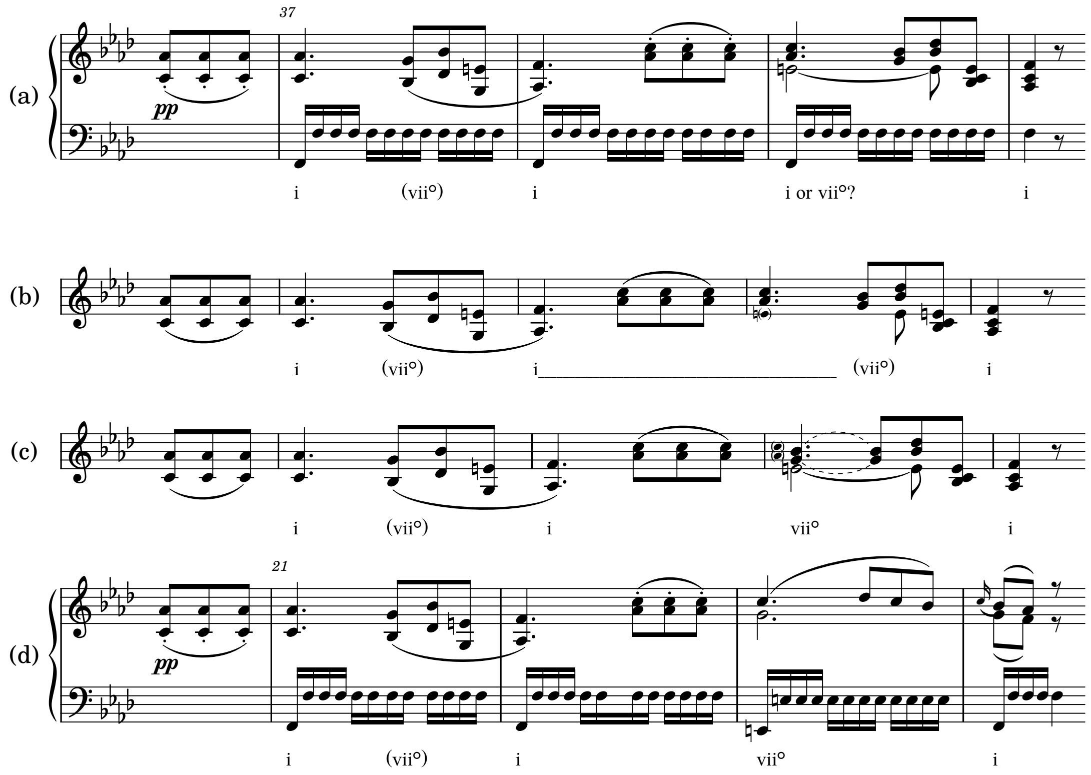


Figure 1.9 – (a) W.A. Mozart, *Sonata* for piano KV 330, ii, mm. 37ff. (b) A reduction where the dissonant sonority on the downbeat of m. 39 is interpreted as a dominant chord ($vii°$) over a tonic pedal, with two non-chord tones ($a\flat_4$ and c_5) as a double *appoggiatura*. (c) A reduction where the same sonority is interpreted as an F minor chord with e_4 added as an extraneous non-chord tone. In the former interpretation, the only chord-tone that expresses the latent dominant harmony is e_4, all other tones being extraneous non-chord tones, whereas in the latter interpretation, e_4 is the only extraneous tone. (d) A parallel passage occurring in mm. 21ff: here, the harmony on the downbeat of the third measure is unambiguously a dominant harmony.

surface with a certain chord c (e.g., the annotations in Figure 1.9a) can be modelled by thinking of observed surface events as the observable manifestation of a latent entity, the chord: from the perspective of the listener/observer, a chord is not unlike a (physically concrete, yet still latent) physical object in the environment that produces a visual sensory signal by reflecting light. In such a model (as proposed by Finkensiep, 2023, Chapter 4; cf. also Harasim et al., 2021 for an analogous approach to mode inference), the generative process is formalised by specifying the (prior) probability $p(c;h)$ that a chord c may be used by a composer at all (e.g, as influenced by the historical context h), as well as, for each such latent chord-entity c, the likelihood $p(s|c)$ that a given surface s is produced by that chord-entity: surface events can then be "generated" (alongside the latent entities that "cause" them) by sampling from the joint probability distribution $p(s,c;h) = p(s|c)p(c;h)$. This formalises the (internal) explanatory value of the statement that the surface *is* a certain chord $\hat{c} = \arg\max_c p(c|s;h)$, with $p(c|s) \propto p(s|c)p(c;h)$ after Eq. 1.1: the inferred chord $\hat{c}$ is the latent entity that offers the best explanation for why the surface looks the way it does.

Let us consider from this perspective the passage displayed in Figure 1.9a, and in particular the sonority occurring on the downbeat of m. 39. Possibly the most intuitive hearing is one where the f in the bass together with the $a\flat_4$ and c_5 constitute a F minor triad, relative to which e_4 is dissonant. Under this analysis, the e_4 is *heard as* an elaboration of the reduction displayed in Figure 1.9c. However, another plausible reduction of this passage is the one displayed in Figure 1.9b. Under this analysis, the sonority on the downbeat is assumed to be a *dominant* sonority (e° in the key of F minor), whose only expression is the e_4 – all other tones being non harmonic tones. These two interpretations proposed for m. 39 in Figure 1.9 are both captured by a generative model where the musical surface is generated as the expression of latent chord-entities. The greater plausibility of the intuitive interpretation where the pitch collection $\{f_2, e_4, a\flat_4, c_5\}$ expresses an F minor chord may be formalised in terms of the likelihoods $p\left(\{f_2, e_4, a\flat_4, c_5\}|\text{chord} = f\right) > p\left(\{f_2, e_4, a\flat_4, c_5\}|\text{chord} = e^\circ\right)$ that chord f or e° express the observed pitches. However, the actual posterior probability

$$p\left(\text{chord}|\{f_2, e_4, a\flat_4, c_5\}\right) \propto p\left(\{f_2, e_4, a\flat_4, c_5\}|\text{chord}\right) \cdot p\left(\text{chord}; \text{context}\right)$$

of the chord interpretation is also shaped by other contextual cues that make the presence of a e° chord more plausible *a priori*: this additional knowledge is formalised in the prior $p(\text{chord}; \text{context})$. In this case, for example, parallelism with the thematic material presented in mm. 21ff (Figure 1.9d) may bias the listener to assume $p\left(\text{chord} = e^\circ; \text{context}\right) > p\left(\text{chord} = f; \text{context}\right)$. As a result, the interpretation in Figure 1.9b, and not Figure 1.9c, may turn out to be the one that is actually inferred.

Overall, this discussion exemplifies how internally-explanatory interpretations can be formulated as generative models of the musical surface – a more systematic account in terms of generative grammars will be given in Chapter 3. Here, we just intend to point out that generating and explaining can be seen as two sides of the same coin. From this perspective, aspects of the structural identity of individual events are the result of an inference based on the inversion of a generative model. When a surface is presented to the listener, the outcome of the inference is a representation of the virtual, unobserved "environment" of entities and relations that generates that sensory input. Such representations, in turn, are what is reflected in perception.

1.3.4 Structural hearing as inference

To some extent, it is trivially true that "brains" infer internally-explanatory interpretations: minimally, they do so as part of the rational endeavour of music theorists. However, the rational process of explaining, as reflected in the discipline of analysis, may in principle be completely unrelated to "hearing" as a perceptual and cognitive experience, which is our focus here. Analytical understanding is the result of volitional rational thought, and can be supported by explicit domain knowledge as well as by outsourcing representations and computations to external memory and processing infrastructure, including bodily representations (e.g., fingering or articulation decisions in piano performance; cf. Schenker and Esser, 2000), mathematical models (e.g., the geometric properties of a particular space; Tymoczko, 2011), algorithms (as, e.g., in computer-assisted composition; Dean, 2011). In these cases, analytical understanding can in principle even exceed the cognitive, perceptual, and representational capacities of the individual.

Nevertheless, as discussed in Section 1.2.4, when theorists talk about analytical interpretations, they entail something beyond the level of rational understanding. In *Composition and Cognition* (Lerdahl, 2019), theorist and composer Fred Lerdahl attributes certain structural features to his own compositions, proposing that listeners *hear* music *as* bearing those structural features. Dmitri Tymoczko, approaching Lerdahl's music as an analyst, comments:

> [Lerdahl's] diagrams do not correspond to anything I recognize from my own experience. [...] *Composition and Cognition* asserts [that these diagrams] show how "listeners" do hear his piece, while GTTM and TPS both offer a theory of the perceptions of "experienced" listeners, of which I am one. So if he is right, then either I am wrong about my own musical experience (not just of Lerdahl, but also of Bach, Beethoven, etc.) or my hearing is defective to the point where I should be considering another line of work. This is perhaps why I tend to experience Lerdahl's writing as [...] contradicting my beliefs about my own experience. (Tymoczko, 2020, par. 29ff)

This dispute strongly indicates an underlying belief that the association between structure and surface has a psychological reality that is not fully captured by the rational understanding of an analysis. In fact, on one hand, there is probably a shared rational understanding that Lerdahl's analyses are internally explanatory towards his own music (so that, in principle, anyone could draw a graph explaining the surface of Lerdahl's music as Lerdahl intended). On the other hand, though, there seems to be a shared belief that the musical surface (1) has the affordance to elicit *some* experiential manifestations that emerge to some extent automatically and implicitly, as opposed to fully explicitly and under volitional control in the same way as rational thought does; and (2) cannot elicit (or is unlikely to elicit) *some other* experiential manifestation, not even under volitional effort. The disagreement is exactly about what these (non-)affordances of the musical surface are: which music theory offers a better explanation of the fact that a given passage of music has *some* experiential manifestations that emerge spontaneously, and some that cannot emerge at all? To be sure, different listeners may experience the same music differently, but an effective music theory (for a given musical idiom) should be able to account for all possible interpretations that some listener can "recognise from their own experience" (Tymoczko, 2020): in our terminology, music theory aims at offering internally-explanatory interpretations that are also externally explanatory towards aspects of some listeners' experience that emerge *automatically and implicitly*.

Based on the Bayesian-cognition perspective, we propose that the reason why internal and external explanatory power go hand in hand is that, as part of their ordinary effort of *making sense* of the signals they receive from external world, listeners

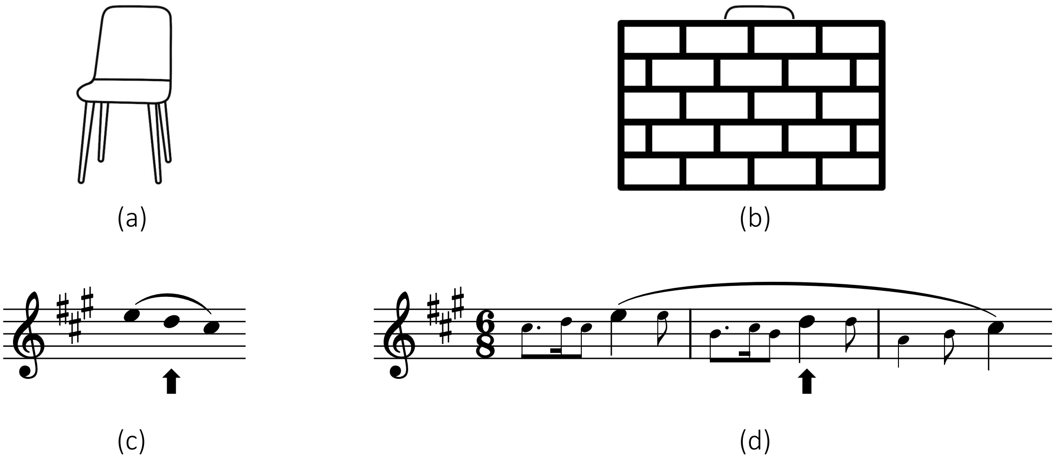


Figure 1.10 – (a) A chair. (b) The same chair, hidden behind a wall. In both figures (a) and (b), we may infer that the presence of a chair is a likely cause of the visual stimuli, hence *see* (part of) the figure *as* (part of) a chair. (c) A passing tone (arrow). (d) The same passing tone (arrow) "hidden" in an elaborated surface. In both examples (c) and (d), we may infer that the a likely (internal) explanation for the presence of d_5 is its being a connection between e_5 and $c\sharp_5$, hence *hear* d_5 *as* a passing tone.

(I) LEARNING: acquire a generative model for musical surfaces, based on which they

(II) REPRESENTATION: infer (representations of) internally-explanatory interpretations, and

(III) PROCESSING: this inference happens in real time, automatically and implicitly, to some extent at least.

More precisely, as listeners are exposed to a repertoire of surfaces, they face the problem of making sense of all these observations by inferring their latent causes. As part of this endeavour, brains acquire a generative model for the entire repertoire of a style, including its (potentially infinitely many) unobserved instances. What the acquired generative model looks like, given exposure to a repertoire, is determined by (1) how the generative model and the inner representational states it implies contribute to cognitive functions including prediction, memory, generalisation, action, or reward with respect to the stimuli that are actually encountered by the listener, and by (2) constraints of the neural and cognitive computational infrastructure, which may limit what models are learnable and what inference computations are feasible in the first place.

Such generative models provide generalisable internally explanatory interpretations to all surfaces in the repertoire. In other words, referring to the definition of interpretation given in Section 1.2.3, the representations resulting from the inference process are encodings of a labelled graph of entities and relations. We terms such a representation a *structural representation*. We then think of the representational states implied by the generative model as contributing to shape perceptual experience, behaviour, and brain activity. In vision, we "see" a certain arrangement of coloured pixels "as" (part of) a chair (i.e., endowed with the *quale* of "chairness") based on (largely) automatic inference mechanisms relying on a generative model

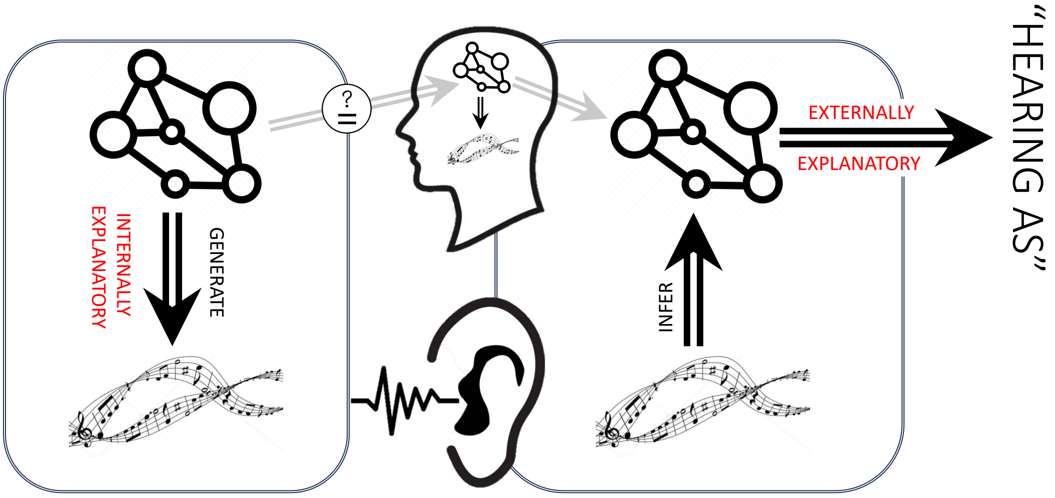


Figure 1.11 – The relationship of surface, structure, and perception (cf. Figure 1.2) revisited in the framework of Bayesian cognition (cf. Figure 1.7). Music-theoretical structural interpretations are internally explanatory towards the musical surface and, as such, they are candidate generative models that listeners may infer as part of the cognitive processing of sensory inputs. To the extent that aspects of the listener's perceptual experience reflect the inferred representations, structural interpretations are then both internally and externally explanatory.

that "explains" visual signals based on objects as latent causes (Figure 1.10a,b). Similarly, we may "hear" a given surface event "as" an entity that is part of a certain interpretation (e.g., as having the *quale* of "passingness", or of "dominantness", or of "syncopatedness") if our brain, having acquired a generative model as a cognitively useful (e.g., parsimonious and generalisable) tool to assign causes to stimuli, "explains" the presence of the given event by attributing to it structural relations with other events (Figure 1.10c,d). The last component of structural hearing, together with (I-III) above, is then that

(IV) EXPERIENCE: the experience of "hearing as" reflects the information encoded in the inferred representations, as well as possibly the processing mechanisms that implement the inference.

As a consequence, a music theory that appeals to the same entities and relations as such a cognitively acquired generative model would be guaranteed to be both internally and externally explanatory (Figure 1.11).

We suggest, then, that music theory fulfils its twofold function by both reflecting and influencing the generative model that "brains" acquire and exploit as part of their normal effort of making sense of the musical surfaces as auditory inputs. On one hand, music theorists discover the features of such inner generative models just like vision scientists discover object-models of vision (Marr, 1982) and linguists discover language-models of speech (Chomsky, 1965). In this sense, music-theoretical models *reflect* the listener's inner generative models. On the other hand, music theory also has the potential to *influence* the acquisition of internally-explanatory generative models in its capacity of also being prescriptive towards the listener and the composer. Specifically, music theory can be prescriptive towards the listener in the sense that it more or less explicitly encourages listening strategies that may facilitate detecting specific structural cues, hence the inference of generalisable internal explanations that are constrained by those cues (Polth, 2006; Salzer, 1962). Music theory may also be formulated in a way that is prescriptive towards composers (Wason, 2002), thereby directly influencing the repertoire based on which inner generative models are acquired through implicit learning (Rohrmeier, 2010; Rohrmeier et al., 2012). The adherence between music-theoretical interpretations and structural hearing is then reinforced by this feedback loop between listening, theorising, and composing.

Such a framing of structural hearing allows us to interpret music-theoretical introspection, as faithfully as possible, as a cognitive phenomenon. However, at this stage, there is nothing *factual* in this perspective. The notion of structural hearing outlined above is only the definition of a putative phenomenon, while its existence is to be established empirically. In particular, at this point, the characterisation is too underspecified to make predictions towards empirically observable effects of human cognition and behaviour. Questions to be answered include (1) what are the characteristics of the hypothesised representations, (2) how is the inference performed, and (3) how is it constrained by the cognitive architecture and the temporality of listening. In Chapter 3, we will discuss how these aspects can be addressed in terms of grammar-based parsing, before testing some implications of this approach in the body of the thesis. As we make the theoretical model more concrete and operationalise it for empirical testing, one challenge lies in making sure that the reductionist assumptions introduced in the process remain faithfully compatible with the phenomenology of structural hearing. In the remainder of this Chapter, we discuss this methodological challenge with some concrete examples, justifying the top-down perspective we commit to in the modelling and empirical approach.

1.4 The divergence of the empirical and the music-theoretical view

1.4.1 Bottom-up vs. top-down

Empirical research in music psychology typically proceeds by unpacking the music-theoretical discourse into elementary phenomena that are operationally well-defined in the psychological discourse, like the psychoacoustic consonance of simultaneous tones (Parncutt and Hair, 2012; Terhardt, 1984), prediction (Pearce, 2018; Rohrmeier and Koelsch, 2012; Vuust et al., 2022), boundary perception (Sloboda and Gregory, 1980), the hierarchy of tonal stability (Krumhansl and Kessler, 1982), or auditory streaming (Bregman, 1994). Individually, each of these phenomena (under their empirical definitions) can be investigated to establish what cognitive mechanisms underlie its emergence. The challenge of investigating structural hearing as it is envisaged in the music-theoretical discourse is then to integrate all these individual perceptual and cognitive phenomena into a model of how correlates of the music-theoretically relevant structures emerge from elementary perceptual and cognitive phenomena. Ultimately, this *bottom-up* approach is crucial to achieve a mechanistic understanding of the workings of the mind at Marr's (1982) algorithmic and implementational level.

A complementary strategy is to start with a faithful characterisation of the general phenomenon itself – the emergence of interpretation in perception – as a computational task, at Marr's (1982) computational level of description. In our case, this would be a mapping from the musical surface to its interpretations (and vice-versa: from a Bayesian perspective, generation and inference are two sides of the same coin). By identifying what are the plausible features of structures, as well as how these may be matched to musical surfaces, music theory provides us with insights that can inform the construction of such a model. In turn, the definition of a computational problem constrains what algorithmic strategies and implementational infrastructure may in principle "solve" that computational problem. A *top-down* research approach would then derive – starting from the computational description – hypotheses about what representations may be useful to solve the computational task and how those representations may be manipulated algorithmically to solve the task. When we test these hypotheses, then, we learn something that is directly relevant to the phenomenon at hand: algorithmic and implementational hypotheses, being derived top-down from a computational-level description, are guaranteed to be (components of) well-defined "solutions" of the computational task itself.

For the present purposes, we are in the following situation. We have a phenomenon defined as the computational task of mapping surfaces to structural interpretations, and we intend to support or falsify the existence of such a phenomenon by making empirical observations. The burden of proof we bear is to ensure that the the array of *known* elementary phenomena that we identify in empirical work compose to form a computationally viable solution of the general computational problem itself. For example, if we intend to understand whether and how an unknown piece of electronic hardware computes sums (among other things), we can start by probing the activity and interactions of its electric components. Most of the behaviour we will observe, though, is likely to be unrelated to the specific computational problem of computing sums. Even if we do identify some patterns of electric behaviour that *occasionally* match inputs and outputs *as if* the machine was computing sums, this will not really be sufficient to support the statement that the machine computes sums. To this end, we would need to show that some specific pattern of (electric) behaviour implements a mapping from inputs to outputs that shares the algebraic properties of sums in *all* (observed and unobserved) cases where the machine is expected to be capable of performing sums. Ideally, we would also need to show that the particular pattern of behaviour we identified can in principle be extrapolated to implement such a mapping in *all* cases where sums are mathematically well defined, even when this exceeds the capacity of the particular machine: for example, even if the machine lacks the capacity to store numbers beyond 3 digits, we could specify how the same pattern of behaviour could compute 999 + 1 with a well-defined adaptation (e.g., by adding some extra bits of memory). In summary, we need to show that the observed behaviour implements the computation in a way that generalises to both observed and unobserved cases where the computation is well-defined (possibly including some well-defined adaptation). Without satisfying these constraints, we can still identify some true phenomena of the machine's behaviour, but with no guarantee that they collectively reflect the one particular aspect we are interested in.

In the following examples, Sections 1.4.2-1.4.5, we highlight some putative effects of perceptual experience that are suggested by music-theoretical insights. These examples confront us with two issues. One is the question of whether such differences are cognitively *real.* This, we argue, is an empirical question that should not be settled out of principle – in fact, in this thesis, we will address this kind of question empirically for some specific scenarios. An epistemological and, more pragmatically, methodological issue remains: what approach should we take to model these phenomena, and possibly investigate their existence and emergence? Based on these examples and the preceding discussion, we argue that currently established elementary phenomena as defined empirically stand in no trivial relationship to the general phenomenon of structural hearing as characterised in music theory. As a consequence, it is unclear how a bottom-up approach can lead to claims that are directly relevant to support or falsify different degrees of structural hearing as cognitive phenomena.

1.4.2 The structural identity of sounded events

Mechanistic accounts of rhythm perception suggest that the experience of metrical weight emerges in terms of periodic peaks of hightened predictability (Vuust et al., 2022) and hightened attention (Large and Jones, 1999). Under these accounts, the rhythmic identity of sounded events is characterised by their temporal location relative to an underlying grid of metrical weights. In particular, at a merely descriptive level, rhythmic syncopations occur when an event occupies an off-beat temporal location while nearby on-beat temporal locations carry no events: syncopated events are then characterised as being unexpected and "surprising" (Vuust et al., 2022).

Under this account, the rhythm in Figure 1.12(a) shows some degree of syncopation, as both beats in the second bar are not occupied by sounded events whereas nearby weaker metrical locations are. However, this rhythm affords (at least) two different structural interpretations,[14] as exemplified in panels (b) and (c). We could *hear* this rhythm *as* a syncopated version of the famous ♪♪♪|𝅗𝅥 opening rhythm of Beethoven's Fifth Symphony, where the downbeat is delayed by an eigth-note duration; or we could *hear* this rhythm *as* a syncopated version of the simple template displayed at the bottom of panel (c), where the entire pattern is anticipated by an eigth-note duration.

In order to account for the difference between these two hearings, we need to define the identity of each event not only in terms of its temporal location (relative to an underlying metrical grid), but also in terms of the structural relations events entertain with one another

[14] Other interpretations of the same surface that imply a different alignment with the metrical grid (e.g., ♪|♪♩♩) are also possible.

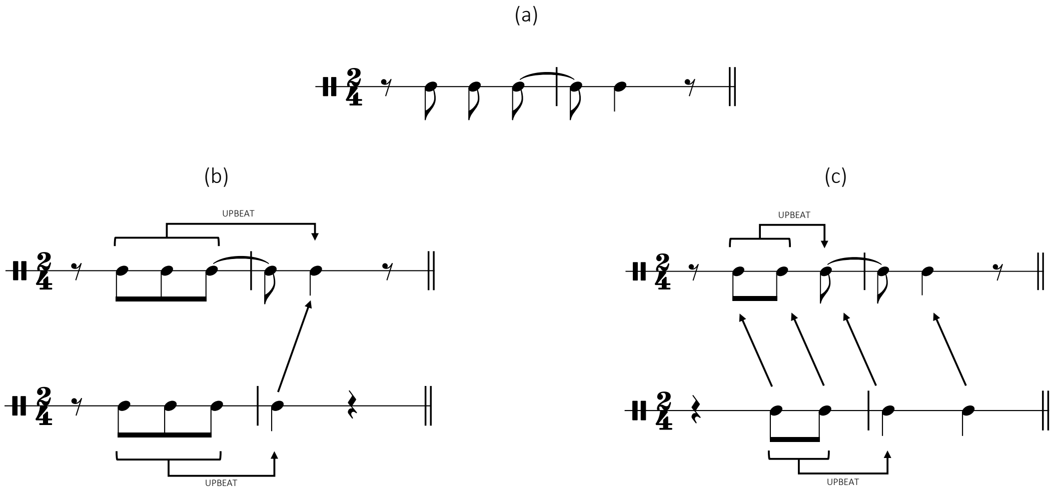


Figure 1.12 – (a) A syncopated rhythm that affords two structural interpretations. (b) The rhythm in (a) is understood as a syncopated version of the rhythm displayed at the bottom: the first three events are upbeats to a delayed downbeat. (c) The rhythm in (a) is understood as a syncopated version of the rhythm displayed at the bottom: the first two events are understood as an upbeat to the third, and the entire pattern is anticipated by an eigth-note duration.

(cf. Rohrmeier, 2020a). For example, the question "which event *is* the first event of the second bar?" can only be answered by recognising that some events have the character of *leading towards* other events as their "upbeats": depending on whether we attribute this "upbeat" character to the first three events (leading towards the fourth), or to the first two events (leading towards the third), we end up giving different answers, corresponding to hearings (b) and (c) respectively. Even the question "which events are syncopated?" affords different answers in the two hearings: in (b), only the last event is syncopated, while in (c) the entire pattern is syncopated. The character of being "syncopated", then, is again not a property of the events' temporal location, relative to the metrical grid, but it reflects the relationship between an event as it is observed in the musical surface and a counterfactual hypothesis about where that event should have occurred: such a counterfactual hypothesis is represented here by the simplified rhythms (reductions) at the bottom of Figure 1.12. Events that are considered *syncopated* are *displaced* instances of events in the reduction, and the identification of a displaced event in the actual surface with a non-displaced event in the reduction reflects the structural relations they entertain with other events: two events from different rhythms (or from a rhythm and one of its possible reductions) have the same structural identity if they entertain the same structural relations with other events in the respective rhythms. In summary, the phenomenon of rhythm perception seems not to be best characterised as a mapping from the rhythmic surface to a representation of temporal locations relative to a metrical grid, but rather to some representation of the relations that link events with one another – i.e., a structural interpretation.

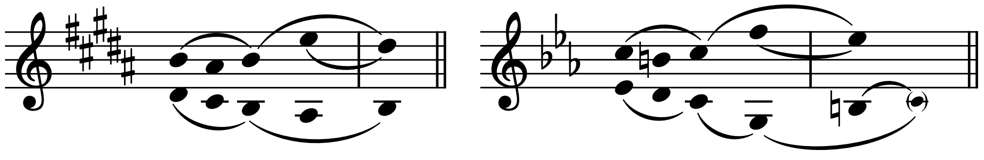


Figure 1.13 – In the two-voiced passage (a), the interval on the downbeat of the second bar is a consonant major third between scale degrees $\hat{1}$ and $\hat{3}$ in the key of B major. In (b), instead, the same two pitches (in a different enharmonic spelling) form the dissonant interval of a diminished fourth between scale degrees $\sharp\hat{7}$ and $\hat{3}$ in C minor.

1.4.3 Latency of structural entities

The notion of consonance, as a psychoacoustic feature of simultaneously sounding pitches enriched by culture-specific connotations (Di Stefano et al., 2022; McDermott et al., 2016; Milne et al., 2023), does not faithfully characterise the music-theoretical notion of consonance in the context of common-practice harmony: for example, in music theory, consonance is not (only) an absolute property of a vertical sonority as a whole, but it is a *relative* attribute of something being *consonant* or *dissonant* relative to something else. Let us consider a few examples.

In Figure 1.13a, the simultaneity on the last downbeat is a *consonant* major third, while it is a *dissonant* diminished fourth in Figure 1.13b, although the sounded sonority is acoustically identical in the two cases. The tonal listener would experience the first interval as consonant and the latter as dissonant – in some sense – irrespectively of the degree of psychoacoustic consonance which is the same in the two cases (if the bichord is taken in isolation). This observation points to a discrepancy in the definition of consonance and dissonance in music theory and music psychology (Parncutt and Hair, 2012). Since both instances of the bichord give rise to the same degreee of psychoacoustic consonance, any perceptual difference between the two musical excerpts is rather to be explained by their different (*structural*) *interpretation.* In Figure 1.13b, the lower note of the interval is an *appoggiatura* of the tonic, meaning that it is interpreted as an extraneous tone that *replaces* an occurrence of the tonic. Note how *extraneousness* (hence *dissonantness*) is a feature that is attributed to a specific instance of the tone based on the relations it entertains with other tones: *hearing* the b_4 *as* extraneous entails encoding information about that particular b_4 that is attributed to it but is not present in the auditory signal. As mentioned before, whether the difference in perception is *real* should not be established out of principle, it is rather an empirical question: what this example shows, though, is that any mechanistic explanation of this phenomenon (if real) should have structural interpretation and not (psychoacoustic) consonance as its *explanandum.* In other words, the perceptual phenomenon to be explained *is* the structural interpretation.

Psychoacoustic consonance might still play a role as part of a mechanistic account of this putative phenomenon: one such account might invoke, e.g., a memory buffer that integrates sensory consonance over time to explain any perceptual difference between these two par-

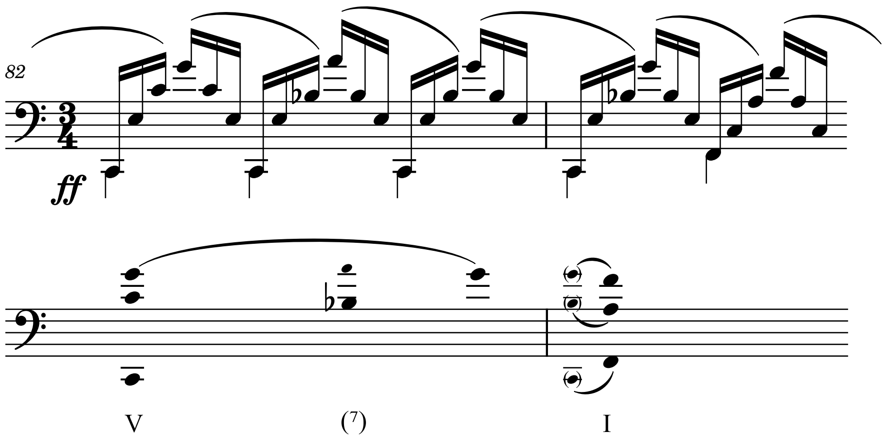


Figure 1.14 – A. Piatti, *Capriccio* for cello op. 25 n. 7, mm. 82f. The chord on the downbeat of m. 83 is a dissonant suspension of an F major chord, where the suspended chord is completely absent from the surface.

ticular musical excerpts at the algorithmic level (after all, pitch b_3 is more dissonant within an overarching *c* minor sonority than in an overarching *B* major sonority). However, this kind of explanation does not easily generalise. Let us consider the example from Figure 1.9a again. Here, the sonority on the downbeat of m. 39 is (psychoacoustically) dissonant as a whole. However, there are two ways to hear this sonority, exemplified in Figures 1.9b,c. The hearing in Figure 1.9c, where the downbeat is taken as the expression of an *f* minor chord, is well explained by the fusion of the tones comprising a root-position triad into a common sonority with a shared implied fundamental pitch (Terhardt, 1974). In the competing hearing, the downbeat chord is taken to be the expression of a dominant harmony instead. In this case, not only are *f*, *a*♭, and *c* all dissonant, but they are so for different reasons. The *a*♭ and *c* are both suspensions of the following *g* and *b*♭, respectively, while the *f* in the bass is a tonic pedal point. Note how integrating consonance information over a longer temporal context does not account for the (putative) experiential difference between these two hearings, since the musical surface is exactly the same in both hearings. Similarly, an account of harmony based on the psychoacoustic notions of implied pitch (Terhardt, 1974; Terhardt et al., 1982) would not capture the plausibility of this second hearing, since *f* – *a*♭ – *c* are likely to be collectively integrated into a chord-like sonority expressing fundamental pitch *f* (Parncutt, 1988), leaving e_4 as the extraneous tone. The difference between the two hearings is rather captured by inferring a partially latent vii^o sonority somewhat counterfactually, irrespectively of the available psychoacoustic cues, and attributing to *a*♭ and *c* the character of being *suspensions of g* and *b*♭, rather than chord-tones over the bass *f*. An even more striking example is displayed in Figure 1.14. On the surface, the sonority on the downbeat of m. 3 looks like a root-position C^7 chord, moving to a *F* chord in the following beat. However, the sonority on the downbeat is rather to be *heard* counterfactually as a suspension of the *F* chord, so that all of the notes on the downbeat are non-chord tones of an entirely latent yet already functionally active *F* major chord. Here, the harmonic entity that is implied on the downbeat is not expressed by any note on the musical surface, and can only be experienced by attributing to the notes that are sounded the character of being "suspensions" of other notes that – counterfactually – should have been there.

In both these examples, the putative existence of alternative hearings is not in contradiction with the established psychological fact that pitches $f - a\flat - c$ or $c - e - g - b\flat$ engender a chord-like percept with a given fundamental pitch. Overall, these examples rather show that a mechanistic account of harmonic perception based on psychoacoustic consonance, pitch inference, and auditory grouping is not a viable implementation of the general computational task of forming structural interpretations: such a mechanistic account does not produce the "right" kind of representation, i.e., the kind of representation that reflects the subtle differentiations of putative phenomenal experience exemplified in Figures 1.13, 1.9, 1.14. Specific psychoacoustic phenomena like consonance or implied pitch may rather be seen as cues that contribute (alongside other cues) to bias the likelihood of one hearing over another.

1.4.4 Categoriality and non-locality of structural relations

In accounts of melodic perception, expectation plays a prominent role both from a music-theoretical and an empirical perspective (Huron, 2006; Margulis, 2007; Meyer, 1956; Narmour, 1990). Expectation can be modelled quantitatively in terms of the probability $p(e|C(t))$ of the occurrence of event e at time t, conditional on some preceding context $C(t)$. A structurally agnostic approach might estimate such probabilities by (loosely speaking) counting the co-occurrences of e and $C(t)$ in some training data: such an approach is structurally agnostic in the sense that it only takes into account features of the musical surface as they are presented. Figure 1.15 shows estimates of such probabilities computed through IDyOM for the last bars of a theme by Schubert (Pearce, 2018). What aspects of the listeners' experience do these probabilities capture?

The sequence of numerical probabilities outlines a contour of higher or lower expectedness or surprisal that unfolds alongside the music, as shown in the plot above mm. 23-24 in Figure 1.15. Listeners may find this contour resonates with their own subjective experience, for example by finding the a^4 following the large descending skip somewhat unexpected, as reflected by a low probability, and the c^5 rather expected, as reflected by a higher probability. However, the experience of (tonal) music as characterised in music theory cannot be faithfully reflected by the continuous modulation of a real-valued quantity, since many aspects of this experience are qualitative and categorical.

For example, the skip between f^5 and a^4 marks the temporary transition of the melody from one voice (descending from g^5 to f^5 and then over to $e^5 - d^5 - c^5$) to a different, inner one (ascending from a^4 to c^5). The effect of switching from a voice to another may well manifest itself as a surprising event. However, *hearing* $a^4 - b^4 - c^5$ as a different voice than $g^5 - f^5 - (e^5 - \ldots)$ has to do with the *absence*[15] of a linear-melodic connection between f^5

[15] Note that, even though f and a are not linearly related, they are still harmonically related as being part of the

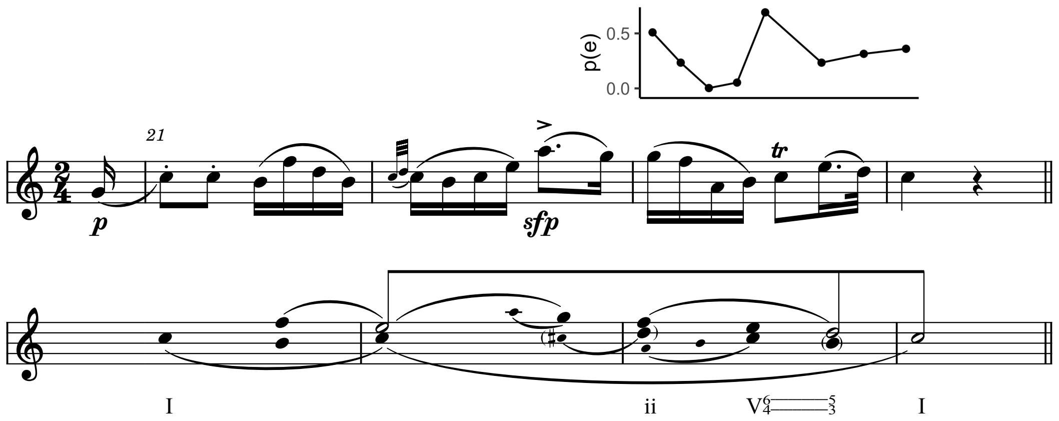


Figure 1.15 – F. Schubert, *Octet* D803, iv, mm. 21ff. Above the score, note probabilities as estimated by IDyOM are reported for mm. 23-24 (Pearce, 2018). In the lower staff, an analysis of the melody is provided, highlighting an implied polyphonic texture.

and a^4 (visualised by the absence of a slur connecting the two tones in the analytical graph in 1.15), and this (categorical) absence may manifest itself in, but is not *constituted by*, its being (continuously more or less) surprising.

In summary, an internally-explanatory account of this passage would explain the intentional disposition of a^4 in terms of its being part of an inner voice linked to the following c^5 rather than the previous f^5. From this perspective, the surprisingness of the resulting skip is merely a consequence of the apparent melodic interval that is formed in the musical surface when the implied polyphony is "collapsed" in a seemingly monophonic texture. The size of the interval and its surprisingness may well serve as a cue to infer the shift to a different voice, which in turn may be part of a mechanistic account of voice separation (cf., e.g., the approach proposed in Sauvé, 2018). Surprise itself, though, does not *constitute* the experience of a^4 *as* part of a different voice, which would be the content of structural hearing.

The latter may rather be reflected in *how* the expectedness of an event, leading to surprise, is estimated by the listener. Let us consider a listener who has access to the structural representation exemplified in Figure 1.15 (bottom), and who has heard the f_5 at the beginning of m. 23. For this listener, the probability of tone e_5 occurring later in m. 23 is not generically conditional on all preceding events. Since e_5 is understood as a passing tone between f_5 and d_5,[16] after hearing f_5 the listener knows with probability

$$p(e_5|f_5) = p(e_5|f_5, d_5) \cdot p(d_5|f_5) \tag{1.4}$$

that an e_5 will come at *some* point in the future. This probability can be factored into the probability $p(d_5|f_5)$ of a skip from f_5 to d_5, times the probability $p(e_5|f_5, d_5)$ of this skip being filled by a passing tone, as expressed in Equation 1.4. Note how this probability is independent

same harmony, which is visualised in the analytical graph through their vertical alignment.

[16] For simplicity, we ignore any possible alternative interpretation here.

of all other tones occurring between f_5 and e_5, which are syntactically unrelated to either of these tones. The probability $p(e_5|C(t))$ that is actually reflected in the listener's surprise rating for encountering event e_5 at time t, would then need to take into account (1) the time-independent probability $p(e_5|f_5)$ that an e_5 (as opposed to, say, a g^5) will occur at *some* point in time after the f_5 is encountered, as well as (2) the time-dependent probability distribution $p(C(t_{f_5},t)|e_5)$ quantifying how likely it is that the additional material $C(t_{f_5},t)$ is interpolated between f_5 and the current time t if an e_5 is to occur: in other words,

$$p(e_5|C(t)) \propto \underbrace{p(e_5|f_5)}_{\text{Eq. 1.4}} \cdot p(C(t_{f_5},t)|e_5). \tag{1.5}$$

It is in such a partitioning of the function that computes the probability of an event, not in its numerical value, that we may find a reflection of structural hearing.

1.4.5 *Which* expectancy

Accounts of musical expectation typically differentiate between *what*-expectancy and *when*-expectancy (Huron, 2006; Rohrmeier and Koelsch, 2012): the former refers to the nature of the expected event (e.g., expecting the occurrence of a tonic chord, as opposed to a Neapolitan-sixth chord, following a dominant chord), whereas the latter refers to the temporal location of the expected event (e.g., expecting something to happen on the next beat as opposed to a sixteenth-note earlier than the next beat). Both these kinds of expectancy give rise to meaningful perceptual effects, such as the experience of reward when the expectations are fulfilled (Salimpoor et al., 2015a), and a surprise response when the expectations are somehow violated (Cheung et al., 2019; Koelsch, Gunter, et al., 2000; Maess et al., 2001; Patel, 1998).

In common-practice tonality, cadential settings marking formal boundaries are typically associated with a specific pattern of expectancy (Bergé and Neuwirth, 2015; Caplin, 1998; Sears et al., 2014; Sears, Spitzer, et al., 2018): a root-position dominant chord V supporting scale degree $\hat{2}$ in the topmost voice induces the expectation of a root-position tonic chord I supporting scale degree $\hat{1}$ (*what*-expectancy) to come at a specific metrical location (*when*-expectancy). Weakening and evading cadential resolution is a common compositional device for introducing formal expansion (Schmalfeldt, 1992): it is a stylistic requirement that, if a cadence is in some way unsatisfactory, the cadential process is to be repeated to eventually achieve closure (Caplin, 1998; Hepokoski and Darcy, 2011). The weakening of cadential closure is often achieved through the non-normative resolution of some of the voices comprising the contrapuntal texture (Neuwirth, 2015; Wall et al., 2020): in this way, the music *deceives* the listener by failing to fully uphold the promise of cadential resolution. However, other factors may also contribute to weaken a cadence, even in those cases where the cadence is otherwise contrapuntally normative (apparently, at least).

In Figure 1.17a, the dominant established in m. 122 marks the first attempt to an Essential Expository Closure (EEC; Hepokoski and Darcy, 2011): at first glance, one may see in the

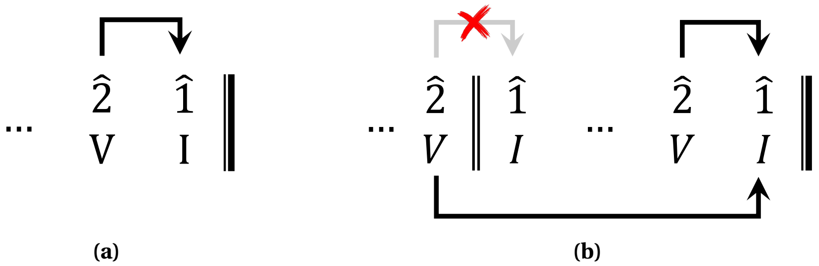
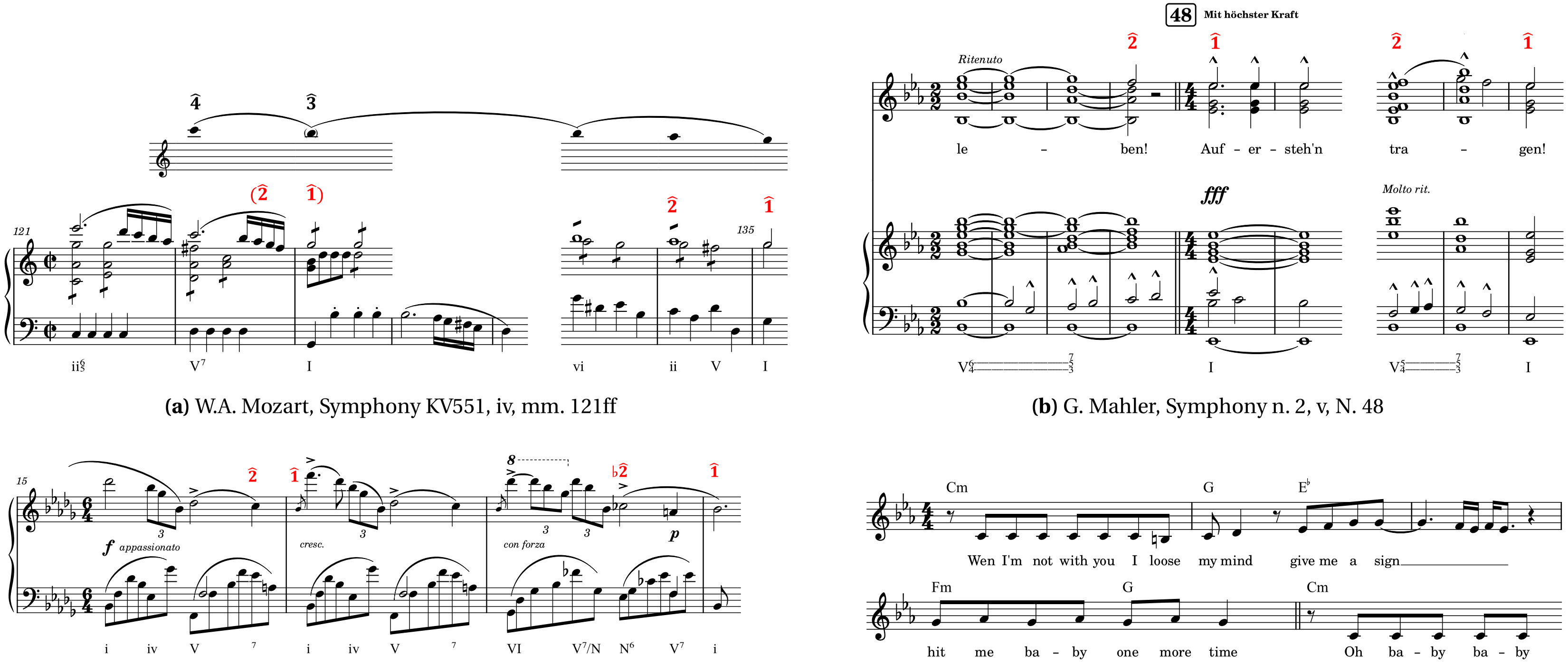


Figure 1.16 – (a) A minimal requirement for a fully satisfactory cadence in common-practice tonality is the resolution of a root-position dominant into a root-position tonic, supporting a descent from $\hat{2}$ to $\hat{1}$ in the topmost contrapuntal line. (b) Cadential weakening or evasion can result in formal expansion, whereby a new cadential process is initiated. A full resolution of the expectation induced by the first dominant may only occur with the arrival of the final tonic, rather than on the immediately following tonic.

violin line descending to the tonic $\hat{1}$, supported by a V – I harmonic progression, a satisfactory cadential motion (a Perfect Authentic Cadence, PAC; Caplin, 1998). However, a second cadential attempt is repeated in m. 135: if structural closure had already been achieved in m. 123, why repeat the EEC? Indeed, the listener may recognise in their own experience a feeling of unresolved expectancy after the first cadential attempt, which reflects in the necessity of repeating the cadence. This feeling is to be ascribed to the presence of an implied b_5 ($\hat{3}$) as the top-most voice in m. 123, descending from the prominent c_6 of the preceding measure (cf. top staff): this makes the resolution to g_5 only the motion of an inner voice, with the structural line still being suspended on scale degree $\hat{3}$ all the way to the proper EEC in m. 135.

The experience of an incomplete resolution is even stronger in cases of genuine cadential elision or evasion (Schmalfeldt, 1992). In Figure 1.17b, the tonic harmony on rehearsal mark 48 can only partially be considered as a resolution of the preceding dominant harmony. The tonic chord does harmonically resolve the dominant, as highlighted in particular by the rising line of the horns (g_3 to e_4). However, the distribution of the text indicates that the dominant harmony ends a formal section, whereas the tonic harmony initiates a new one with the peremptory *Aufersteh'n*, the *fff* and the *marcato* markings in the vocal part. The dominant should be understood, to some extent at least, as unresolved, its cadential promise being interrupted by new material with initiating function. This elusion of cadential closure is crucial for the experience of form, as it allows the music to keep some tension until the arrival of the true final tonic, aligned with the end of the text on the word *tragen*.

Similarly, in Figure 1.17c, the cadential dominant established in m. 15 does lead to a fleeting $b\flat_4$ as its tonic resolution, but this is rather to be understood as an embellishment initiating the following *one-more-time* phrase (Schmalfeldt, 1992) than as the resolution of the preceding dominant harmony (Gran, 2017). The expectation engendered by this dominant is then only resolved with the arrival of a tonic in m. 18. Cadential evasion, where a cadential dominant is followed by a tonic that does not resolve the preceding dominant but rather initiates a new



(a) W.A. Mozart, Symphony KV551, iv, mm. 121ff

(b) G. Mahler, Symphony n. 2, v, N. 48

(c) F. Chopin, Nocturne op. 9 n. 1, mm. 15ff

(d) B. Spears (M. Martin), *Baby One More Time*

Figure 1.17 – Four examples of cadential weakening or evasion.

formal section, is also common in Pop music, where it may be used to hit the tonic onset of the verse "one more time" without fully resolving all harmonic tension (Figure 1.17d).

In all examples, the surface does seemingly provide us with the promised root position tonic supporting $\hat{1}$, satisfying the *what*- and the *when*-expectancy elicited by the cadential dominant. Nevertheless, for one reason or another, the arrival of the tonic does not mark satisfactory formal closure: even after hearing the tonic chord, the expectation induced by the dominant is to some extent still active and awaits resolution into *another* tonic to come in the future. In order to *hear* this state of things, besides a notion of *what* entity is expected and *when* it is expected, listeners should also assess *which* instance of that entity is the one that is expected. To be sure, different features of the surface may serve as cues that allow listeners to assess whether the tonic that immediately follows a dominant *is* the instance of the tonic that was expected, or whether another instance of the tonic is to be awaited. However, the experience of a lack of resolution is not to be understood as a surprise response to the surface event itself (i.e., to the expected *what*), but rather as the result of attributing a certain structural identity to the surface event – i.e., assigning to that event a structural relation to some event *other* than the preceding dominant.

1.4.6 Music theory as a computational-level theory of cognition

These examples show that there is a large discrepancy between our understanding of what the phenomenon of structural hearing is – based on music-theoretical introspection – and our knowledge of concrete elementary phenomena of auditory perception that may provide mechanistic explanations for the emergence of that phenomenon. Current mechanistic accounts do not explain or predict the kind of experiential effects that are possible under music-theoretically inspired accounts. Fundamentally, this is because the *structural* identity of events is typically not considered part of the *explanandum* of most existing theories: these rather focus on explaining the emergence of observable phenomena (e.g., particular responses to consonance or syncopation) based on the surface identity of events, irrespectively of their structural identity.

In the present work, we propose to take a different approach. Music-theoretical introspection may be seen as a computational-level characterisation of a cognitive task: that of mapping the musical surface into some specific phenomenal experience – one that reflects the effects and differentiations we have exemplified in the pevious sections. Our interest, then, is to investigate the nature and the very existence of this *specific* phenomenon for which we have a definition at Marr's (1982) computational level, yet no full-fledged mechanistic account.

By adopting a top-down approach, we are guaranteed to formulate algorithmic and implementational hypotheses that are compatible with the computational-level understanding of the phenomenon. By falsifying or supporting such hypotheses, we will be learning whether we can interpret some aspect of cognitive behaviour as a manifestation of an algorithmic and implementational strategy for performing the computation we assume as a definition of the phenomenon. While this is an advantage of a top-down approach, it comes at a cost. Namely, the hypotheses we formulate, being driven by computational-level arguments, are virtually guaranteed to be factually false when interpreted as models of actual cognitive behaviour. Specifically, it is unlikely that actual cognitive and brain machinery can implement the mapping from surfaces to interpretations exactly as we define it through, e.g., a grammar. For example, the cognitive system may have processing limitations that only allow for the mapping to be performed under limited circumstances (e.g., due to memory constraints). This distinction reflects the analogous distinction drawn in linguistics between the *competence* of the ideal speaker and the *performance* of the actual speaker in specific tasks (Chomsky, 1965). The burden of proof for supporting a given computational-level hypothesis and its algorithmic and implementational implications is then to specify how the observed behaviour (reflecting the listener's performance) can be seen as an approximate implementation of the given computation (reflected in the competence model). In summary, while both bottom-up and top-down perspectives present difficulties, we commit here to a top-down approach based on the observation that, at present, we have greater knowledge of the phenomenon at the computational level than we have knowledge of its plausible implementation in terms of elementary empirical phenomena.

In Chapter 3 we will discuss how specifying the notion of inference (as introduced in this Chapter) into a computational- and algorithmic-level theory based on generative grammars and incremental parsing may account for many aspects of the phenomenology of structural hearing. Before that, in the next Chapter, we overview current empirical findings that have been interpreted as unveiling aspects of structural hearing in music. Since individual chapters in the body of the thesis will present their own literature review, the discussion here is limited to highlighting the relationship between existing empirical approaches and the theoretical perspective we adopt. In particular, we focus on the core cognitive capacities that, based on the discussion from Chapter 1, we take as defining of structural hearing, namely (1) the acquisition of a generative model for a given musical idiom based on exposure to its repertoire, (2) the cognitive reality of representations encoding specific kinds of structural information, and (3) the listeners' capacity to form such representations in real time during listening.

2 Empirical approaches

In Chapter 1, we have proposed an understanding of structural hearing as the kind of music-induced experience that results from inferring representations of structural interpretations upon exposure to the musical surface. Such representations would encode the network of structural relations that link surface events with one another according to the music theory of a given musical idiom. This approach being primarily theory-driven, we face the challenge of making it empirically testable. As a first step in this direction, in this Chapter we overview established empirical observations about how the listener's perceptual experience is influenced by musical structure. In doing this, we focus on whether these observations support, contradict, or are compatible with our notion of structural hearing.

2.1 The listeners' implicit knowledge

Since most aspects of musical structure are cultural conventions, familiarity with a given repertoire is the minimal precondition for listeners to develop a capacity for structural hearing in the corresponding musical idiom. In particular, listeners who are not familiar with a musical idiom may respond differently from "native" listeners to the structural features of musical pieces (Castellano et al., 1984; Lartillot, 2011; Popescu et al., 2021; Rohrmeier and Widdess, 2017). Through exposure to musical surfaces comprising a representative repertoire, listeners can rapidly acquire knowledge of the structuring principles of the corresponding idiom (Popescu et al., 2021; Rohrmeier, 2010; Tillmann, 2005). Such knowledge is largely *implicit*: listeners are not consciously aware of what structuring principles are at play, but those structuring principles are nonetheless deployed as part of the cognitive operations involved in processing music (Reber, 1989).

Establishing what kinds of idioms are implicitly learnable based on exposure has been addressed empirically both in computational and behavioural studies (see Rohrmeier and Rebuschat, 2012 for an extensive review). Processes of statistical learning allow listeners to become sensitive to statistical regularities present in a repertoire of musical surfaces (e.g., Pearce, 2018): in turn, such statistical regularities can be exploited to infer representations (e.g., of tonal stability; Tillman et al., 2003) and improve prediction (Rohrmeier and Koelsch, 2012). While *statistical learning* refers, per se, to any kind of statistical regularity, individual models typically commit to specific kinds of statistical features. Most commonly, these include the frequency of (co-)occurrence for individual surface entities (e.g., the frequency of individual pitches Moss et al., 2022; Tillman et al., 2003) or the frequency of sequences of surface entities (e.g., chord-transition statistics, Moss et al., 2019; Rohrmeier and Cross, 2008; White and Quinn, 2018 or Markov-like features of melodies Pearce, 2018; Pearce and Wiggins, 2012). As a consequence, studies showing effects of statistical learning do not exhaust the range of possible statistical features that listeners may acquire from exposure to the repertoires, and in particular what kinds of rule systems underlie the statistical regularities that are observable in th emusical surfaces.

Artificial-grammar-learning paradigms (AGL) explicitly account for the process of learning specific rule systems (Fitch and Friederici, 2012): listeners are exposed to a training set of surfaces reflecting a certain unknown idiom, and their acquisition of the underlying structuring principles is probed by testing their capacity to deploy such knowledge in a task involving novel stimuli. While the artificial nature of the tested grammars precludes the ecological interpretation of the results, evidence from AGL can help determine what are the features of an idiom that make it learnable *in principle* given a certain type of exposure. In the musical domain, several AGL paradigms have shown that listeners acquire the statistical features induced in the musical surface by specific rule systems (Jonaitis and Saffran, 2009; Koelsch et al., 2016; Loui and Wessel, 2008; Loui et al., 2010; Prince et al., 2018; Rohrmeier et al., 2011). However, a limitation when interpreting results from AGL paradigms is their typical reliance on explicit grammaticality or familiarity ratings: the acquired implicit knowledge is probed by asking participants to distinguish between surfaces that "belong" to the idiom and those that are not. This means that these experiments only allow us to conclude that some grammar within a class of grammars with the same language (i.e., with the same weak generative capacity; Chomsky, 1965) has been acquired. In other words, AGL paradigms often do not directly demonstrate the attribution of specific structural representations, or of structural representations with specific features, to specific surfaces. This limits the capacity of such studies to inform our understanding of the *complexity* of grammars that are learnable, since grammars with different complexity can be weakly equivalent (i.e., generate the same language).

The problem here is that surface statistical features (such as frequency of occurrence or transition probabilities) do not reflect the kind of implicit knowledge that a structural listener, as outlined in Chapter 1, may be expected to possess. From a music-theoretical perspective, part of music's putative complexity has to do with its capacity to express structural relations that are arbitrarily non-local, as opposed to local statistical features of the surface (Rohrmeier and Pearce, 2018; cf. Tymoczko, 2020 for a contrasting view). This is particularly evident in the examples discussed in Section 1.4.4 and 1.4.5: the explanation for the presence of some events in the surface is their being related to events that are arbitrarily distant from them. Structural interpretations, that we hypothesised to be the object of mental representation on part of the listener, are expected to encode such non-local relations. As a consequence, a core question to be addressed – one that cannot be fully addressed with AGL paradigms relying on explicit grammaticality judgements – is whether a grammar that is complex enough to account for such non-local dependencies can be implicitly acquired by listeners through exposure to the repertoires (Pearce and Rohrmeier, 2018; Rohrmeier, 2013; Rohrmeier and Pearce, 2018).

In an extra-musical auditory domain, listeners exposed to artificial languages allowing for different kinds of hierarchical embedding have been shown to be sensitive to syntactic violations that break non-local dependencies (Rohrmeier et al., 2012). In the musical domain, Rohrmeier and Cross (2009) showed that context-free rules predicted listeners' familiarity ratings better than bigram probabilities in a musical AGL paradigm, supporting that listeners can acquire a competence for center-embedding and recursion in the musical domain too. Convergent computational evidence comes from studies on unsupervised grammar induction of Jazz harmony (Harasim et al., 2020): training on a limited corpus of chord sequences allows for automated parsing of their syntactic structure, achieving above-chance performance when assessed against expert-annotated syntactic trees. Complementing a behavioural AGL paradigm with analysis of fMRI data, Cheung et al. (2018) showed evidence that listeners are sensitive to hierarchical depth in a newly acquired idiom based on repetition structure. However, such studies support learnability *in principle*, leaving the question open whether (tonal) listeners *actually* acquire this kind of grammar.

Psychological evidence for the cognitive reality of the acquisition of a capacity for processing complex syntax is conflicting. On one hand, several studies demonstrate that listeners are generally only marginally sensitive to manipulations of global structure, especially at large time scales (Atalay and Tekman, 2006; Cook, 1987; Levinson, 1997; Tillmann and Bigand, 2001; Tillmann and Bigand, 2004). On the other hand, listeners have been found to be sensitive to scrambling the order of segments at different hierarchical levels in a musical piece (measure, phrase, section), indicating that the brain processes the coherence of musical structure up to the timescale of minutes (Farbood et al., 2015). Behavioral and neural responses further reflect the violation of the non-local dependency between the initial and final tonic of a tonal piece, even when local relations were preserved (Koelsch et al., 2013; Zhang et al., 2018): in particular, the persistence of a sense of "home key" after modulation has been shown to extend up to the timescale of ~20s (Farbood, 2016; Woolhouse et al., 2016).

An additional implication of hierarchical models of harmonic structure is that multiple nested structural relations may be embedded simultaneously, resulting in different degrees of structural complexity. The processing of harmonic progressions has been shown to reflect their depth of embedding (Cheung et al., 2018; Herff, Bonetti, et al., 2023; Ma et al., 2018a; Ma et al., 2022), which has been further related to emotional responses such as perceived tension (Farbood, 2012; Lerdahl and Krumhansl, 2007; Rohrmeier, 2013).

The hypothesis that multiple structural relations may simultaneously coexist further entails that, in principle, multiple expectancies towards different targets may be active at the same time. This diverges from many expectancy-based accounts of music processing that treat expectedness as a single scalar quantity (quantified, e.g., in terms of information-theoretical surprisal of an event conditional on the preceding context). Preliminary evidence for the cognitive relevance of multiple expectancy streams comes from probing listeners' expectations about the *number* of expected chords when a chord progression is interrupted. Listener's responses are better captured by a model that also accounts for hierarchical expectancies towards multiple goals, as opposed to a sequential model that only accounts for a single stream of expectation (Herff, Harasim, et al., 2021).

It is plausible that refined musical expertise may be a precondition for hierarchical hearing. On one hand, in general, untrained listeners typically exhibit non-trivial sensitivity to musical structural features (Koelsch, Gunter, et al., 2000; Tillmann, 2005). Consistently, Koelsch et al. (2013) found no difference between experts and non-experts in a non-local-dependency violation paradigm. In contrast, Ma et al. (2018b) rather indicate that musical proficiency is necessary for acquiring the capacity to deal with complex structure, with Ma et al. (2018a) further suggesting that different processing mechanisms may underlie structural perception in experts and non-experts.

Overall, whether and under what conditions listeners develop and deploy a capacity to recognise musical structure beyond local statistical (e.g., Markovian) regularities is only partially settled by the empirical literature, and seems generally unclear. Part of the reason for this knowledge gap is that the theoretical framing of most empirical observations pertaining to structural features does not rely on the existence of structural relations between (possibly distant) events, while also not being incompatible with this understanding.

2.2 Tonality

One core set of organisational principles in Western music falls under the umbrella-term *tonality*. Such principles reflect how individual tones acquire meaningful attributes when being functionally embedded in a context that establishes a given key. They include the notion of a tonal center and a hierarchical "ranking" of the other tones relative to one another and to the tonal center (Krumhansl, 1990; Lerdahl, 2001). This hierarchy is further associated with tones serving different functions relative to one another (Agmon, 1995; Jacoby et al., 2015; Rohrmeier and Cross, 2008; White and Quinn, 2018).

Such notions have been operationalised in empirical research in order to unveil perceptual correlates of tonal organisation. One fundamental observation from tonal-priming paradigms is that the perceptual attributes of individual tones or chords are strongly influenced by the preceding context. For example, individual tones acquire the quality of being more or less stable (Krumhansl and Kessler, 1982; Krumhansl, 1990), as well as several more subtle perceptual *qualia* that listeners consistently attribute to tones depending on their relationship to the preceding context (Arthur, 2018). Furthermore, tonal context influences the listener's performance in behavioural tasks such as memory (Farbood and Mavromatis, 2018; Krumhansl, 1979), pitch (e.g., Marmel et al., 2008; Sears et al., 2021) and timbre discrimination (e.g., Prince et al., 2015; Sears et al., 2023) , as well as the temporal and cortical distribution of neural responses to tonal stimuli (Janata et al., 2002; Marmel et al., 2011; Tillmann et al., 2003).

These phenomena correlate with aspects of the structural identity of tones that are distinct from their surface identity: in a tonal context, for example, they reflect the identity of tones as *scale degrees* relative to the tonal center. These patterns are interpreted as the manifestation of a representation of tonal relations that is available to enculturated listeners as a consequence of psychoacoustic principles (Huron and Parncutt, 1993; Milne et al., 2015; Milne et al., 2023; Parncutt, 2011) and statistical learning from the repertoires (Bharucha and Stoeckig, 1987; Tekman and Bharucha, 1998; Temperley and Marvin, 2008; Tillman et al., 2003; Tillmann et al., 2000). For example, individual units in a neural-network model like MUSACT (Bharucha, 1987; Bharucha and Todd, 1989; cf. also Tillman et al., 2003) represent surface tones (i.e., pitch classes as they appear in the musical surface) as well as chords and keys as abstract harmonic entities. Activation of the surface-tone units spreads towards the chord and key units and resonates backwards to the surface-tone units until an equilibrium is reached. In this kind of model, representations are modelled in terms of the weightings of the connections between units, and the weightings themselves are learnt through training over the repertoires. Overall, psychological accounts based on tonal priming and the resulting key profiles envisage a kind of mental representation that reflects the relationship of tones with underlying latent entities (in the form of stability profiles, probability distributions over pitch-classes, or of symbolic representations of chords and keys). However, the resulting picture is incomplete in two respects.

First, it is widely acknowledged that tonal-priming effects are the result of both sensory and cognitive representations of the tonal context (Bigand et al., 2003; Sears et al., 2019): the former encode information about the specific pitches that are presented, whereas the latter encode some abstraction that reflects the key established by the context. However, it is unclear to what extent the emerging representations (even cognitive ones) are still tied to the sensory content of the context: specifically, the available empirical evidence does not allow us to determine to what extent contexts differing in their sensory content necessarily result in distinct (e.g., key-specific) representations (Figure 2.1a). An alternative scenario is that, besides such key-specific representations, listeners also form key-independent representations that specifically encode the tonal relations between events, abstracting away the sensory nature of the events themselves (Figure 2.1b). The latter kind of representation would be compatible with the notion of structural representation introduced in Section 1.3.4: there, we posited that a listener "hearing" a surface "as" comprising certain structural relations would form a representation of precisely such structural relations. Importantly, evidence from tonal-priming is not inconsistent with either scenarios 2.1a or 2.1b, but it cannot disambiguate between the two. In this thesis, we will address this issue with a dedicated experimental paradigm in Chapter 5.

Second, the hierarchy of stability or relatedness of tones and chords within keys (Krumhansl and Kessler, 1982; Lerdahl, 2001) is only one component of the music-theoretical notion of

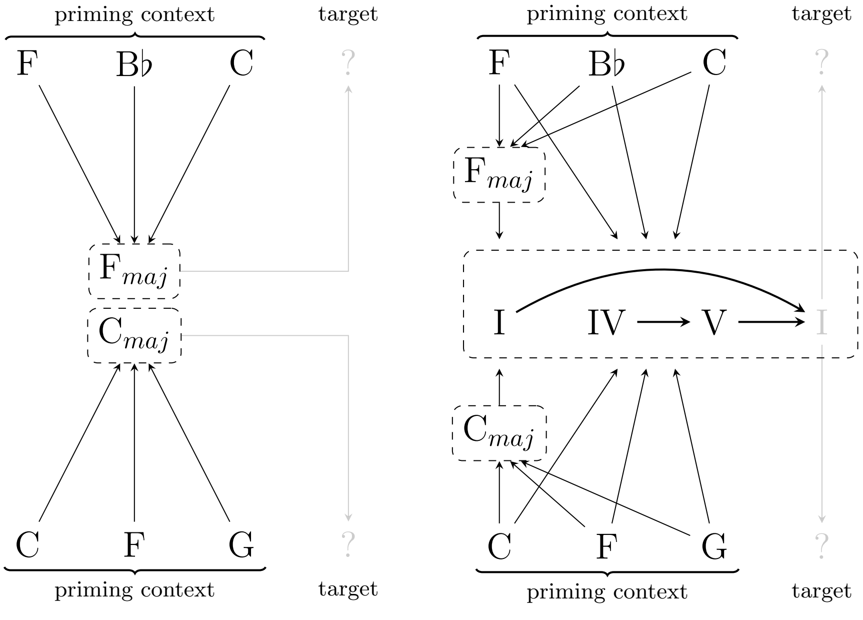


Figure 2.1 – In a typical tonal priming paradigm, some key-establishing context primes the perception of a target (a tone or a chord). An influence of the priming context on the target (gray) is compatible with two underlying scenarios. **(a)** Exposure to some context-establishing material activates a key-specific representation that is responsible for priming the target. In this scenario, a context establishing the key of C major (bottom) and a context establishing the key of F major (top) activate distinct representations. **(b)** Exposure to some context-establishing material activates a key-independent representation that encodes the relations between the events in the context, irrespectively of the absolute pitches involved. In this scenario, a shared representation is activated by both the C major (bottom) and the F major (top) context.

tonality. The latter is also, and possibly primarily, grounded in processes of contrapuntal and harmonic elaboration. While the tonal hierarchy is an a-temporal representation, the temporal unfolding of a contrapuntal texture putatively results in the establishment of dependency relations between events as they are arranged syntactically over time (Lerdahl and Jackendoff, 1983a; Rohrmeier, 2020b; Schenker, 1935). One may wonder, then, whether tonal listeners also form representations of such temporally situated syntactic relations, besides having access to a-temporal representations of global statistical features (cf. Sears et al., 2023 for preliminary evidence in this direction). In order to address this question, we need to look into the available empirical evidence on the temporal organisation of music, in particular with respect to rhythm and grouping (Sections 2.3-2.4), expectancy (Section 2.5), and syntax (Section 2.6).

2.3 Metrical and rhythmic structure

Two core aspects of the temporal organisation of music are metrical structure and grouping structure (Lerdahl and Jackendoff, 1983a). The former refers to a hierarchy of pulse layers that identify time-points in the musical surface characterised by increased "metrical weight" (R. Cohn, 2020). In empirical research, these are operationalised as moments of hightened attention (Fitzroy and Sanders, 2015; Large and Jones, 1999) or predictability (Vuust and Witek, 2014) which are observable in terms of behavioural (Clayton, 2012) and neural (Nozaradan et al., 2012; Stupacher et al., 2016) entrainment. Importantly, the pulse patterns comprising meter do not necessarily coincide with sounded events in the musical surface. As a consequence, meter stands in a twofold relationship with respect to the musical surface. On one hand, the perception of meter is induced by the temporal regularities and accent patterns of the musical surface (Ellis and Jones, 2009; Large and Snyder, 2009; Longuet-Higgins and Lee, 1982; Toiviainen and Eerola, 2004). On the other hand, meter provides a temporal lattice, the "metrical grid", relative to which the temporal location of upcoming rhythmic events is gauged: for example, events can be perceived as "belonging" to a beat (Danielsen et al., 2023) or as (unexpected) anticipations or delays of an expected beat (Vuust et al., 2022; Vuust and Witek, 2014).

As discussed in Section 1.4.2, though, the temporal alignment of sounded events with an underlying metrical grid is not sufficient to establish some of their (putatively perceptual) attributes, even in the domain of rhythm alone. Part of the listener's experience is hypothesised to reflect the relations that events entertain with one another: for example, one event could be understood as an upbeat of another event, with the two events thus being related to one another. From this perspective, the events' temporal location as determined by the metrical structure only serves as a heuristics guiding the attribution of such relations.

2.4 Grouping structure

While metrical structure characterises the "temporal container" where events take place, grouping structure characterises the way events themselves merge to form separate chunks, possibly based on general *Gestalt* principles (Deliege, 1987; Deutsch, 1999) or criteria of predictability (Hansen et al., 2021; Pearce, Müllensiefen, et al., 2010). Grouping structure represents a plausible perceptual manifestation of structural relatedness: events that are mutually related are minimally perceived as "bound together" to form a group, and as somehow separated from other events to which they are not, or are more distantly, related. In this sense, evidence pertaining to grouping directly speaks in favour of the idea that representations of structural relations are formed during listening.

One kind of empirical evidence pertaining to the perception of grouping is obtained by probing the listeners' capacity to detect segment boundaries. The capacity of listeners to detect sequential (Sloboda and Gregory, 1980) and hierarchical (Krumhansl, 1996; Martínez, 2018; Popescu et al., 2021; for evidence supporting the relevance of hierarchical segmentation in analysis, cf. McFee et al., 2017) segment boundaries in music has been extensively investigated in empirical research. However, identifying boundaries does not necessarily imply that the entire segments of the music comprised between the boundaries are perceived as forming

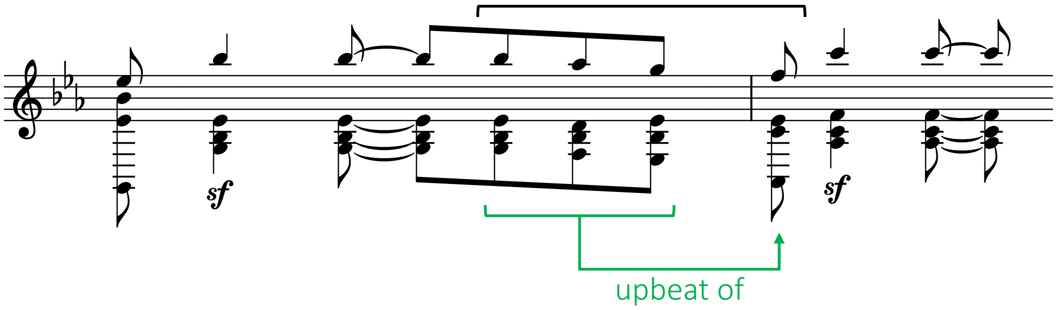


Figure 2.2 – The bracket above the score highlights a particular group in this passage from L. van Beethoven, String Quartet op. 18 n. 4, i, mm. 60f. The unitary nature of this group can be seen as the result of the existence of a structural relation with a specific interpretation (being an *upbeat of...*) linking the events comprising the group, as displayed below the score.

groups. For the latter purpose, listeners would need to experience the events comprising each group as being reciprocally related. On the contrary, for the purpose of detecting boundaries, listeners may simply learn to pick up more or less local cues (e.g., long rests, large pitch skips, or metrical location; Hamaoui and Deutsch, 2004; cf. Temperley, 2001a) that allow them to respond selectively to those moments in the music, without being sensitive to the specific structural relations holding together the events that comprise groups. As a consequence, evidence pertaining to boundary perception cannot directly address the issue of whether representations of structural relations are formed during listening.

In turn, phenomena of perceptual chunking provide implicit support for the cognitive reality of groups as representational units. For example, listeners are facilitated in recognising a previously-encountered melodic segment when it is included with the boundaries of a melody, compared to when the segment spans across a boundary (W. J. Dowling, 1973; N. Tan et al., 1981). Furthermore, the reported temporal location of an auditory click is distorted when the click occurs in the proximity of a group boundary (Sloboda and Gregory, 1980): this has been interpreted as an indication that groups are perceptual units that resist interruption. The temporal dynamics of motor action during music performance also shows evidence that musicians conceive of the musical surface in terms of separate perceptual units (Van Vugt et al., 2012). Overall, these results support that surface events are minimally linked by a relation of group-belongingness. However, a music theory for a given idiom may comprise more than one kind of structural relations: for example, Figure 2.2 shows how the existence of a group of events is associated with the presence of a structural relation of "upbeatness" between those events. The additional connotation of "upbeatness" cannot be reduced to group-belongingness. This leaves the question open whether listeners represent both grouping structure as well as the qualitatively distinct forms of structural relations that a music theory may consider relevant in a given idiom, and how different forms of relatedness may interact with one another. For example, grouping structure may simply reflect the constituents implied by rhythmic relations, one example of which is depicted in Figure 2.2 (Rohrmeier, 2020a).

Overall, the literature we have overviewed sofar directly investigates the representations listeners form during listening. Both with respect to tonal and to temporal relations, the available evidence does not allow us to disambiguate between the existence and the non-existence of mental representations encoding such relations specifically. In Chapters 4 and 5 we will provide evidence supporting the cognitive relevance of such representations, in particular whether they are abstracted from sensory information (e.g., key-independent) and whether the encoded structural relations are merely local or rather hierarchical. An alternative, indirect approach to investigating structural representations is to inquire into the processing phenomena that occur in real time during listening. If representations are the result of processing mechanisms, we may learn something about the nature of the former by investigating the latter. In the following Sections 2.5 and 2.6, we discuss the available empirical evidence pertaining to two aspects of real-time processing: expectancy and structural violations.

2.5 Expectancy and prediction

The notion of expectancy is traditionally at the heart of many empirical approaches to musical structure (Huron, 2006; Rohrmeier, 2013), ever since the seminal proposal by Leonard Meyer (1956, 1957). Music theory often explicitly addresses how structural features of the music manipulate listeners' expectations, such as through subtle degrees of (non-)resolution in the context of classical cadences (Caplin, 1998), and this introspective characterisation of the psychological experience can be observed empirically (Sears et al., 2014; Sears, Pearce, et al., 2018; Sears et al., 2020).

The prominence of expectancy as a psychological correlate of musical phenomena is likely due to expectancy being involved in most aspects of musical experience, while also being a very accessible object of empirical investigation. First, once a space of relevant entities has been defined, it is straightforward to model and quantify expectancy in probabilistic and information-theoretical terms as a time-dependent probability distribution $p(e|c(t))$ over entities e, conditional – in the general case – on the current overall musical context c (Meyer, 1957; Pearce & Wiggins, 2012; Temperley, 2007). Furthermore, expectancy has clear behavioral (e.g., anticipation and surprise; Huron, 2006; Margulis, 2007) and neural (e.g., Event Related Potentials; Koelsch, Maess, et al., 2000; Maess et al., 2001; Sammler et al., 2013) manifestations that can be targeted with dedicated experimental paradigms. For example, the way music communicates and evokes emotion is thought to be largely shaped by patterns of creation, resolution, and violation of expectancy (Egermann et al., 2013; Sauvé et al., 2018; Steinbeis et al., 2006), possibly through the mediation of reward-related brain mechanisms (Gold et al., 2019; Salimpoor et al., 2015a, 2015b). Expectancy also plays an important role in the context of rhythm and meter perception, where temporal regularities afford predictions towards the temporal location of metrical beats (Vuust & Witek, 2014) or sounded events (Large & Jones, 1999), with prediction mismatch resulting in syncopation and groove (Vuust et al., 2018).

2.5.1 Predictive coding

The role of expectancy across all these musical phenomena can be accounted for within the unified framework of predictive coding (Koelsch et al., 2019; Vuust et al., 2022). Predictive coding accounts formalise the relationship of sensory inputs, perception, and higher-order cognition in terms of an inference problem over the latent causes of sensory inputs that are produced by some generative process in the external environment. As the brain has no access to the generative process itself, it implements (possibly multiple levels of) predictive models that generate predictions about (current and future) sensory inputs. Such predictive models reflect the prior knowledge of plausible generative processes – knowledge that is assumed to be acquired through statistical learning over prior exposure. Predictions, weighted by their precision, are then compared against the incoming sensory inputs to quantify a prediction error that is fed forward to update the predictive model. Action can also be modelled within a predictive-coding perspective as a way to minimise prediction error by intervening on the sampling of sensory data from the external world rather than updating the internal predictive model directly (Friston, 2010).

Note how, in this account, sensory inputs only provide a bottom-up contribution through prediction errors, whereas representations of the external world are not directly acquired from the sensory data and are rather modelled as part of the internal top-down predictive model(s) (Gładziejewski, 2016). Importantly, the predictive coding framework *per se* is agnostic with respect to the specific nature of the predictive models that operate in each domain, as well as to the nature of the representations such predictive models operate on. In principle, any formal model that is capable of producing predictions over sensory inputs is a viable instantiation of the top-down component of a predictive coding model. This flexibility is a great strength of predictive coding as a general framework for human cognition, as it allows for applications on different domains (within and beyond music) each calling for its own predictive model. However, predictive coding remains underspecified exactly on the issue we intend to address, namely, the nature of representations that are formed as part of processing music.

2.5.2 Beyond expectancy

Overall, research on expectancy has proven extremely fruitful and has rightfully attracted large attention in the field of music cognition. Nevertheless, expectancy does not exhaust *all* of the musical phenomena that are theoretically associated with structural hearing. In particular, we are interested here in whether and how listeners represent structural relations specifically. Whether this is the case can be understood as an open empirical question: however, research focusing on expectancy can only partially address this issue because expectancy-related phenomena are neither necessary nor sufficient manifestations of structural relations.

Relations do not imply expectancy. First, not all music-theoretically relevant harmonic relations have clear implications towards the creation and resolution of expectancy. In particular, functional harmonic relations in the context of Western tonality (such as the relation of a dominant towards its tonic) can be and are often explicitly associated with patterns of creation and resolution of expectancy: e.g, after hearing a dominant, listeners are meant to expect the arrival of a tonic. However, even within the theory of Western tonality, not all structural relations are functional in nature, nor do they necessarily relate to expectancy. Examples of non-functional harmony can be found in (extended-)tonal repertoires in the form of “coloring” harmonies that do not engender goal-directed expectations but may still participate in other kinds of structural associations (e.g., DeVoto, 2004; Waters, 2005). Similarly, voice-leading transformations between chords as characterized by Neo-Riemannian theories (R. Cohn, 2012), hexatonic-pole “contrasts” (R. Cohn, 2016; Polth, 2011), transformations of *Quintenreihe* in *Tonfeld* theory (Haas, 2004; Rohrmeier and Moss, 2021), and other pitch-class-set-theoretical transformations (Forte, 1973) also determine relations between harmonic entities that may have perceptual relevance but do not necessarily engage with the creation and resolution of expectancy. In order to investigate the perceptual reality of these relations, methods that do not rely on expectancy-related phenomena need to be adopted. Possibly due to the difficulty of identifying observable manifestations of non-functional harmonic relations, these are notably neglected in empirical research compared to those characterizing functional harmony: a few exceptions include investigations on extended-tonal harmony by Bisesi (2017), Milne et al. (2016), and Krumhansl (1998).

Expectancy does not imply relations. Second, even in the cases where harmonic relations are in principle associated with expectancy (e.g., functional harmony), observing the manifestation expectancy is not sufficient to demonstrate the perceptual relevance of structural relations. From a structural perspective, expectancy could be understood as arising whenever an event is encountered during listening that is linked via some relation to some other event that has not occurred yet – in other words, when a relation is opened but not closed yet (Rohrmeier, 2013). Based on this theoretical correspondence between structural relations and expectancy, it is possible in principle to define “structural” models of expectancy that proceed by first (1) inferring relations between events, and, as part of the process, (2) making predictions on future events based on the relations that are currently opened but not closed. However, on the other hand, it is also possible to define models of expectancy that do not rely on the notion of relation at all, as they operate by tracking statistical features of surface events as they occur (e.g., IDyOM, Pearce, 2005). Such “statistical” models of expectancy do not infer an explicit representation of harmonic relations. In other words, for the “statistical” listener, relations are at best epiphenomenal to the statistical regularities of surface events as learnt from the repertoires: for example, they can be seen as the theorist’s way to make sense of those statistical regularities. For the "structural" listener, in turn, inferring structural relations is causally primary to casting predictions about future events. Overall, both statistical and structural models predict expectancy-related phenomena: the difference between the two approaches rather lies in the information that is assumed to be encoded in the mental and neural representations based on which predictions are made. As a consequence, it is not trivial how to disambiguate between statistical and structural perspectives by looking at expectancy alone.

2.6 Syntactic processing in music "as" language

The empirical investigation of on-line processing of musical structure has largely focused on observing real-time responses to the breaking of structural norms. Structural violations, being less expected than normative continuations, are hypothesised to elicit surprise in the listener, and to require extra-ordinary cognitive efforts in the attempt to make sense of the unexpected event and to continue the parsing process. For example, listeners are slower in reacting to an auditory click presented while listening if the music presents a sudden unexpected modulation: this has been interpreted as evidence of the additional cognitive load due to the increased structural complexity associated with the modulation (Berent and Perfetti, 1993).

Putative mechanisms of structural processing are also reflected in brain activity. In particular, encountering events that are unrelated to pre-existing harmonic context results in well-established Event Related Potentials (ERP), including an early right-lateralised anterior negativity (ERAN) observed selectively in response to harmonic violations, as well as a later anterior negativity (N5) observed both in response to harmonically normative chords and to harmonically deviant ones (Koelsch et al., 2013; Koelsch et al., 2003; Koelsch, Gunter, et al., 2000; Ma et al., 2018a, 2018b; Zhang et al., 2018). The latter is especially interesting insofar as it has been interpreted as a result of the cognitive effort of integrating a newly-encountered event into the pre-existing harmonic context (Koelsch, 2011; Koelsch, Maess, et al., 2000). Over the course of a normative stimulus (without structural violations), the amplitude of the N5 decreases as the integration of new expected material is facilitated by the increasing context; however, when a deviant chord is encountered, an N5 with larger amplitude is elicited. This may reflect the increased difficulty of integrating an unexpected event in the pre-existing context, and possibly the necessity to revise the context itself (Koelsch et al., 2003; Koelsch, Maess, et al., 2000). Overall, the N5 ERP component is likely related to the attribution of structural interpretations (musical "internal" semantics; Koelsch, 2011, 2013) to chords, by virtue of their syntactic relatedness to other chords.

A later positive ERP (P600) previously associated with the processing of complex syntax in language was also shown to be elicited by out-of-key chords in harmonic stimuli (Patel et al., 1998). Besides evidence from ERPs, modulation of brain activity in the alpha (Ruiz et al., 2009) and theta (Herff, Bonetti, et al., 2023) frequency ranges has been observed in response to structural violations or increased structural complexity. Furthermore, in performing musicians, embodied mechanisms of action planning and motor control have been shown to be involved in processing musical structure (Bianco et al., 2016; Bianco et al., 2015; Sammler et al., 2013). Collectively, this evidence is consistent with the view that musical events are integrated into some kind of structural representation based on processes that are distinguishable from those that identify (the violation of) merely sensory regularities (Koelsch, 2013).

Both musical and linguistic structural violations have been shown to result in increased activation in overlapping areas of the brain (Koelsch et al., 2005; Levitin and Menon, 2003; Maess et al., 2001). This increase is further strengthened when linguistic and musical violations occur jointly, supporting that processing of language and music interfere with each other (Kunert and Slevc, 2015). Such interference manifests itself in behaviour as well: joint violation of musical and linguistic syntax results in impaired linguistic comprehension (Fedorenko et al., 2009), reading times (Slevc et al., 2009), working-memory performance (Fiveash et al., 2018; Fiveash and Pammer, 2012), and cognitive control (Slevc and Okada, 2015; Slevc et al., 2013). These effects are distinguishable from those associated with processing of linguistic semantics (Hoch et al., 2011; Koelsch et al., 2005) and phonology (Fiveash et al., 2018), suggesting that they specifically reflect the processing of syntax.

Bringing together this body of evidence, the Shared Syntactic Integration Resource Hypotheses (SSIRH; Patel, 2007) and the Syntactic Equivalence Hypothesis (SEH; Koelsch, 2013) propose that structural representations in music and language are distinct, but that activating and manipulating such representations during processing recruits common cognitive and neural resources. Evidence in this direction would strongly support that musical structure is processed in a somewhat similar sense as linguistic structure is, i.e., by integrating structural information incrementally resulting in the activation of structural representations. However, the degree to which common brain mechanisms are involved in both music and language is still debated (Chen et al., 2023; Peretz et al., 2015).

2.7 Open issues

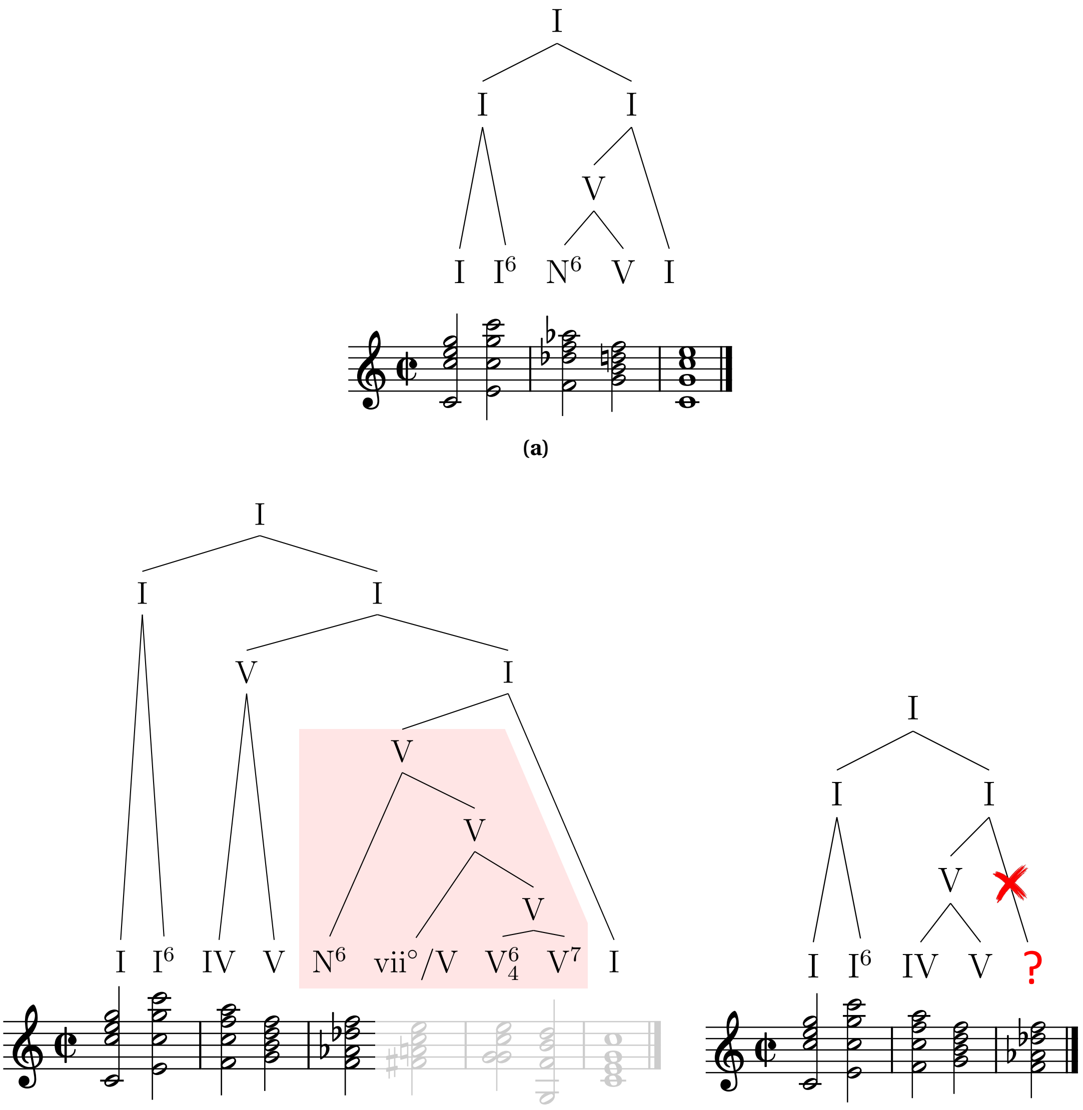
Overall, empirical evidence suggests that structural features are the object of dedicated (yet not necessarily domain-specific) processing mechanisms that build on top of lower-level sensory processing (Koelsch, 2011; Peretz and Coltheart, 2003). However, the methodological reliance on syntactic violations limits the interpretability of these results with respect to the more specific notion of structural processing that we are interested in, namely, the incremental emergence of structural representations. By comparison, experimental manipulations in linguistic paradigms can selectively control specific aspects of processing complexity in fully grammatical sentences, such as depth of embedding or dependency locality. On the contrary, musical experiments typically present outright unidiomatic chord progressions that sound "weird". Even when the manipulation is relatively subtle (e.g., comprising an unexpected but in-key chord, cf. Tillmann et al., 2006) the observable effects of ungrammaticality demonstrate the existence of *some* form of structural processing that is being disrupted, but are not necessarily informative about the nature of the underlying "ordinary" processing mechanisms as they operate on idiomatic stimuli.

Another difficulty in interpreting available empirical evidence lies in the fact that different kinds of structural violations are adopted in experimental stimuli (cf. Featherstone et al., 2012). In some cases, harmonic violations comprise chords that are unlikely to occur in the given context (e.g., they contain out-of-key tones), yet syntactically correct (Figure 2.3a).[1] In other cases, syntactic violations may also lead to *recoverable* parsing failures: for example, a Neapolitan sixth chord following a dominant chord, as unexpected as it is,[2] may signal the beginning of a prolonged dominant region that will eventually lead to a new dominant chord resolving to the tonic (Figure2.3b). However, the syntactic violation may turn out to be unrecoverable: for example, a deviant chord in final position cannot be integrated in the previous context (Figure 2.3c). The specific time-course of processing may differ depending on whether listeners completely suppress interpretation (b) (e.g., if they know veridically that the N^6 chord is the *last* chord), or whether the cognitive processor still attempts to recover by hypothesising the structure (b), which is only abandoned when the following chord fails to come. More generally, the processing difficulty in these three scenarios may result from different kinds of mechanisms, including a difficulty to access or activate a representation of an infrequent chord (a), the necessity to update beliefs about the existing context (b), and, eventually, outright processing breakdown (c). All these three mechanisms may contribute to the observed ERPs: for example, the observation that N5 amplitude in scenario (a) is lower than in scenario (c) is compatible with the understanding that (a) exhibits a different kind of processing complexity than (c) (Koelsch, Gunter, et al., 2000). Similarly, Featherstone et al. (2013) showed differentible ERP responses to resolved vs. unresolved structural violations. Despite such cues, a systematic investigation of possible processing mechanisms specified at the algorithmic level, which may provide a computational interpretation to the observed ERPs, has not been undertaken.

That the processing mechanisms underlying music perception operate in some way analogously to structure-building processes in language is supported by the processing overlap between language and music. It should be noted, though, that musical processing may still operate identically to linguistic processing at the computational level of description without there being any overlaps at the implementational level (Fedorenko and Shain, 2021). In particular, overlaps between music and language may be due to stages of processing that are only accessory to the mechanisms that specifically implement the inference of representations. For example, observations pertaining to musical syntactic violations and their interference with linguistic syntactic violations may well be related to the effect of surprise, and possibly mere sensory surprise (Bigand et al., 2014), rather than processing itself. In this respect. most paradigms control for non-syntactic sources of surprisal, for example by including control conditions that are deviant in timbre, consonance, or loudness (e.g., Fedorenko et al., 2009; Fiveash and Pammer, 2012; Ishida and Nittono, 2022; Koelsch, Gunter, et al., 2000; Koelsch et al., 2007; Kunert and Slevc, 2015; Slevc et al., 2013). Taking this kind of controls into account may show that structural violations are intrinsically more salient than timbral or loudness-related violations in the specific experimental tasks, or, at best, that the observed effects of

[1] Among the twelve published corpora of the Digital and Cognitive Musicology Lab (DCML; cf. Hentschel, Moss, et al., 2021), which comprise tonal music from the XVIII and XIX centuries annotated after Hentschel, Neuwirth, et al. (2021), Neapolitan chords are ~ 9 times less likely to occur in major than in minor.

[2] In the DCML corpus, only 0.3% of all dominants are followed by a Neapolitan chord, which drops to 0.07% in major keys.



(a)

(b)

(c)

Figure 2.3 – Three kinds of structural violation in the stimuli by Koelsch, Gunter, et al. (2000), following the syntax of tonal harmony by Rohrmeier and Neuwirth (2015). **(a)** The Neapolitan chord (N^6) is somewhat unusual, especially in a major tonal context, but it is still employed normatively as a pre-dominant. **(b)** The N^6 following a dominant chord may be interpreted as initiating an interpolated dominant region (red) leading to a cadence on the tonic.**(c)** Only once it is clear that the N^6 occupies a final position, syntactic integration with the previous context becomes impossible, leading to an unrecoverable violation.

interest have indeed to do with syntactic unexpectedness specifically, as opposed to generic unexpectedness (Fiveash, 2018). However, this still does not rule out that the specific effects that are observed may reflect, e.g., the simple acknowledgment of something syntactically unusual happening, as opposed to the online processes that are specifically involved in the incremental inference of structure. An exception to this is the observation that the N5 ERP component is also observed in non-deviant chords, thus possibly reflecting structural integration (Koelsch, Gunter, et al., 2000). However, an N5 ERP has been recently observed in response to loudness deviants, which challenges its interpretation as a specific marker of syntactic processing (Ishida and Nittono, 2022).

Overall – differently from how syntactic processing is addressed in linguistics – musical paradigms are mostly concerned with notions of in-keyness (Krumhansl and Kessler, 1982) and (un)expectedness (Huron, 2006). The underlying representations take the form of (e.g.) distributions of relative stability or expectedness of events in context, not of networks of relations between events. Consistently with the expectancy-based framework, recent work on a large corpus of Pop chord progression has attempted to predict ERP amplitudes at different latencies based on cognitive and sensory models of surprisal (Goldman et al., 2021). Interestingly, neither the amplitude of the ERAN and of the P600 (which typically reflect violations) nor the amplitude of the N5 (which has been observed in normative chords as well) have been found to be linearly related to the degree of cognitive or sensory surprisal. This may indicate that the association between ERPs and expectedness does not fully generalise from the artificial stimuli adopted in previous studies to the large set of idiomatic Pop chord progression adopted in the later study, but it may also indicate that expectancy does not fully capture the phenomenon of musical syntax as reflected in the observed ERPs.

In summary, it is unclear what the available empirical evidence tells us about musical syntax and the way it is experienced and processed by listeners, beyond proving that listeners are generically sensitive to *some* structural regularities and expectations. In particular, what seems to be missing is empirical evidence that is directly traceable to a specific form of representation, and to a computational- and algorithmic-level description of processing. Such a description would account for some well-specified class of phenomenal effects (as opposed to the generic surprise due to structural violations) and would specify hypotheses about how such effects of experience may emerge from plausible parsing computations, as well as what representations may be involved in those computations. This top-down methodological approach is more common in other domains, such as vision (Marr, 1982) or language (Chomsky, 1957; Griffiths et al., 2010; Jackendoff, 2002a).

For example, in the language domain, linguists propose explicit hypotheses about the shape linguistic representations may take, in the form of grammars, and test the reality of these representations by means of, e.g., structural priming paradigms (Branigan et al., 1995). The persistent effect of the activation of one specific representation is strong evidence that some encoding of that information, abstracted from the sensory stimulus that primed it, is available to the speakers. Based on such hypotheses about the shape of representations, linguists can formulate hypotheses about how processing can produce those representations under the constraints that come from the nature of the sensory inputs and of the cognitive architecture. For example, Gibson (1991) proposed a left-corner parsing model that executes specific operations at specific times while processing a sentence incrementally, word by word. Since each of these operations recruits cognitive resources, such a model comes with testable hypotheses on the time-course of the comprehender's performance, including processing difficulty (Gibson, 2000) and garden-path effects (Gibson, 1991). Psycholinguistic models, for example, differ depending on whether they assume a *serial* parsing algorithm (the parser only produces one representation at a time) or a *parallel* one (the parser produces multiple competing representations). Empirical research then looks for observable behavioural and neural effects predicted by either model (Gibson and Pearlmutter, 2000; Lewis, 2000).

In music, on the contrary, empirical results tend to be agnostic with respect to the particular form or even the very relevance of hidden representations of structural relations, let alone specific parsing mechanisms, so that they can neither prove nor falsify computational-level models that assume their existence. Importantly, by not committing to specific representational and processing hypotheses, or by relying on expectancy-based frameworks, the available empirical evidence is not *in contrast* with any specific model of underlying structural-processing mechanisms, just agnostic to it. In this thesis, we move some initial steps towards a top-down approach to music processing that strongly commits to the reality of structural representations, with the goal of turning this theoretical commitment into empirically testable hypotheses of the kind we see discussed in psycholinguistics for the language domain. In the next chapter, we outline a general approach for making prediction about computationally specified processing mechanisms that are compatible with the definition of structural hearing given in Chapter 1. In particular, we conceptualise the process of inferring structural representations as a form of grammar-based incremental parsing, and we show how this perspective captures several aspects of the phenomenology of (structural) listening.

3 Syntax and the phenomenology of listening

In Chapter 1, we have outlined a computational-level characterization of structural hearing, the component of the listening experience that (putatively) reflects music-theoretical structural relations. We suggested that listeners infer (internally-explanatory) representations of such structural relations upon exposure to the musical surface. Examples in Section 1.4, and the overview of empirical literature in Chapter 2, indicate that the cognitive relevance of specific structural representations, and of the processes that may result in their emergence, is elusive of direct empirical investigation. As a contribution to bridging this gap between music-theoretical introspection and empirical research, we now attempt to make the computational-level understanding proposed in Chapter 1 more concrete. In particular, we aim towards an algorithmic characterization of the processes that implement the inference as it (putatively) occurs during listening. Such an algorithmic account needs to be consistent with multiple kinds of constraints: (1) the proposed algorithm should be a viable implementation of the underlying computational-level mapping, i.e., the representations it outputs should be consistent with the music-theoretically predicted ones; (2) the time-course of processing should reflect the constraints that come from the temporal nature of the input, which is presented incrementally, and from its immanent ambiguity; (3) while (part of) the operation of the processor may be implicit, there should be a correspondence between the operation of the processor and the aspects of phenomenal experience that are introspectively associated with structural hearing.

In this Chapter, we explore how grammar-based incremental parsing is a viable algorithmic characterization of inference that has the potential to satisfy these three sources of constraints. In particular, we will discuss how (idiom-specific) grammars can fulfil the role of (internally-explanatory) generative models of the musical surface, before exemplifying the operation of a cognitively-plausible parsing algorithm for tonal harmony inspired by psycholinguistic accounts. Finally, we will show how specific phenomena of the listening experience, as identified in the phenomenological literature in music theory, may reflect features of the parsing process and of the representations it constructs incrementally. In the remainder of the thesis, we will then provide empirical evidence pinpointing some cornerstones of such an algorithmic theory of parsing: namely, the relevance in the musical domain of notions of structural representation (Chapters 4, 5), syntactic priming (Chapter 5), syntactic categories (Chapter 6), online incremental parsing (Chapter 7), and retrospective reanalysis (Chapter 8).

3.1 Music and language: a methodological analogy

In language, as well as other domains such as narrative (A. J. Cohen, 2002; N. Cohn, 2020) and action (Novembre and Keller, 2011), users face the problem of finding the most plausible explanation for sensory inputs that are presented sequentially over time. Both in music and in language, we have expert introspective accounts of a *syntax*, i.e., what structural relations may be attributed to instances of entities as they appear in a sequence of symbols. The sequential organisation of words in sentences is *governed* by syntactic principles. This does not necessarily entail that users *cannot* violate such principles. Rather, the existence of syntactic principles that are shared between speakers and comprehenders reflects the fact that the capacity of the signal to have the desired effect (in the case of language, the communication of semantic meaning) largely relies on the comprehender's capacity to attribute syntactic relations to words. The sharedness of syntactic principles provides a viable strategy for ensuring successful communication given the constraints of the human cognitive system (Christiansen and Chater, 2008).

These insights are formalised through grammars, which model the mapping between strings of symbols (in language, sentences as sequences of words) and syntactic interpretations. Grammars (Chomsky, 1957) and psychologically plausible models of incremental parsing (e.g., Frazier and Fodor, 1978; Gibson, 1991) constitute computational- and algorithmic-level characterisations of the phenomenon of language comprehension. The core question in (psycho)linguistics is, then, which grammar better captures the language user's interpretation of sentences, and how do language users implement the mapping from sentences to interpretations in real time.

Based on the discussion in Chapter 1, the phenomenon of structural hearing seems to emerge from an analogous computational task as language comprehension (Asano and Boeckx, 2015; Cecchetti et al., 2020; Jackendoff, 2009; Katz and Pesetsky, 2011): inferring structural interpretations from sequences of symbols that are presented incrementally over time. The computations underlying this process likely include the integration of newly encountered events into a pre-existing representation, as well as the storage of incomplete representations while new events are awaited to fulfil expectations. Differently from music research, though, psycholinguistics has a long track record in directly probing computationally well-specified phenomena such as processing complexity (e.g., in terms of dependency locality; Gibson, 1998, syntactic priming (Branigan and Pickering, 2017), and garden-path effects (Frazier, 1987). In this work, we draw inspiration from the methodologies adopted in psycholinguistics to investigate the very same cognitive computation in the domain of language, as outlined in the following.

Music and language have been compared on multiple levels as cognitive domains (Asano and

Boeckx, 2015; Jackendoff, 2009; Katz and Pesetsky, 2011; Patel, 2003). As discussed in Section 2.6, it has even been suggested that aspects of processing may be shared between the two domains (Patel, 2010). In the following, we remain agnostic with respect to the issue of the sharedness of cognitive functionality and processing resources, but we do rely on the analogy between musical and linguistic structure from a methodological perspective. Specifically, we approach the modelling of musical syntactic processing as sharing some analogy at the computational level of description to its linguistic counterpart, and we approach the empirical enquiry in a similar vein as psycholinguists approach language.

The theoretical underpinnings of this approach are not new: they are rooted in the research program initiated by the first grammar-based models of musical structure (Baroni et al., 1983; Keiler, 1978; Lerdahl and Jackendoff, 1983a; Steedman, 1984). Jackendoff (1991), in particular, outlines the requirements of a theory of music perception. Such a theory should minimally account for

1. the kind of structural intetrpretations that listeners may *in principle* attribute to surfaces,
2. the (possibly probabilistic) principles that influence the attribution of an interpretation to a surface,
3. an algorithmic-level description of incremental parsing, and
4. the identification of specific cognitive capacities and neural resources that are involved.

Generative models of musical structure specifically address points (1) and, in different forms, (2). For example, the Generative Theory of Tonal Music (GTTM; Lerdahl and Jackendoff, 1983a) specifies *well-formedness* criteria that determine what structures are possible in principle, as well as probabilistic criteria that determine the *preference* of an interpretation over another where multiple alternatives are possible in principle. In a similar vein, but without hard-coding preference rules, probabilistic implementations of generative grammars explicitly model the probability distribution over all derivations for a given surface, and allow for the automated learning of such probabilities (Finkensiep, 2023; Harasim, 2020). The probabilistic nature of such models is crucial, since it allows for modelling ambiguity.

These models are meant to capture the *competence* of listeners, and they only provide broad computational-level constraints for cognitively plausible processes that may implement parsing in real time. Computationally viable parsers for individual generative models have been proposed (Finkensiep, 2023; Granroth-Wilding and Steedman, 2014; Hamanaka et al., 2006; Harasim et al., 2018), but they have not been explicitly put forward as algorithmic-level models of cognitive processing. In particular, implications towards observable processing effects that are specific to individual parsing models have not been tested.

From a purely theoretical perspective, accounts of processing addressing point (3) have been proposed by David Lewin (1986) and Ray Jackendoff (1991), who independently discussed in great detail how the experience of musical structure may be shaped on a moment-by-moment basis during listening. Although both approaches are presented in the form of quasi-formal computational models of real-time processing, and they both aim to capture common phenomena, they start from somewhat different perspectives: the former is rooted in Husserlian phenomenology, while the latter in (psycho)linguistics (cf. also Lerdahl, 2014). Both authors unpack the incremental construction of a structural interpretation as it unfolds over time and, in both cases, the theoretical proposal is formulated at a very abstract computational level: the assignment of structural representations to incremental segments of the surface is treated as a "black-box" mapping, with minimal algorithmic characterisation. As a consequence, the involvement of specific cognitive functions and the time-course of their deployment remains unspecified. Even more importantly, such theoretical proposals have remained outside of the purview of empirical research: neither candidate operations that implement the integration of new events into preexisting partial representations, nor mechanisms to resolve ambiguity, nor the very existence of the kind of structural representations putatively produced by the parser have been tested. More generally, despite the influential tradition of acknowledging computational-level analogies between linguistic and musical processing on a theoretical basis, there are few empirical approaches that commit to language-inspired representational and processing models for music. In the following, we discuss how a grammar-based model of parsing, in the spirit of Jackendoff's account, may capture many aspects of the phenomenal experience of structural hearing as it is characterised in accounts of music-theoretically inspired phenomenology, such as Lewin's.

3.2 Syntax and grammars

The examples in Section 1.3.3 illustrate how surface entities can be modelled as being *generated by* latent entities, i.e., as the observable expression of those entities. Generally speaking, a generative model of musical structure should specify operations for transforming entities into other entities, including transformations for transforming latent entities into observable surface entities. These transformations would then formalise the possible relations that entities may entertain with one another: in some sense, knowing how a surface is generated is "the same as" knowing one of its interpretations.

Formal grammars naturally model the attribution of internally explanatory, compositional, and generalisable interpretations to musical surfaces, when the latter are understood as temporal sequences of events (each event being an instance of an observable entity). They do so by modelling the surface "as if" it was the final state of a process which iteratively applies transformations, or *production rules*, to entities. Each production rule introduces new instances of entities ("children") as subordinate to previously existing ones ("parents"), until eventually all observed entities are generated. Crucially, production rules formalise the relatedness between entities: each rule application is associated with relations being established among the children as well as between children and parents of that rule application. The sequence of rule applications that generate a specific surface is a *derivation* of that

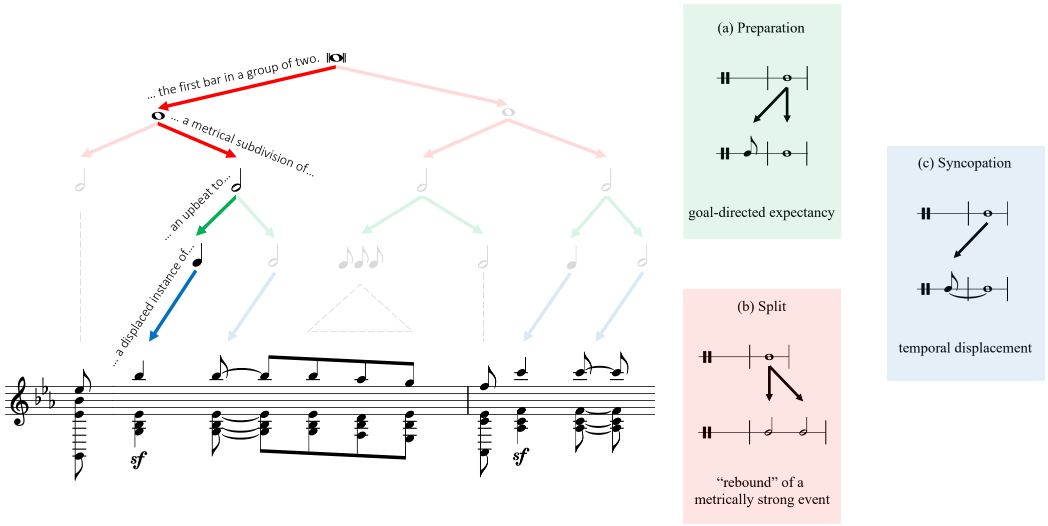

... the first bar in a group of two.
... a metrical subdivision of...
... an upbeat to...
... a displaced instance of...
sf
sf
(a) Preparation
goal-directed expectancy
(b) Split
“rebound” of a metrically strong event
(c) Syncopation
temporal displacement

Figure 3.1 – Derivation of the rhythmic surface of L. van Beethoven’s String Quartet op. 18 n. 4, i, mm. 60f (left, adapted from Cecchetti, Tomasini, et al., 2023), based on the three types of production rules (right) comprising the generative grammar proposed by Rohrmeier (2020a). At each node in the tree, one of the thee rule types is applied (as reflected by the colour), and the surface is generated by applying the production rules recursively. Since each rule is associated with a structural relation (e.g., (a) being preparatory, (b) being a metrical subdivision or rebound, or (c) being displaced) the derivation can be read "bottom-up" as an interpretation of the surface: for example, the second event is interpreted as "*a displaced instance of an upbeat to a metrical subdivision of the first bar in a group of two*".

surface.[1] Since each rule application within a derivation establishes relationships between the entities that are involved, it is possible to map a derivation into a network of such relationships. In other words, given a derivation of a surface, a structural identity can be attributed to each event in the surface in terms of the relationships it entertains with other entities. The derivation as a whole, then, corresponds to an interpretation of the surface.

Figure 3.1 exemplifies the derivation of a rhythmic surface under a generative grammar for tonal rhythm (Rohrmeier, 2020a). Three different types of generative rules, reflected in different colours, formalise three different types of structural relations. For example, if two events are generated through the application of a "split" rule (red), they inherit a structural relationship: the event corresponding to the right child is a *metrical subdivision of* the event corresponding to the left child. The structural identity of each event is then encoded in the derivation of the surface, as we can see by "reading" the derivation from the leaves up: the second event in the surface, for example, is interpreted here as a displaced instance of an upbeat to a metrical subdivision of the first bar of the excerpt.

By fixing the class of possible production rules, a grammar determines (1) a language, i.e., what surfaces can be generated (*weak* generative capacity), and, for those that can, (2) a particular

[1] It is worth reiterating that the sequence of rule applications that generates a given surface is not meant to reflect a generative process that occurs in reality, such as the compositional process of a piece as reflected, e.g., in a composer’s sketches: it is just a way to formalise a pairing between surfaces and interpretations.

mapping between each surface and its possible derivations (*strong* generative capacity). Thanks to its strong generative capacity, the grammar assigns derivations, represented, e.g., as tree graphs, to surfaces. Each rule invoked by a derivation implies the establishment of structural relations between entities, so that a derivation implies an interpretation of the musical surface. Therefore, grammars are widely used to formalise internally-explanatory interpretations. Since the grammar relies on the same set of production rules to assign derivations to all the surfaces in its language, these interpretations are also generalisable in the sense of Section 1.2.3: the grammar captures the structuring principles that are common to an entire repertoire.

Under the Bayesian-listener hypothesis, when exposed to a repertoire, listeners implicitly acquire a grammar that allows them to infer explanations of the observed surfaces by attributing structural interpretations to each surface. The result of this inference is a representation of the derivation of the surface. Among all grammars that generate the same language, the one particular grammar whose rules can be directly associated with the structural relationships that are explanatory towards the surfaces in the corresponding idiom is the *competence* grammar for that idiom. The goal of a grammar-based music-theoretical model is to faithfully capture the competence grammar for that idiom (Lerdahl and Jackendoff, 1983a; Rohrmeier, 2020b). For instance, the rhythm grammar exemplified in Figure 3.1 (Rohrmeier, 2020a) is a candidate competence grammar for tonal rhythm, since its rules are directly matched with the interpretive notions of *preparation*, *split*, and *syncopation* – which, in turn, are proposed as the fundamental kinds of structural relations that "explain" the temporal arrangement of surface events.

Overall, grammars represent powerful tools allowing theorists to express their beliefs about what interpretations can be heard in a given style. However, formalising music-theoretical frameworks as we just discussed is not sufficient to model structural hearing as a cognitive phenomenon. Minimally, the following aspects also need to be accounted for.

Probabilistic inference. First, while a grammar tells us *that* a given surface affords *some* interpretations in a given style, and identifies the set of these interpretations, this does not really correspond to the kind of statements theorists typically make. In particular, when more than one interpretation is possible for a given surface, theorists often consider one more plausible than the others, as we discussed in the case of Mozart's theme (Figure 1.9). Furthermore, some surfaces may sound less plausible than others as expressions of a given style, in the sense that even though it is possible to find some interpretation for those surfaces, these interpretations are, in some sense, collectively implausible. In order to capture this kind of analytical statements, a generative model should be complemented with a probabilistic implementation (e.g., as a probabilistic grammar; Manning and Schütze, 1999; cf. Abdallah et al., 2016 for a review in the musical case).

In the influential GTTM, such probabilistic constraints are specified as *preference rules* that are hard-coded in the theory. In principle, this approach allows the theory to make predictions about what interpretation listeners would favour given a certain surface. However, specifying preference rules at the theory level prevents the theory to account (even in principle) for individual differences in hearing and, more importantly, for the acquisition of statistical preferences through implicit learning. While this exceeds the declared scope of the GTTM, which solely intends to model the *ideal* tonal listener, a cognitive theory of structural hearing – even for a restricted idiom – may aspire to capture stylistic enculturation more flexibly.

Furthermore, it should be noted that the GTTM is not formalised as a grammar: in particular, preference rules are typically expressed as bottom-up criteria (e.g., given that a note has a longer duration, prefer assigning strong metrical weight to it). From a strictly generative perspective, at least some such constraints (as characteristic of a style) may rather be expressed as top-down constraints on the generative process, as opposed to bottom-up constraints on perception. When perception is understood as "inverse generation", i.e., as the inference of a generative derivation, the perceived structure would still reflect the probabilistic preferences that constrain the generative process. As a simple example, each rule of a grammar may be associated with a probability, so that the plausibility of an interpretation of a given surface is the product of the probabilities of the rule applications that comprise that derivation (*tree probability*), whereas the plausibility of a surface as expression of the idiom is the sum of the tree probabilities of all its derivations (*string probability*). Such a probabilistic model, e.g., as implemented in Harasim et al. (2018) for the musical case, would then capture, at the computational level of description, the learning and the inference component of a Bayesian listener (Chater and Manning, 2006).

However, note that when a human analyst uses the grammar as a tool to express their own interpretation, the probabilistic aspects of the inference are implicitly reflected in the choices of the analyst. In practical applications, then, grammar-based models of musical structure can be used without a full computational implementation of the probabilistic model as long as the grammar is intended as a formal notational language for a human analyst to express the outcome of their inference. In this sense, the value of a given grammar formalism lies in its capacity of allowing a human analyst to faithfully express their interpretation, rather than in its capacity to *predict* or *simulate* the analyst's interpretation.

Incremental parsing. Second, the pairing of surface and interpretation, as formalised through a grammar, abstracts away the temporal nature of listening from the perspective of the listener. A grammar specifies possible *final* interpretations for a given surface *as a whole*, provided that the surface belongs to the grammar's language. However, listeners are exposed to the surface incrementally, so that at every moment in time they have only access to a segment of the musical surface which is not guaranteed to be a full idiomatic surface itself. In order to account for what happens during listening, and in particular what representations are produced at every moment in time while the surface is presented incrementally, a model of processing should complement the grammar formalism. Such a model would characterise how the grammar's rules are exploited to incrementally construct a derivation for the entire surface, what intermediate representations are formed in the process, and what is the time-course and the computational cost of the operations involved in the process.

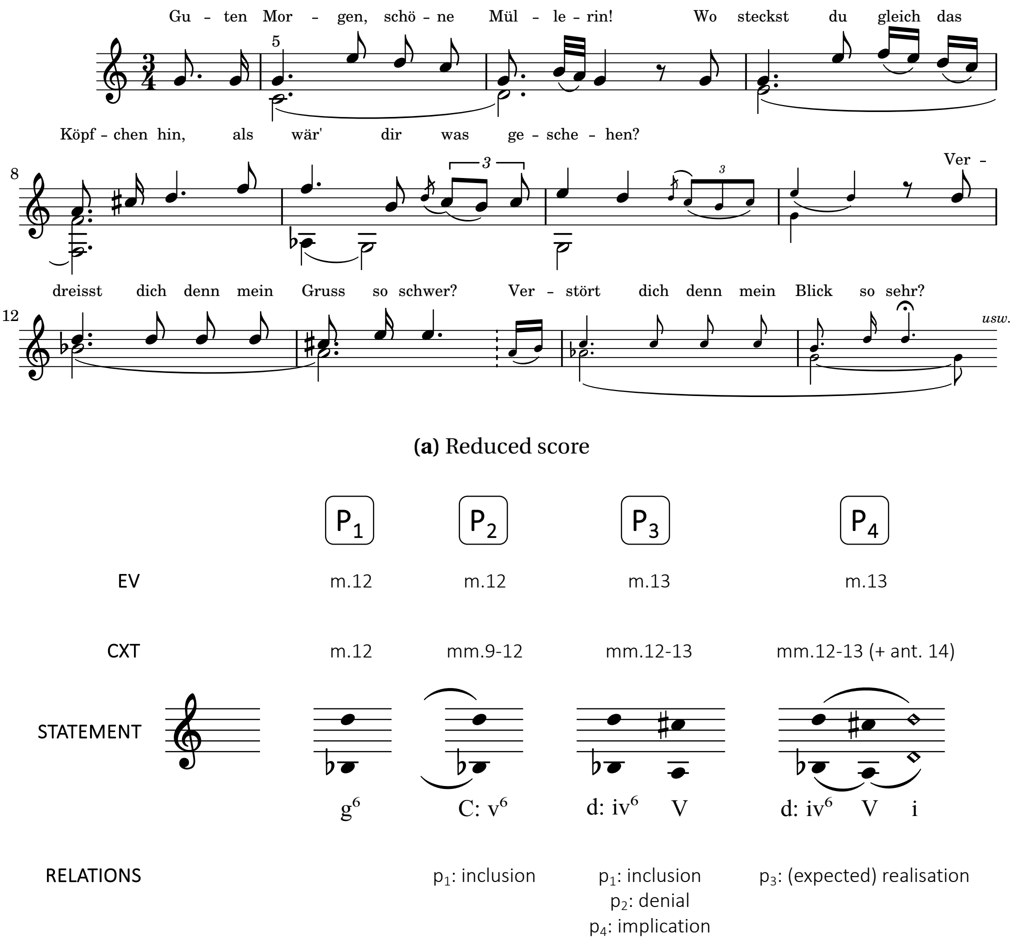
Processing limitations. Finally, by adopting a music-theoretically inspired competence grammar as the core of a formal model of listening, we are accepting some degree of approximations. A grammar or an individual derivation formalising music-theoretical insights reflects the result of an inference that is largely achieved offline as part of a discovery process driven by rational thought and explicit knowledge. However, (1) listeners may acquire implicit knowledge of a grammar that is in some respects different from the music-theoretically motivated one, even if the latter happened to be optimal in terms of generalisation and parsimony with respect to the given repertoire, and (2) the actual inference process, as implemented in real time, may be limited by constraints of cognitive architecture that lead to sub-optimal inference or prevent the inference of a complete derivation.

In summary, a characterisation of structural hearing (for a given repertoire) would require to specify a competence grammar – accounting for the space of possible representations – together with a model of processing that accounts for the constraints imposed to the inference process by the temporality of real-time listening and its cognitive implementation. In the following, we overview how these components can concur to reflect many aspects of the phenomenology of listening.

3.3 Modelling interpretation as parsing

Sofar, we have characterised structural hearing as a form of inference (Section 1.3) and we have suggested that the such inference may be performed by inverting the generative derivation of each surface under a competence grammar for a given repertoire (3.2). Such inverse generation should occur probabilistically and incrementally during listening, in order to account for the phenomena that listeners can introspectively recognise from their own experience. For example, Lewin (1986) describes a parser that (1) traverses the surface event by event (EV), (2) for every event selects one or more contexts (CXT), (3) for each context, conjures a percept in the form of an analytical statement as well as (4) establishes relations with other percepts. As exemplified in Figure 3.2, upon listening to m. 12 of F. Schubert's *Morgengruß*, listeners would first perceive the chord in isolation in its uninterpreted sensory nature of g^6. This percept P_1 may be included in a more elaborate percept P_2 where the g^6 chord is interpreted as an unusual minor dominant in C major, but the following chord in m. 13 disproves this hearing. At this point, a new percept P_3 may emerge, where P_2 is denied in favour of a different interpretation where the g^6 chords takes the role of a pre-dominant in d minor, implying percept P_4 as a resolution.

Lewin's account outlines the skeleton of a parsing process for music. However, despite being presented in the form of a quasi-formal computational model, it remains underspecified with


Gu - ten Mor - gen, schö - ne Mül - le - rin! Wo steckst du gleich das
5
Köpf - chen hin, als wär' dir was ge - sche - hen?
8
3
3
Ver -
dreisst dich denn mein Gruss so schwer? Ver - stört dich denn mein Blick so sehr?
12
usw.
P1
P2
P3
P4
EV
m.12
m.12
m.13
m.13
CXT
m.12
mm.9-12
mm.12-13
mm.12-13 (+ ant. 14)
STATEMENT
g6
C: v6
d: iv6 V
d: iv6 V i
RELATIONS
p1: inclusion
p1: inclusion
p2: denial
p4: implication
p3: (expected) realisation

(a) Reduced score

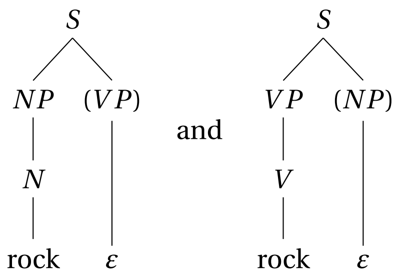
(b) Four percepts P_1-P_4 after Lewin (1986)

Figure 3.2 – F. Schubert, *Morgengruß* mm. 5ff. Each percept is produced at a certain moment in time (EV) by considering some surrounding context (CXT). The phenomenal content of percepts is expressed in the form of analytical statements, as well as in terms of relations among percepts.

respect to the nature of the statements comprising percepts (cf. Lerdahl, 2014). An alternative account of online parsing based on the GTTM is the object of Jackendoff's (1991) article: here, the nature of the inferred representations at every moment in time is constrained by the GTTM well-formedness and preference rules. In Jackendoff's account, the GTTM preference rules are applied to incremental segments of the surface as new events are encountered, in order to determine which well-formed structures are possible for that segment. Since, in principle, multiple structures are admissible at every moment in time, a *selection function* is invoked that singles-out the most likely interpretation, which manifests itself in the listener's conscious experience.

Both Jackendoff's and Lewin's theoretical proposals are formulated at a very abstract computational level: the assignment of structural representations to incremental segments of the surface is treated as a "black-box" mapping, with minimal algorithmic characterisation. As a consequence, the role of specific computations and the time-course of their deployment remain unspecified. More recently, computationally viable parsing algorithms for individual generative models have been proposed (Finkensiep, 2023; Foscarin et al., 2023; Granroth-Wilding and Steedman, 2014; Hamanaka et al., 2006; Harasim et al., 2018), but they have not been explicitly put forward as algorithmic-level models of cognitive processing. In particular, implications towards observable processing effects that are specific to individual parsing models have not been tested. In contrast, cognitively plausible models of online parsing in language fully specify the time-course of computations that are deployed when comprehenders are exposed to sentences. For every new word that is encountered, these computations produce and manipulate partial derivations until, eventually, a derivation for the entire sentence is constructed. This is achieved by reverse-engineering the generative process as modelled through the competence grammar for the particular idiom: this requires the processor to have access to the grammar's rules.[2]

To give a concrete example, parsing operations in Gibson's (1991) left-corner parsing model include *node projection*, *pushing* to a memory stack, and the *attachment* of two partial derivations. Node projection models the experience of a single word as the expression of a given syntactic category: it exploits the grammar's rules to determine possible ways in which that word may be embedded in some syntactic structure.[3] For example, upon encountering the word *rock*, its node projections include



(3.1)

corresponding to its interpretations as a noun or a verb, respectively. Note how the syntactic category is reflected in relating the observed word *rock* with unobserved nodes that are hypothesised to be filled in by words to be encountered in the future.

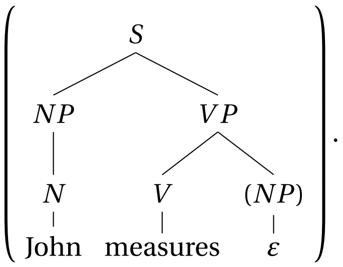
To illustrate the operation of the parser, let us consider what happens at the time the word *the* in the sentence

John measures the rock's weight. (3.2)

[2] In principle, it is not a logical requirement for the processor to operate precisely under the competence grammar (Berwick and Weinberg, 1984): the processor may rely on an alternative grammar for online parsing, and map the resulting representation to the one implied by the competence grammar. However, it is unlikely that the processor's grammar is completely unrelated to the competence grammar: the "strong competence hypothesis" (Savitch et al., 1987) that assumes the identity of the processor grammar with the competence grammar is then a reasonable and parsimonious assumption to make as a starting point for investigation (Steedman, 2000).

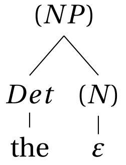
[3] In Gibson's (1991) formalism, (maximal) node projections are defined in terms of $\bar{X}$-theory, which we omit for the purpose of this informal presentation.

is encountered. After parsing the previous words, *John measures...*, the memory stack contains the partial derivation



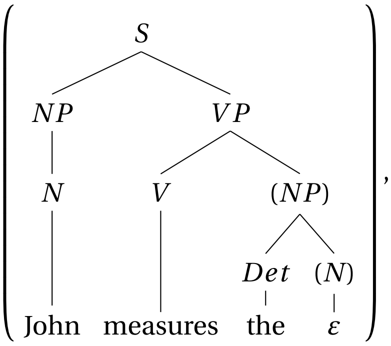
(3.3)

Upon encountering the word *the*, the *node projection*



(3.4)

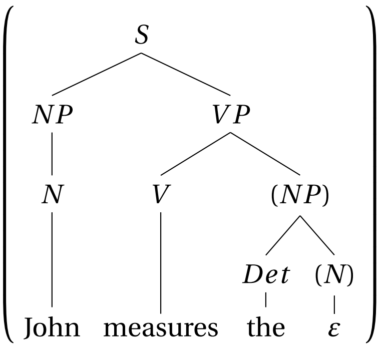
is computed, and the processor assesses whether the projection 3.4 can be attached to the partial derivation 3.3. One possibility is to identify the hypothesised (NP) node in 3.4 with the analogous node in 3.3: attaching the two partial derivations 3.4 and 3.3 on this common node yields one possible partial derivation for *John measures the...*, namely



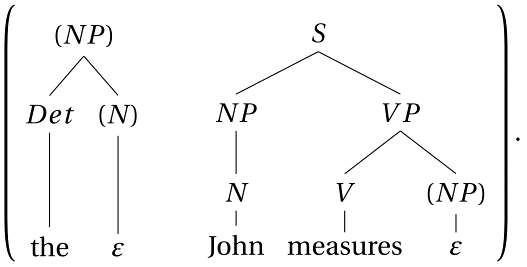
(3.5)

which is pushed to the stack awaiting a noun to fill the hypothesised node (N) as the object of the verb *measures*.

Another possibility, though, is that *the* stands for the beginning of a genitive noun-phrase: in this case, the (NP) node in 3.4 is not to be identified with the (NP) node in 3.3, as the latter is related to the verb *measures* as its object. Since, in this case, attachment is not possible, the parser has to *push* the projection 3.4 to the stack. In summary, after reading *the*, there are two alternative states of the stack, one containing a single partial derivation and the other storing two separate partial derivations that cannot be attached yet:

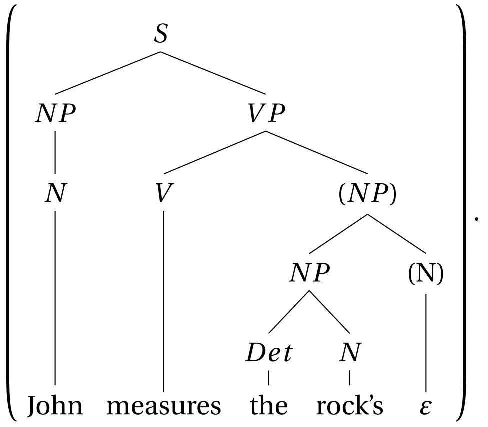


(3.6a)



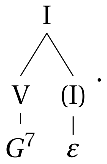
(3.6b)

The twofold state of the stack (3.6b and 3.6a) reflects the ambiguity of this segment of the sentence, which the parser has no way to disambiguate until the next word, *rock's*, clarifies that only the interpretation reflected in 3.6b is tenable, while 3.6a has to be abandoned. The resulting state of the stack, obtained by attaching *rock's* as the hypothesised (N) node in 3.6b (left) and attaching the resulting (NP) as a genitive-NP node in 3.6b (right)

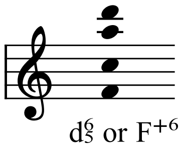


(3.7)

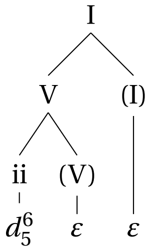
Somewhat speculatively, we could envisage an analogous parsing process for music, based on Rohrmeier's (2020b) grammar. Node projections may be taken as modelling harmonic functionality. For example, *hearing* a G^7 chord *as* expressing the functional category of a dominant may be modelled as the projection of a structure that includes a hypothesised tonic to come in the future



(3.8)

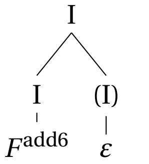
Similarly, upon hearing the chord 

, the projection



(3.9)

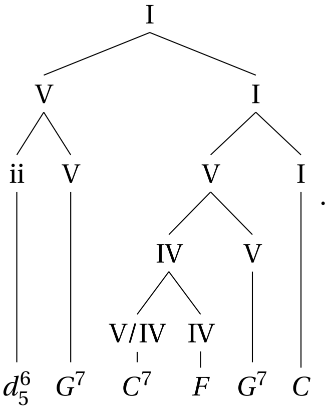
captures the hearing of the chord as a pre-dominant (d^6_5), whereas the projection



(3.10)

reflects a hearing where the same chord is understood as a stable added-sixth harmony (F^{add6}), commonly expressing tonic function in the Jazz idiom (H. Martin, 2023).

By analogy with the linguistic case, Figure 3.3 exemplifies the time-course of processing for a simple chord progression, whose complete derivation is



(3.11)

Similarly to the linguistic example, the insertion of a local tonicisation (with the applied domi-

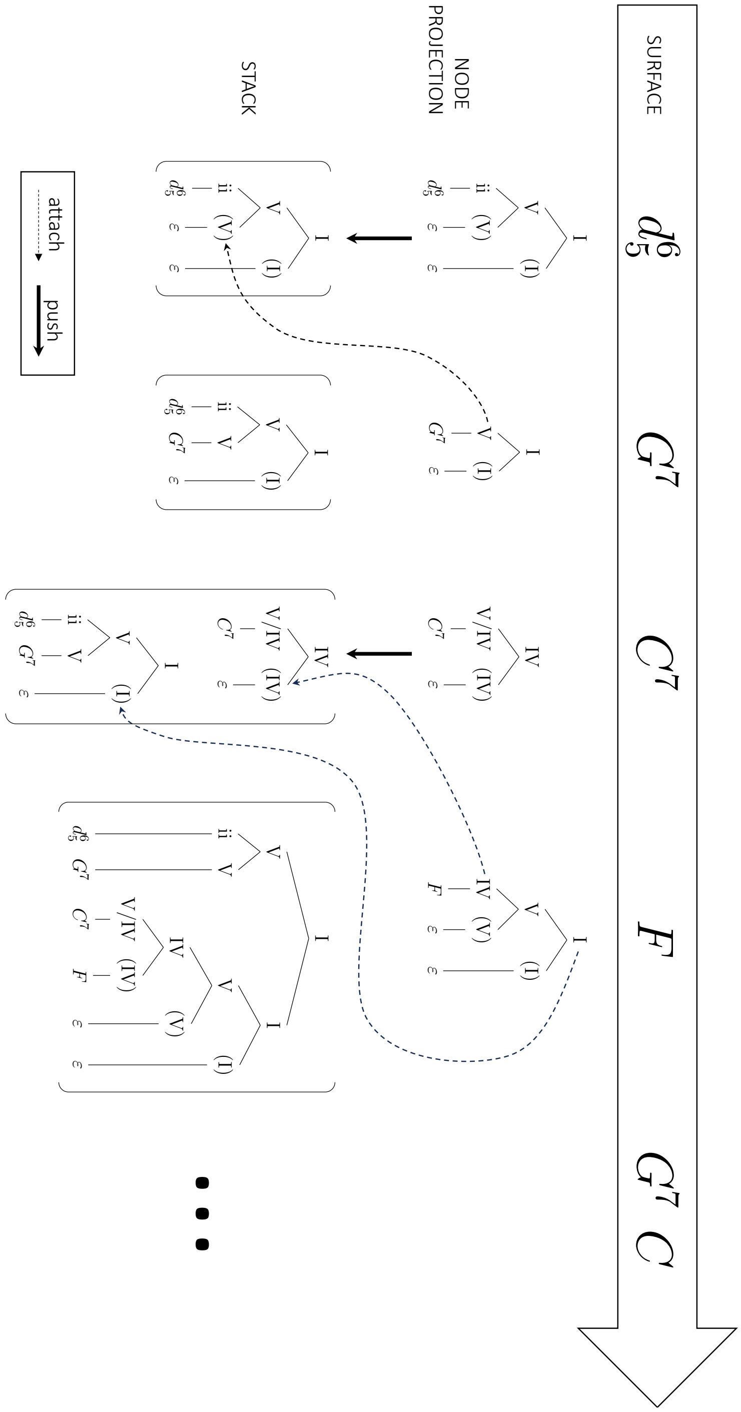


Figure 3.3 – Incremental parsing of the chord progression in Equation 3.11, after Gibson (1991). Under each chord in the surface, the corresponding node projecton as well as the state of the stack after parsing that chord are displayed. For every new chord that is encountered, the grammar rules (Rohrmeier, 2020b) are deployed to infer a node projection for the observed chord. The resulting projection is either attached to a partial derivation stored at the top of the stack (dashed arrows), or otherwise the projection itself is pushed to the stack (solid arrows).

nant C^7) results in the embedding of a subordinate phrase. In turn, this puts the completion of the pre-existing partial derivation on hold (by pushing it deeper into the stack).

We presented this toy model of parsing as an example inspired by a psycholinguistic approach: the cognitive plausibility of this specific parsing model has no empirical support. Crucially, though, so does the cognitive plausibility of any other parsing model, since no aspects of such an algorithmic-level account have been tested in music. We argue that grammar-based modelling makes it possible to – at least – formulate such models as hypothetical accounts of processing, bringing them within the purview of empirical research as it has been the case in linguistics. In other words, this modelling approach makes it possible to investigate specific questions about the computational and algorithmic nature of inferring structural representations.

3.4 Experience as representation

We think of the structural listener as inferring representations of interpretations during listening, in the form of derivations under the competence grammar for a given idiom. More precisely, for every newly encountered event, listeners minimally form a representation of a partial derivation (a *structural representation*) that accounts for all past events by providing an (internal) explanation for their presence. Such a derivation is "partial" in the sense that it does not account for the entire musical surface, just the portion that has already been parsed.

Crucially, the attribution of phenomenal character to events is influenced by the existence of retentions, protentions, and counterfactuals within a phenomenal state (Augustine, 2012; Hoerl, 2013; Husserl, 1964; Lewin, 1986; Moshaver, 2012). For example, *hearing* a note *as* a passing tone entails retaining in the current phenomenal state, i.e., as part of a currently active representation, information about an event in the past, which originates the passing motion. Vice versa, *hearing* a rhythmic event *as* an upbeat entails projecting the coming of another event in the future, and the experience of the upbeat as such is only determined by the relation between an event that has occurred and one that has not but is still represented in the current phenomenal state as a protention (or, in more standard music-psychological terminology, an anticipation; Huron, 2006). Similarly, the experience of an event as syncopated is determined by advancing the counterfactual hypothesis that this event is the manifestation of an event that “was supposed to be” somewhere else. In summary, in order to faithfully model the experience of listening, structural representations ought to encode the structural identity of past and putative future events, in terms of the relationships they entertain with other events as well as with latent entities that merely manifest themselves in the observed events, without being present in the surface.

Grammar-based derivations fulfil this modelling role in several ways. First, representations of derivations encode information about past events that is available in perception at the present moment in time, in the form of phenomenal attributes of events that are (now) *heard as* standing in some relation to events in the past. In this sense, the emergence of (partial) derivations may be seen as capturing the aspect of phenomenal experience that Lewin, after Husserl (Miller, 1984), refers to as *retention.*

Furthermore, the presence of some among the past events may only be explained by assuming the occurrence of specific events in the future: in this case, the derivation will also account for relations linking events in the past with events in the future (Figure 3.4a). Partial derivations, then, also encode present information about future events, in the form of phenomenal attributes of events that are (now) *heard as* standing in some relation to events in the future. The existence of such *open* relations captures the phenomenological notion of *protention* (Lewin, 1986; Miller, 1984), the music-theoretical notion of implication (Narmour, 1990), and, in more standard music-psychological terminology, anticipation and expectancy (Huron, 2006; Rohrmeier, 2013).

Finally, structural representations also encode the relationship of observed events with unobserved latent entities as well as possibly counterfactual hypotheses about how the musical surface *may* look like. This information is encoded in the hierarchical nature of derivation trees. For example, Lewin (1986) describes how the phenomenal experience of a $V - I$ progression, which is represented in the listener's mind as an anticipation at the time the dominant is presented, is still counterfactually "real" after the dominant is resolved (deceptively) to a submediant harmony instead. A derivation would capture the counterfactual reality of the expected resolution to the tonic by preserving the relatedness of the dominant to the tonic upon encountering the deceptive resolution (Rohrmeier and Neuwirth, 2015): the corresponding reduction is obtained by pruning the tree above the hierarchical level where the submediant is introduced in the derivation (Figure 3.4b). In this sense, the hierarchical nature of derivations reflects the music-theoretically predicted influence of counterfactuals on experience. TO consider a different example, a chord may appear in a derivation as the descendant of a harmonic entity, say, a hexatonic *Tonfeld*: the phenomenal character of the chord being *heard as* expression of a hexatonic sonority is reflected in the chord's ancestry in the derivation tree (Figure 3.4c). In other words, "pruning" a derivation tree at some hierarchical level yields a reduction of the musical surface: reductions constitute simpler templates that are never sounded, yet *hearing* the actual musical surface *as* a deformation of such templates determines a certain phenomenal character of experience. In Figure 1.12, for example, two different reductions, corresponding to derivations diverging at some hierarchical level, correspond to two different percepts. Similarly, Schenker (1987) describes how the experience of a chromatic passage is shaped by the listener having access to a representation of its diatonic reduction, which is (counterfactually) never observed ("It is astonishing how rapidly our perception functions, how it rushes with lightning speed through so many intervening stages and grasps the abbreviation", p. 149; cf. discussion in Dubiel, 1990).

Overall, the emergence of derivations models the condition of the structural listener that, at any moment in time, does not simply experience events under their surface identity, but rather *hears* events *as* having a certain structural identity. However, while some aspects of the listener's experience reflect information that is encoded in the representation itself, other as-

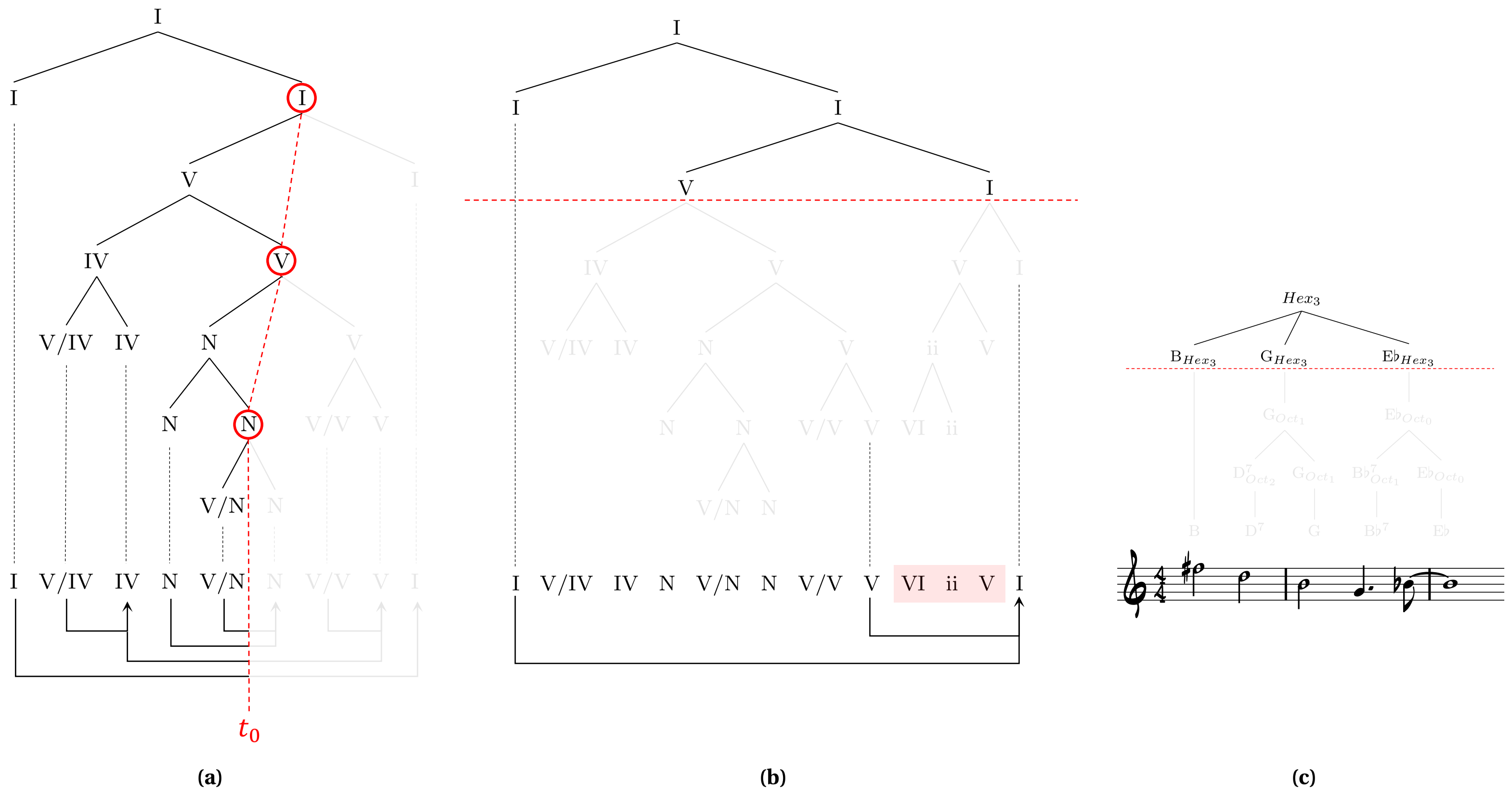


Figure 3.4 – Expectations and counterfactuals are encoded in syntactic trees. **(a)** The tree encodes structural relations among chords in a progression from F. Schubert, *Erlkönig* D328, mm. 131-148. The shaded arrows across time t_0 indicate that, at a given moment in time, past events entertain structural relations with events in the future: these relations are incomplete and await resolution, engendering expectancy towards three different harmonic goals that are encoded in the tree (red circles). **(b)** The final cadence from progression (a) is delayed by a deceptive resolution of the dominant on VI. The "counterfactual" structural connection between the dominant and the final tonic is still captured by the reduction obtained by pruning the tree along the dashed red line. **(c)** In the beginning of J. Coltrane's *Giant Steps*, chords B, G, and E♭ are locally tonics of functional progressions (Oct), yet they are to be understood as expressions of a Hexatonic sonority (Hex) as reflected in the the tree above the dashed red line (see Rohrmeier and Moss, 2021 for details).

pects may rather reflect effects of processing as byproducts of the very cognitive computations that lead to the emergence of representations.

3.5 Experience as processing

3.5.1 Updating expectations

The core functionality of a parser is to constantly update an active representation of structure. In order to integrate a newly encountered event, the preexisting derivation may require changes both with respect to the part that accounts for events in the past, as well as with respect to its projection towards putative future events. We will discuss the latter case first.

Updating expectations is reflected in changing the hypothesised nodes in the currently active partial derivation. The experience of surprise, deceptiveness and cadential evasion may be seen among the phenomenal byproducts of such a processing effect (Rohrmeier, 2013). In Figure 3.3, for example, the expected tonic resolution upon encountering the dominant G^7 is delayed by the insertion of a harmony (C^7) that (1) fails to resolve the dominant, and (2) engenders its own expectations (towards F) that require resolution (Rohrmeier and Neuwirth, 2015). On one hand, some aspects of the listener's experience are reflected in the representation that is formed once deception has occurred: for example, the experience of the sustained expectation towards the tonic promised by the cadential dominant, as well as of the embedded expectations engendered by the deceptive or evaded resolution, are encoded in the partial derivation. However, the experience of *surprisingness*, *deceptiveness*, or *evasion* itself does not reflect something that is encoded in the representation, and rather reflects the processing operation that *repairs* the mistaken expectation of a tonic right after the dominant, by embedding subordinate sub-structure upon encountering C^7. In this sense, the overall experience elicited by cadential evasion may be seen as the result of both features of the representations and features of the underlying processing.

3.5.2 Ambiguity and retrospective revision

If structural interpretation in music was deterministic, then updating expectations towards future events would be the only kind of possible manipulation: once parsed, events in the past would maintain their structural role. However, a core feature of musical structure is its being ambiguous, even within the constraints of a single musical idiom: multiple interpretations are typically possible for any given surface. This entails that a cognitively plausible processor for musical syntax should complement an incremental parser with mechanisms to deal with ambiguity. This may include provisions for handling multiple parses in parallel (such as the multiple stacks in Gibson's model, cf. Equation 3.6), as well as an *oracle* component (Steedman, 2000) that decides heuristically which of the multiple possible representations to prune and which to prefer.

Psycholinguistic models differ with respect to the contributions of these components to the parsing process (Gibson and Pearlmutter, 2000; Lewis, 2000). In *parallel* models of parsing, all possible derivations are computed simultaneously and are only abandoned when they become untenable. In particular, in *ranked* parallel models, the oracle still identifies one derivation as more plausible, corresponding to the interpretation that is actually experienced during listening. Vice versa, in *serial* models of parsing, the oracle commits to a single preferred derivation at all times.

In both ranked-parallel and serial models of parsing, one derivation is salient to conscious experience (although effects associated with suppressed or dispreferred derivations may still be observed; Gibson and Pearlmutter, 2000; Hickok, 1993). Crucially, though, such parsers are prone to error, as they rely on the heuristic operation of the oracle. For instance, when encountering "the song" in the sentence

When the band played the song pleased all costumers, (3.12)

the oracle might mistakenly favour a reading whereby "the song" is the object of the verb "played". Upon encountering the verb "pleased", though, the previously preferred interpretation has to be abandoned in favour of a reading whereby "the song" is the subject of the verb "pleased". This so-called *garden-path* effect amounts to updating *retrospectively* the interpretation of events that have happened in the past, either due to a change in the ranking of alternative representations (in a ranked-parallel parser) or the formation of an entirely new representation (in a serial parser).

In the musical domain, garden-path effects correspond to a phenomenology of listening that is not one-directional, being only driven by updating expectations towards future events, but both prospective and retrospective, being also driven by updating the memory of past events. In the theoretical discourse, this aspect of the listening experience is widely acknowledged: in Lewin (1986) this is captured by the retrospective "denial" of a percept (cf. Figure 3.2, in Caplin (1998), Schmalfeldt (2017) and N. J. Martin and Vande Moortele (2014) it is reflected in the notion of "becoming" (e.g., of an authentic cadence in the key of *V* reinterpreted retrospectively as a half cadence in the tonic key), while Jackendoff (1991) explicitly argues in favour of a ranked-parallel parser for musical structure prone to *retrospective reanalysis* (cf. also Rohrmeier, 2013; Temperley, 2001a). However, differently from their linguistic counterpart (Gibson and Pearlmutter, 2000; Lewis, 2000), the existence of musical garden-paths has not been investigated empirically yet: this will be the object of Chapter 8.

3.5.3 Processing complexity and tension

Another effect of processing that may contribute to phenomenal experience is the complexity of parsing computations. Depending on the nature of the representations to be constructed, or the manipulations to be performed, the required algorithmic operations may have higher or lower computational cost. In turn, computational cost is reflected in increased demands on cognitive and neural resources. In the psycholinguistic literature, factors influencing processing complexity at any given moment in time during parsing include

(1) the depth of the memory stack, with higher stack depth being associated with higher working-memory demands (Lewis, 1996; Shain et al., 2022);

(2) the number of hypothesised nodes in the currently active (partial) derivation, corresponding to representations of expected events in the future to be maintained in working memory until they are finally encountered (Gibson, 2000; Grodner et al., 2002);

(3) the execution of specific computations on (partial) derivations, such as attachment or pushing to stack (Gibson, 1991);

(4) the non-locality of dependencies (namely, the sequential distance between events that are linked by a structural relation) which makes structural integration more difficult to perform (Caplan and Waters, 1999; Gibson, 1998);

(5) similarity-based interference in memory retrieval, due to the recursive embedding of similar structures (Jäger et al., 2017; Lewis, 1996);

(6) the contextual frequency of occurrence for events or (sub)structures, which may facilitate the retrieval of representations from memory (Lewis and Vasishth, 2005);

(7) the surprisal of newly-encountered events, conditional on the currently active representation (Hale, 2001);

(8) the coexistence of multiple competing interpretations in ambiguous inputs (Bornkessel et al., 2004);

(9) the necessity to abandon or revise the currently active representation (e.g., in garden-path effects; Ferreira et al., 2001; Ferreira and Henderson, 1998).

The computational cost of parsing is modulated continuously as a result of the above factors, the effect of which may be observed in behaviour and brain activity (Gwilliams, 2020; Tanenhaus and Trueswell, 1995). Importantly, processing cost may exceed the available cognitive resources, resulting in the failure to conjure a unified representation of the entire input. This kind of processing limitations may be responsible for (part of) the discrepancy between the representations postulated by the competence grammar, on one hand, and the observed performance of the parser in conjuring representations that are actually experienced. A model of parsing with precise implications towards estimating processing complexity, together with estimated thresholds of processing breakdown, would then capture both the top-down understanding of perception as inferred representation, as well as the bottom-up characterisation of perception as observed (cf. Gibson, 1991).

In the absence of cognitively-plausible parsing models for music, processing complexity in music is mainly discussed in terms of the complexity of representations (e.g., the presence of embedded modulations; Berent and Perfetti, 1993; Ma et al., 2018a), rather than of specific parsing computations. Nevertheless, processing complexity may underlie important aspects of the listening experience. First, minimisation of processing complexity may be among the criteria that shape the oracle's preference for one among the plausible interpretations of an ambiguous input (Gibson, 1991; Grodner et al., 2002): in this sense, processing complexity contributes to determining which representation is actually experienced at every moment in time.

Furthermore, processing complexity may underlie the experience of *tension*, a generic feeling of sustained, unresolved urgency with a strong emotional connotation that is considered crucial to the experience of music (Farbood, 2012; Krumhansl, 2002; Lehne and Koelsch, 2015; Lerdahl, 2014). Syntactic structure has been proposed among the contributing factors to the build-up of musical tension (Bigand and Parncutt, 1999; Bigand et al., 1996; Farbood, 2012; Lerdahl, 2014), alongside multiple other musical dimensions (Farbood, 2012; Farbood and Price, 2017; Pressnitzer et al., 2000). In particular, modelling the time-course of tension-relaxation dynamics is the central goal of GTTM's prolongational structure (Lerdahl, 2014; Lerdahl and Krumhansl, 2007). There, tension is understood as the phenomenal manifestation of a representation of structure:

> Listeners' unpremeditated awareness is not of hierarchical structure per se but of the patterns of tension and relaxation that arise from it. Theory speaks in terms of, say, the composing out of a tonic prolongation, but the immediate experience is one of rise and fall in tension. (Lerdahl and Krumhansl, 2007, p. 356)

Specifically, right- and left-branching dependencies reflect increasing and decreasing tension, respectively, with the degree of change in tension being quantified in terms of distances in Tonal Pitch Space (Lerdahl, 2001). Tension values propagate recursively from the root of the prolongational tree down to the observed surface events, which inherit tension values from all of their ancestors.

In the GTTM, patterns of tension and relaxation are understood as reflecting the *content* of a structural representation (in particular, of prolongational structure). However, structural representations may reflect a wider range of possible relations beyond tensing and relaxing, including e.g. preparation and prolongations. From this perspective, tension may be seen as the byproduct of processing effects resulting from the cognitive mechanisms that implement the *inference* of a structural representation, rather than as the content encoded in the structural representation itself. In particular, increased cognitive load induced by parsing may contribute to the build-up of tension, whereas a decrease in cognitive load may contribute to the complementary feeling of relaxation. Consistently with this view, the structural features that are typically associated with the phenomenon of musical tension (in principle at least) are also associated with increased processing complexity:

1. tension has been investigated as reflecting the embedding depth of surface events in

derivation trees (cf. point (1) above), showing that deeply embedded structures result in increased tension (Farbood, 2012; Lehne and Koelsch, 2015; Lerdahl and Krumhansl, 2007; Sun et al., 2020);

2. tension is often associated with establishing and delaying the resolution of expectations (Huron, 2006; Margulis, 2007), which in turn translates into prolonged recruitment of memory resources due to the maintenance of representations of hypothesised nodes (cf. point (2) above);

3. tension has been related to surprise (cf. point (8) above), and in particular to both the occurrence of an event that is unexpected *tout-court* (*surprise-tension*), as well as the occurrence of an event that denies the occurrence of a more expected one (*denial-tension*; Margulis, 2005; cf. also Lerdahl, 2001): both these circumstances require the parser to update, abandon, or revise the currently active representation (cf. points (3), (7), and (9) above, as well as discussion in Sections 3.5.1 and 3.5.2).

4. tension has also been associated with the computational cost of maintaining representations of multiple competing interpretations of ambiguous stimuli in a parallel-parsing model (cf. point (8) above; Jackendoff, 1991).

Overall, we have discussed how aspects of the experience of music may reflect (1) the emergence of representations of musical structure, and (2) the processes that construct such representations during listening. In particular, core notions such as tension and surprise, that are observable as scalar quantities, may in principle arise as a result of a variety of processing effects (Lehne et al., 2013), including prospective updates of expectations, violations of expectancy that are recoverable through retrospective revision, non-recoverable violations that lead to processing breakdown. A fine-grained empirical characterisation of the phenomenology of listening requires dedicated experimental paradigms that explicitly target these underlying mechanisms, rather than their shared observable manifestations.

3.6 The nature of structural representations

Before delving into the empirical content of the thesis, let us reconsider the role of structural representations in this context. The notion of representation is somewhat generic, since it boils down to establishing some kind of isomorphism between a representing system and a represented one (O'Brien and Opie, 2004; Shea, 2014). However, in the context of the present discussion, representations acquire more specific connotations that may also be tested empirically.

On one hand, a state of system A (say, a memory-storage system) may represent a state of system B (say, a grammar) if there is a homomorphism between the individual states of A and the individual states of B: in this case, at every moment in time, the state of A encodes information that uniquely identifies (a representation of) some state of B. We may call this a

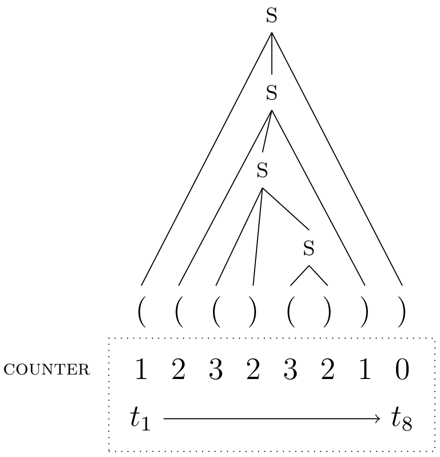


Figure 3.5 – Derivation tree for a string in the Dyck language (inner S nodes that terminate with an empty symbol ϵ are omitted from the tree), and a parser that simply stores a single integer which is increased or decreased by one unit whenever an open or closed parenthesis is encountered (effectively counting the number of open brackets at every moment in time). The sequence of states of the parser from t_1 to t_8 is a *process* representation of the derivation, but the individual states of the parser are not *state* representations of the (partial) derivation.

state representation. On the other hand, system A may still represent system B even if none of its individual states *represents* a state of B: this may happen, for example, if a *sequence* of states of A represents a state of B. In this case, A represents B in the form of a process *as it happened*: we may still reconstruct the represented state of B by looking at the given sequence of states of A *as a whole*, but at no individual point in time does system A encode information about B.

An example of such a *process* (as opposed to *state*) representation is displayed in Figure 3.5, which illustrates the sequence of states traversed by a parser that recognises strings in the Dyck language (the language of strings of opened and closed parentheses, based on the simple rule $\boxed{S \to \epsilon \mid (S)\,S}$). The parser simply counts the number of currently open brackets, so that individual states (encoding the current count) do not represent the structure of the string (Rohrmeier et al., 2014).[4] However, the sequence of states traversed by the parser does encode a representation of the tree structure of the string (loosely speaking, whenever the counter increases, one S node is introduced, and whenever the counter decreases, the corresponding surface symbol is attached to the most recently introduced S node).

"Experiments" probing the individual states of a process representation may still uncover correlates of features of the underlying structure (e.g., by observing that the sequence of counting states in Figure 3.5 correlates with tree depth), yet at no point in time (taken in isolation) does the system have access to an encoding of information that is necessary to reconstruct the derivation. Furthermore, while state representations can be stored as persistent memory

[4] Not even of the part of the string that has already been parsed: for example, there is no way to distinguish the representational content of the state at time t_3 from the one at time t_5.

states, process representations are by definition ephemeral, as they only exist in the form of a process that is concluded.[5] As a consequence, the information that process representations encode about the represented system is not *exploitable* by the representing system: once it is encoded, it cannot be retrieved nor manipulated. This is at odds with the characterisation of structural representations as being prone to manipulation and, in particular, to retrospective revision. For retrospective revision to occur, a representation of the structural identity of an event in the past needs to be accessible at a future moment in time. In other words, structural representations as characterised in music phenomenology are required to be state representations. An implication of this observation is that empirical findings showing responses to individual events (e.g., tension ratings or surprise) are not sufficient to establish the cognitive reality of structural representations: even if such responses correlate with structural features, this can at best demonstrate the existence of a process representation, rather than a state representation, of the underlying structure.

Another characteristic of structural representations is their *abstract* nature. In phenomenological accounts, an event experienced at time t_1 with a given identity may acquire a different identity (effectively becoming a "different mental object", Lewin, 1986, p. 353; cf. Moshaver, 2012) when experienced (as a retention or memory) at time $t_2 > t_1$. In Figure 3.2), for example, percept P_1 (encoding the sensory make-up of the chord g^6) is "included" in both P_3 and P_4, resulting in two different phenomenal experiences of the event g^6.

The observation that the phenomenal character of events is largely independent of their surface features (as defined in terms of, e.g., their pitch, temporal location, etc.) suggests that structural representations – taken as models of phenomenal experience – should be thought of as *abstracted from* surface identity, with structural and surface representations being related by some mapping that attributes structural identity to surface identity. In language, the notion that structural representations are abstracted from surface features is reflected in the phenomenon of structural priming (Branigan and Pickering, 2017): structural-priming effects support that representations encoding a sentence's syntactic structure alone are encoded separately from representations of phonological (and semantic) information. In particular, structural priming is consistent with the view that the structural-representation state that is active when surface a is attributed structural interpretation X is the same, or in some respects similar to, the structural-representation state that is active when surface b is also attributed structural interpretation X. In this sense, structural representations account for what is *common* among surfaces with analogous structure, irrespectively of the sensory input.

In music, evidence for the existence of structural representations that are separable from sensory information is more scarce. The availability of redundant representations has been hypothesised to strengthen memory for musical melodies, which exhibits remarkable robustness against decay caused by the interference of intervening stimuli between encoding and

[5] A process representation may still be turned into a state representation, e.g., in the form of a memory of the process, or in the form of some other persistent change within the representing system (e.g., changed connection weights in a connectionist system).

retrieval (Herff et al., 2019; Herff, Olsen, and Dean, 2018; Herff, Olsen, Prince, et al., 2018). In particular, the observation that such robustness is limited to structured melodies in a musical idiom that is familiar to the listener (e.g., it is not observed for melodies in unfamiliar tuning systems; Herff, Olsen, Dean, and Prince, 2018) suggests that it is structural representations specifically that provide such redundant encodings. After all, structural representations can be seen as generative derivations of the surface they refer to, so that information encoded in structural representations can be exploited to reconstruct (aspects of) the musical surface while memory for surface information decays (and vice versa). Cross-domain priming of repetition structure further indicates that representations of at least some domain-general structural feature are abstracted from sensory inputs during music listening. However, explicit evidence of structural priming in idiomatic music – involving representations of conventional, idiom-specific entities and relations – has not been observed. Chapters 4 and 5 directly address this issue.

In summary, showing the cognitive reality of structural representations entails something more than identifying just any behavioural or neural responses that correlate with the putative interpretation of musical surfaces. If the inference of structural representations is responsible for the phenomenal experience of music as characterised here, then structural representations should have the additional characterisation of being persistent over time as memory states, prone to retrospective manipulation, and abstracted from sensory information. Testing these general features is part of the burden of proof for a model of structural hearing as inference, alongside probing the specific properties of individual representations and parsing models as predicted by generative formalisations of music theory.

3.7 General outline

Considering the breadth of such an endeavour and the limited existing empirical background, the issue of testing a fully specified algorithmic theory of musical parsing exceeds the practical scope of this thesis. In this work, we contribute by (1) proposing a theoretical perspective that sets the ground for future work in this direction, while also (2) providing empirical proofs of existence for some of its critical implications. The remainder of this thesis will be devoted to presenting a series of studies that address different aspects of this approach.

In Part II, we focus on the first burden of proof for a model of a musical syntactic processor: namely, demonstrating that listeners are sensitive to the kind of representations that musical syntactic competence affords. As discussed in Section 3.6, a general assumption of our model of structural hearing is that structural interpretations are abstractions rather than encodings of the musical surface. In **Chapter 4**, we test whether information about the structural identity of sounds is retained in memory over time in addition to sensory information. In order to consolidate these results, we then draw inspiration from (psycho)linguistics by adapting the structural priming paradigm to the domain of music. In **Chapter 5** we therefore investigate the cognitive reality of representations of idiomatic harmony with a structural-priming paradigm that separately targets both local and global structural features.

Having addressed the cognitive reality of representations, we directly tackle the issue of how these are formed by laying the foundations of a cognitively plausible model of processing with specific algorithmic characteristics. This is the subject of Part III, where the ordering of the chapters is organised around the overarching goal of characterising the emergence of structural interpretations in music as the result of the operation of a *processor* for musical syntax. In the language domain, such a processor is assumed to operate based on (Steedman, 2000):

1. a *grammar*, which we take as reflecting idiom-specific principles for constructing structural representations as characterised in music theory;
2. a *parsing algorithm*, characterised by the time-course of executing specific parsing operations as well as by the computational cost of such operations;
3. an *oracle* that resolves ambiguity.

A grammar fulfils its role of matching surfaces with their interpretations by specifying what structural relations listeners are, in principle, sensitive to. In particular, grammars categorise entities in classes whose elements share the potential of forming certain specific kinds of structural relations with other entities. In language, this loosely corresponds to syntactic categories: words expressing the same syntactic category have analogous projections in terms of what relations they may entertain with other words (cf. Equation 3.1). A musical analogue of syntactic categories is harmonic functionality, which is based on the idea that classes of chords are substitutable to one another in that they share the same potential of forming relations with other chords. Provided that listeners have acquired the corresponding competence grammar, hearing two chords as expressing the same harmonic function entails to project analogous expectations towards future events, which differ across different functions. In **Chapter 6**, we probe the cognitive relevance of such a notion of syntactic categories in the musical domain, and specifically in the rather under-studied idiom of extended tonality.

Given a competence grammar, the crucial point of a model of syntactic processing is to specify how the processor exploits the grammar's rules to infer derivations for input surfaces. In **Chapter 7**, we formulate and test a model of real-time parsing based on a grammar for musical rhythm (Rohrmeier, 2020a). In language, the Dependency Locality Theory (Gibson, 2000) hypothesises a parsing algorithm that (1) integrates newly-encountered events in pre-existing partial representations, and (2) stores representations of future events (hypothesised nodes) implied by the current parse. By adapting the DLT to the musical domain, we propose an analogous model of parsing for musical rhythm, and we test the predictions of dependency-locality principles as predictors of processing complexity. This represents a first attempt to model the time-course of specific processing computations as well as their computational cost, and a first application of syntactic modelling to the perception of rhythm.

Finally, in **Chapter 8**, we address the third component of a syntactic processor, the "oracle". Specifically, we show evidence supporting the existence of garden-path effects in music, in analogy with the linguistic phenomenon. This supports the inferential and probabilistic nature of structural processing, and constrains any cognitively plausible algorithmic strategy to account for mechanisms of ambiguity resolution and retrospective revision. Furthermore, it provides additional evidence supporting the cognitive reality of structural representations encoding latent information that is abstracted from the musical surface, as discussed in Section 3.6.

Representation Part II

4 Representations of musical syntax in memory

Abstract

Memories of most stimuli in the auditory and other domains are prone to the disruptive interference of intervening events, whereby memory performance continuously declines as the number of intervening events increases. However, melodies in a familiar musical idiom are robust to such interference. We propose that representations of musical structure emerging from syntactic processing may provide partially redundant information that accounts for this robust encoding in memory. The present study employs tonally ambiguous melodies which afford two different syntactic interpretations in the tonal idiom. Crucially, since the melodies are ambiguous, memory across two presentations of the same melody cannot bias whether the interpretation in a second listening will be the same as the first, unless a representation of the first syntactic interpretation is also encoded in memory in addition to sensory information. The melodies were presented in a Memory Task, based on a continuous recognition paradigm, as well as in a Structure Task, where participants reported their syntactic interpretation of each melody following a disambiguating cue. Our results replicate memory-for-melody's robustness to interference, and further establish a predictive relationship between memory performance in the Memory Task and the robustness of syntactic interpretations against the bias introduced by the disambiguating cue in the Structure Task. As a consequence, our results support that a representation based on a disambiguating syntactic parse provides an additional, partially redundant encoding that feeds into memory alongside sensory information. Furthermore, establishing a relationship between memory performance and the formation of structural representations supports the relevance of syntactic relationships towards the experience of music.

This Chapter is published as: Cecchetti G., Herff S. A., & Rohrmeier M. A. (2021). Musical syntactic structure improves memory for melody: Evidence from the processing of ambiguous melodies. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **43**. GC conceptualised the study, implemented the experiment, analysed the data, and wrote the original manuscript with supervision by SAH and MAR.

4.1 Introduction

Memory for sensory stimuli is generally prone to disruptive interference due to new intervening information and to the passing of time (Eysenck and Keane, 2015). However, specific types of stimuli in different sensory modalities have been shown not to exhibit such a disruptive effect. For example, robust memory with respect to intervening items is observed for drawings (but not for photographs; Berman et al., 1991; Friedman, 1990; Konkle et al., 2010), for poetry (but not for prose; Tillmann and Dowling, 2007), and for melodies in a familiar musical idiom (but not for pitch sequences in unfamiliar tunings;Herff, Olsen, Dean, and Prince, 2018). In order to account for such phenomena, it was proposed under the Regenerative Multiple Representations conjecture (RMR) that some stimuli may afford additional representations that constitute memory traces coding partially redundant information, which can be used to compensate for interference effects (Herff et al., 2017; Herff, Olsen, and Dean, 2018). In fact, redundant information is in general a key tool for robust encoding, since redundancy affords to reconstruct missing or compromised data (MacKay, 2003; Shannon, 1948). As such, redundancy is of great importance for the robustness of computational (e.g., Merkey and Posner, 1984) as well as perceptual and cognitive processing (Barlow, 2001; Puchalla et al., 2005). In the context of memory, the RMR extends previous redundancy-based frameworks such as the multiple-trace (Hintzman, 1988) and the dual-coding theory (Paivio, 1969) to account for the aforementioned phenomena.

4.1.1 Robust memory for melody and musical structure

Predictions from the RMR are supported by converging evidence in the auditory domain, specifically addressing memory for melodies. Memory for novel melodies has been shown not to be disrupted by the passing of time (Schellenberg and Habashi, 2015) nor by the interference of other melodies intervening between first and second presentation (Herff, Olsen, and Dean, 2018). On the contrary, melodies in unfamiliar tuning systems (Herff, Olsen, Dean, and Prince, 2018; Herff, Olsen, Prince, et al., 2018) as well as rhythmic patterns obtained by removing pitch information from melodies (Herff, Olsen, Prince, et al., 2018) do exhibit a significant decay in recognition performance as a function of the number of intervening trials. A direct comparison against words and photographs showed that memory for melody is not generally better but instead deploys a mechanism that, after encoding, makes melodic memories resilient to interference (Herff et al., 2019). This is further supported by the literature on ‘earworms’ (Jakubowski et al., 2017) as well as clinical studies (Baird and Samson, 2014; Cuddy et al., 2015).

Memory performance in music is improved by the presence of structure, as quantified by the degree of adherence to idiom-specific music-theoretical norms (Cuddy et al., 1981; Cuddy et al., 1979; Deutsch, 1980). In particular, previous studies have suggested that the structured organisation of auditory events in time, which is a shared feature of music and poetry, may be responsible for the peculiar behaviour in memory of these types of stimuli (Tillmann and Dowling, 2007). Overall, embedding musical stimuli within a coherent formal structure is necessary for the robustness of memory, but if and how specific syntactic relationships linking musical events are relevant towards memory performance is uncertain (W. J. Dowling et al., 2001). Here, we propose and test the hypothesis that the beneficial effect of musical structure on memory, specifically the robustness in memory for melody, is mediated by the formation of representations of syntactic structure. In particular, we hypothesise that a representation of a stimulus' syntactic structure, distinct from its sensory representation, may constitute an additional representation encoding partially redundant information and hence contribute to robust encoding in memory as predicted by the RMR. For example, memory of the sensory information identifying the pitch of a note may be lost due to memory decay. However, if the note belongs to an idiomatic melody, syntactic relationships link that particular note with those preceding it. Such relationships form expectations (Rohrmeier, 2013) that point towards a specific pitch, thus potentially helping to recover its memory. Note that syntactic relationships would not be perceived within, e.g., melodies in an unfamiliar musical system, which would explain the different behaviour of melodies in an unfamiliar tuning as opposed to idiomatic ones.

4.1.2 Musical syntax as representation

Generative accounts of hierarchical musical structure distinguish between the musical surface, comprising a representation of the sensory events, and its syntactic interpretation, comprising the mutual interpretive relationships that recursively connect events with one another (Cecchetti et al., 2020; Lerdahl and Jackendoff, 1983a; Rohrmeier, 2020b). Examples of such interpretive relationships are preparation and prolongation in the context of tonal harmony (Rohrmeier, 2020b) and rhythm (Rohrmeier, 2020a), or contrapuntual elaborations (neighbouring motion, passing motion, etc.) in the context of monodic or polyphonic structure (Finkensiep et al., 2019; Schenker, 1935; Yust, 2015). Interpretive relationships and the way they can be combined recursively to account for a given musical surface are specific to each musical idiom.

Computational accounts of musical syntactic processing formalise music-theoretical expert-knowledge and also capture many aspects of the experience of musical structure (Herff, Harasim, et al., 2021). This includes predictions for harmonic pattern completions (Herff, Harasim, et al., 2021), expectations arising from hierarchical dependency relations (Cheung et al., 2018; Koelsch et al., 2013), and interference with linguistic syntactic processing (Patel, 1998; Slevc et al., 2009). In particular, a core prediction of hierarchical syntactic models of music cognition is that a representation of the syntactic interpretation is formed through a process of parsing, (Jackendoff, 1991; Rohrmeier, 2013, 2020b), and it is a challenge for both theoretical and empirical research to understand how the availability of such a representation would manifest itself in and impact upon other cognitive functions.

The RMR provides a framework to test the cognitive relevance of syntactic representations by showing their impact on the formation and retrieval of memory. Idiomatic melodies are pecu-

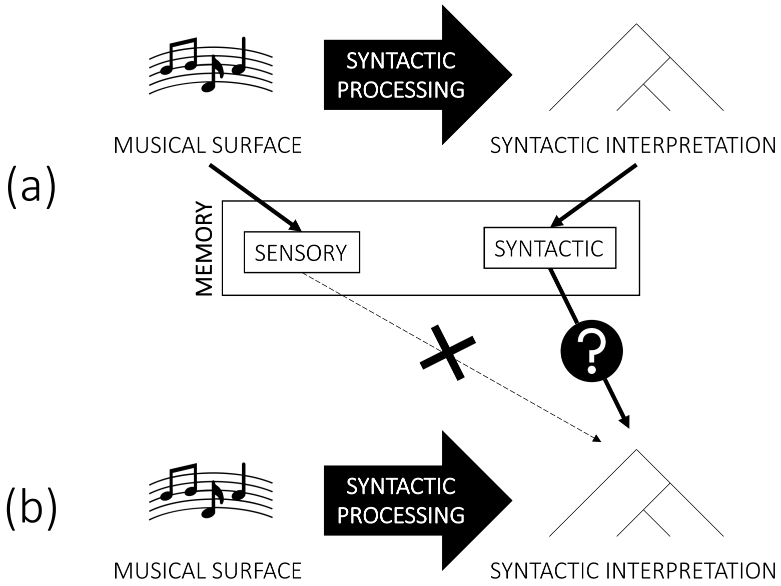


Figure 4.1 – Memory of the first presentation **(a)** of an ambiguous musical surface cannot influence the outcome of syntactic processing in a second presentation **(b)**, whereas memory of the syntactic interpretation, if encoded in memory, could.

liar among non-linguistic auditory stimuli, insofar as they can be perceived as syntactically-interpretable units by listeners who are familiar with the syntactic principles of the given musical idiom. In turn, if information related to the syntactic interpretation is stored in memory alongside sensory information, this may provide the necessary redundancy for the robust encoding of melodies. Furthermore, from a computational perspective, syntactic organisation affords higher encoding compression, resulting in more efficient representations potentially saving memory resources and improving performance (Rohrmeier and Pearce, 2018).

4.1.3 Structural ambiguity: the present approach

In order to test the hypothesis that representations of musical structure contribute to memory, we focus here on a set of novel tonally-ambiguous melodies. These melodies are constructed so that two different syntactic interpretations can be attributed to the same set of sensory events comprising the musical surface. Specifically, each melody may be heard in two different keys in the tonal idiom. As a consequence, a representation of the sensory information alone (e.g., the pitch of each note) is insufficient to uniquely determine a syntactic interpretation. The latter constitutes a separate representation that has to be processed upon listening based on the listeners' syntactic competence.

The presentation of a key-defining chord at the end of a melody, however, may retrospectively bias listeners towards one or the other plausible syntactic interpretation (cf. J. Fodor and Ferreira, 1998 in language). Across multiple presentations of the same melody with different key-defining chords, participants may then change their syntactic interpretation of the melody according to the key-defining chord itself. However, it is also possible that a specific syntactic interpretation is formed during the first presentation and then remains *stable* across successive presentations, even if the key-defining chord presented at the end of the melody changes. Note that, in principle, the stability of a syntactic interpretation characterises the syntactic processing of a melody, not its memory: it indicates that the outcome of syntactic processing on that particular input is the same in two different attempts. However, the outcome of syntactic processing (the syntactic interpretation) may be represented and stored in memory alongside sensory information (the musical surface), forming an additional memory trace for the melody (Figure 4.1a). If such syntactic information from previous parsing attempts complements sensory information in memory, retrieving the memory of syntactic information upon a subsequent presentation of the same melody may influence the subsequent parsing attempt (Figure 4.1b, solid arrow). Specifically, a stronger memory trace of the syntactic information would result in a higher likelihood for the syntactic interpretation to be stable across multiple parsing attempts. Crucially, when dealing with ambiguous stimuli, memory of the sensory information alone would not be able to bias the outcome of subsequent parsing attempts (Figure 4.1b, dashed arrow), sensory information being ambiguous. As a consequence, evidence for a predictive relationship between memory performance and stability of syntactic interpretations in the same melody stimuli supports the existence of a representation of syntactic information in memory.

Furthermore, if such a syntactic representation concurs towards the robustness of melodic memory, for example by sensory and syntactic representations coding partially redundant information that can be used to recover each other, it should also exhibit robustness to interference. Our paradigm affords to test this hypothesis by showing whether the likelihood for syntactic interpretations to be stable across multiple presentations decreases with increasing number of intervening trials.

4.1.4 Aims and hypotheses

In this experiment, we investigate whether the emergence of a syntactic interpretation is related to the formation and retrieval of memory for a melody, as predicted by the RMR conjecture under the additional hypothesis that syntactic interpretations specifically contribute to redundancy in memory for melody. In particular, we hypothesise (1) that the melodies are robust to memory interference, as suggested by previous evidence concerning idiomatic melodies, (2) that stronger memory performance is associated with higher likelihood for stable syntactic interpretations, and (3) that this likelihood does not decay with the number of intervening trials.

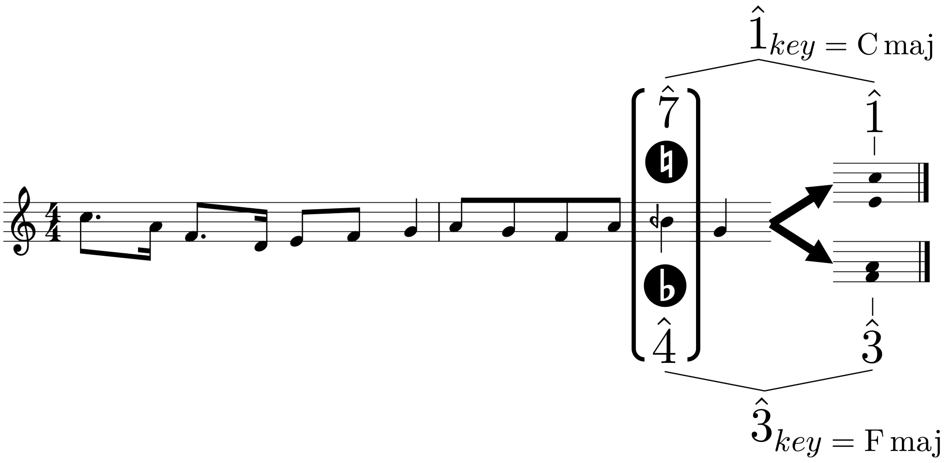


Figure 4.2 – Example stimulus. The quarter-tone B can be interpreted as the lower-neighbour elaboration of $\hat{1}$ in C major, to be tuned upwards as a B natural ($\hat{1} \rightarrow \hat{7}\ \hat{1}$, top), or as the upper-neighbour elaboration of $\hat{3}$, to be tuned downwards as a B flat ($\hat{3} \rightarrow \hat{4}\ \hat{3}$, bottom). In the Structure Task, each presentation of the stimulus is followed by one of the two key-defining chords shown on the right.

4.2 Methods

4.2.1 Participants

Sixty-two participants (median age 25.5, range 18-74) took part in the online experimental session. Participants were recruited among students and professional musicians from several European music academies, as well as through the online recruitment platform Prolific Academic. As a result, various degrees of musical expertise are represented (Goldsmith Music Sophistication Index (MSI), Musical Training subscale: median 0.61, range 0.14-0.92; Müllensiefen et al., 2014), with all participants reporting at least one genre within Western musical practices (e.g., classical, Jazz, Rock/Pop) as their main listening habit. To control for potential mediating effects of musical expertise, we include musical sophistication in our statistical analyses. However, no effects were observed. The participants' involvement was reimbursed with CHF 15, and ethics approval was granted by the research-ethics board of the host institution (HREC 037-2020).

4.2.2 Stimuli

Fifteen original melodies, each spanning 2 bars in 4/4 meter at 120bpm, were synthesized in MuseScore 3.5.0 in the default piano timbre, ranging from C4 to G5 with 440Hz tuning. Melodies were made tonally ambiguous by means of two compositional criteria. First, each melody supports a tonal harmonisation in two different keys (C major and F major) provided that the key-discriminating note B (the only pitch class that is not shared between the two keys) is given the appropriate accidental; furthermore, all occurrences of the key-discriminating pitch class B are de-tuned by a quarter tone, so as to fall halfway between B and B flat (Figure

4.2).

4.2.3 General procedure

Within the online experimental session, lasting 45 minutes, participants were administered two behavioural tasks, a Memory Task and a Structure Task, both comprising the same set of stimuli described above, followed by the Goldsmith MSI (Müllensiefen et al., 2014). The experimental interface was implemented in PsychoPy3 (Peirce et al., 2019) and administered online through the platform Pavlovia (https://pavlovia.org/).

In the Memory Task each melody was presented twice, in random order and transposition, within a continuous recognition paradigm (Shepard and Teghtsoonian, 1961). As a consequence, the Memory Task comprised 30 consecutive trials, and the number of intervening trials between two presentations of the same melody was randomised within and across participants. In each trial, following the presentation of a melody, participants were asked to report whether they believed the melody to be 'new' or 'repeated'. Participants were instructed, by means of an example, to consider the second occurrence of a melody in a different transposition as a repetition.

In the Structure Task, the same melodies were also presented twice throughout the experiment in random order, and each time they were completed with a different key-defining chord. The chord provided post-hoc information to bias the listeners in favour of one out of the two plausible tonal interpretations of the melody. Two behavioural measures were collected in each trial: first, participants were asked to rate how surprising the chord sounded to them; then, participants were asked to reproduce the melody by selecting the 12-equal-tempered tuning of B or B flat for the de-tuned note. This response is taken as a proxy of the participants' syntactic interpretation of the melody. Selecting the sharp or the flat tuning of the quarter-tone note indicates a preference for hearing that note in the syntactic role of an upper-neighbour or a lower-neighbour elaboration (Figure 4.2).

4.3 Results

In order to account for inter-subject and inter-stimulus variability, statistical analyses are conducted with Bayesian mixed effects models (implemented in the R package *brms*; Bürkner, 2018) allowing for cross-random intercepts for individual participants and stimuli. All non-categorical variables are scaled to null mean and unitary standard deviation. Models were provided with weakly informative priors $t(3,0,1)$ (Gelman et al., 2008), and we report coefficient estimates (β), estimated errors in the coefficients (EE), and Evidence ratios ($Odds$) for the individual hypotheses. An asterisk (*) identifies parameters such that $Odds(\beta \lesseqgtr 0) > 19$, corresponding to statistical significance at the conventional 95% confidence level (Milne and Herff, 2020). Data, code and stimuli can be accessed at https://osf.io/ujnef/.

4.3.1 Robustness to interference

In order to test the robustness of memory to the interference of intervening trials, we quantify the predictive power of the number of intervening trials towards the correctness of the participants' recognition responses. In order to account for potential participant- and stimulus-specific biases in how participants' recognition responses vary during the course of the task, we estimate the Dynamic Response Tendency for each participant and use it to correct participant-wise for false-alarm rates over the course of the experiment (DRT; Herff, Olsen, and Dean, 2018). The DRT is the probability for the first presentation of a melody to be recognised (incorrectly) as a repetition just based on the time elapsed since the beginning of the experiment. This is estimated with a linear mixed effects model predicting the recognition response based on the trial number. The DRT, alongside the number of intervening trials, appears then as a predictor in a Bayesian mixed-effects model predicting recognition responses to the second presentations of melodies. As hypothesised, the number of intervening trials separating the repetition of a melody from its first occurrence in the experimental task carries no predictive power towards the participants' recognition responses to the second presentations of melodies ($\beta = -.04$, $EE = .05$, $Odds(\beta < 0) = 3.59$).

4.3.2 Linking memory and structure

We then assess whether memory performance for a given melody carries predictive power towards the stability of the syntactic interpretation of the melody itself, i.e. whether the Tuning Response remains the same across the two presentations of the melody in the Structure Task irrespective of the key-defining chord. To this end, a Bayesian mixed-effects model predicting the stability of the Tuning Response for a given melody is provided with several predictors: the memory performance for that melody from the Memory Task; the participant's musical training score from the musical-sophistication questionnaire, and its interference with memory performance; the difference in Surprise Rating between the two presentations of the melody in the Structure task; finally, the number of intervening trials between the two presentations of the melody in the Structure task. Specifically, the memory performance is expressed as a categorical predictor indicating which presentations of the melody (none, the first only, the second only, or both) were correctly identified.

As hypothesised, strong evidence supports that memory performance in the Memory Task carries predictive power towards the stability of syntactic interpretations in the Structure Task (Figure 4.3). Melodies that are correctly identified as new or repeated at least once in the Memory Task are predicted to exhibit stable syntactic interpretations in the Structure Task with significantly higher likelihood compared to melodies that are never identified correctly in the Memory Task (First: $\beta = .78$, $EE = .34$, $Odds(\beta > 0) = 121.45^{*}$; Second: $\beta = .66$, $EE = .37$, $Odds(\beta > 0) = 31.88^{*}$; Both: $\beta = .87$, $EE = .33$, $Odds(\beta > 0) = 377.95^{*}$).

No evidence is found for an effect of any other predictor. Specifically, the stability of syntactic interpretations is not influenced by the difference in Surprise Rating between the two presen-

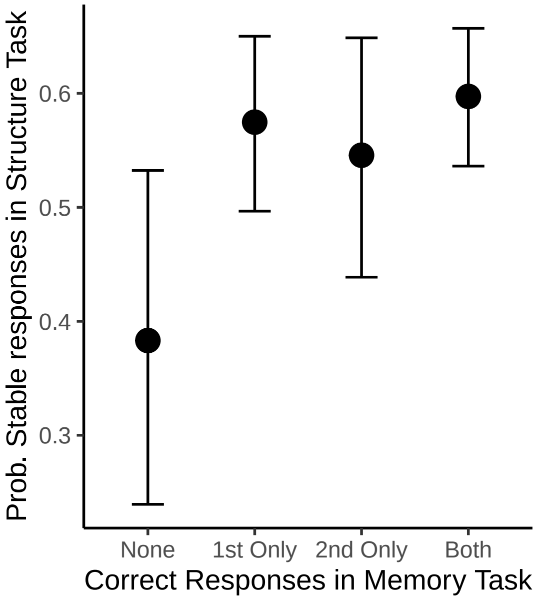


Figure 4.3 – Correct response in First presentation only, Second presentation only, or Both presentations of a melody in the Memory Task predicts higher probability of stable responses in the Structure Task (estimates with 95% Confidence Interval).

tations of the same melody in the Structure Task ($\beta = -.005$, $EE = .07$, $Odds(\beta > 0) = 1.11$), nor by the number of intervening trials separating them ($\beta = -.07$, $EE = .07$, $Odds(\beta > 0) = 5.07$). Furthermore, musical training does not influence the likelihood of stable Tuning Responses ($\beta = -.12$, $EE = .29$, $Odds(\beta > 0) = 2.02$) and also does not modulate the effect of memory performance (all $Odds(\beta > 0) < 10$).

4.4 Discussion

In this experiment, we investigate the relationship between memory performance, as captured in a continuous recognition task, and the stability of syntactic interpretations of tonally ambiguous melodies across multiple presentations. Our results, obtained over two experimental tasks involving a novel set of tonally-ambiguous melodies, support our first hypothesis and previous evidence that the recognition of previously-heard melodies is robust to the interference of intervening trials (Herff et al., 2019; Tillmann and Dowling, 2007). As an explanation for this phenomenon, it has been proposed that multiple partially redundant representations concur to compensate for disrupted memory performance. Here, we further tested the hypothesis that representations emerging as a result of syntactic processing contribute to robust memory encoding. Our results indicate that increased memory performance in a melody predicts higher stability of syntactic interpretations in the same melody, suggesting that the outcome of syntactic processing does play a role in the formation of memory traces.

While this evidence does not directly identify a causal relationship between the formation of syntactic interpretations and memory performance, it does indicate that memory for melody includes a representation of a syntactic interpretation beyond the sensory representation of the stimulus. In fact, if increased syntactic stability is the byproduct of a stronger memory trace of the melody, this memory trace must include a representation of the syntactic interpretation itself, since sensory information does not point to a single syntactic interpretation in presence of ambiguity. In other words, a strong memory of the syntactic interpretation generated during the first presentation primes the perception and interpretation of the second presentation.

While this observation parallels analogous phenomena explored in the psycholinguistic literature (cf. Branigan and Pickering, 2017), evidence for this manifestation of syntactic priming in music is still scarce. Previous priming paradigms in music have shown effects of processing facilitation that cannot be explained in terms of sensory information alone (Bigand et al., 2005; Tekman and Bharucha, 1998), and demonstrated that the perception of subsequent syntactic structures can be influenced by abstract features of priming and target stimuli such as harmonic (Bharucha and Stoeckig, 1986) or stylistic (Vuvan and Hughes, 2019) relatedness. However, the present results specifically support the hypothesis that syntactic representations formed at different moments in time influence one another. Such an effect has only been previously observed in the cross-domain interaction of simple, non-idiomatic musical stimuli and linguistic sentences (Van de Cavey, 2016). As a consequence, our results provide new evidence for an effect of musical syntactic priming based on the tonal idiom, which may be further investigated in future studies.

We further observed that the number of intervening trials separating two presentations of the same melody in the Structure Task does not influence the likelihood for the syntactic interpretation of that melody to be stable. This suggests that any influence on the second parsing attempt due to the syntactic memory trace from the first parsing attempt does not decline with increasing number of intervening trials. As discussed above, sensory information alone declines over time and, for ambiguous melodies, it is not sufficient to determine the stability of a syntactic interpretation, yet both melody recognition and the stability of syntactic interpretations do not decline with the number of intervening items, when both are available. This is consistent with the hypothesis that the additional existence of representations of syntactic information in memory is robust to such interference and may account for the peculiar behaviour of memory for melody in this respect, either on its own or because of the reciprocal regenerative interaction with sensory information when both are available.

Overall, syntactic structure is shown to be a viable candidate in the role of an additional, partially redundant representation explaining the peculiar robust behaviour of memory for melody under the RMR. Critically, the experimental paradigm based on syntactically ambiguous stimuli affords to discriminate sensory and syntactic information, so that the latter can be shown to constitute an additional representation which is not reducible to the sensory one, i.e. the musical surface. The observed impact on the operation of memory highlights a specific cognitive function of musical syntactic structures which has been suggested on theoretical grounds (Rohrmeier and Pearce, 2018).

Finally, it is important to note that memory is subject to expertise effects, with expert musicians showing better memory for music (M. A. Cohen et al., 2011; Herff and Czernochowski, 2019) especially when presented with a familiar idiom (Halpern and Bower, 1982). Nevertheless, while our study involved a wide spectrum of participants comprising musically naive listeners as well as highly sophisticated musicians, results suggest that generic familiarity with the Western tonal idiom seems to be sufficient to determine the observed interplay between syntactic processing and memory, and musical expertise does not mediate the strength of this relationship. Further analyses on data from this and future studies may shed light on the role of formal training and explicit domain-specific knowledge.

4.5 Conclusion

We have shown evidence that musical syntactic processing and memory performance are mutually predictive. While supporting converging evidence concerning the robust behaviour of memory for melody, our results further substantiate the hypothesis that representational redundancy plays a role in the formation and retrieval of such memory. Specifically, results are consistent with the hypothesis that syntactic interpretations arising as the outcome of musical syntactic processing constitute an additional memory representation that is involved in the resilience of memory for melody towards interference.

5 Priming of hierarchical harmonic structure

Abstract

Structural priming is a well-established methodology in psycholinguistics to investigate mental representations of linguistic structure which are independent of the specific sensory (e.g, phonological) features of sentences. In this study, we implemented a structural-priming paradigm in the musical domain to investigate whether structural representations are formed during listening that encode music-theoretically relevant structural features at different levels of abstraction ("shallow" and "deep"). Ninety-nine participants were presented with pairs of prime and target stimuli while engaging with a visual-flash reaction task as well as a memory task. Bayesian mixed-effects models showed a marked difference in participants' performance in the two tasks depending on the sharedness or non-sharedness of shallow and deep structural features between primes and targets. This difference was further shaped by increased exposure to the stimulus materials over the course of the experiment. These results support that representations of idiomatic harmonic structure are formed spontaneously during music listening, and that such representations encode information about both shallow and deep levels of structural abstraction as predicted by hierarchical models of tonal structure.

This Chapter is currently under review as: Cecchetti G., Herff S. A., & Rohrmeier M. A. (in review). Priming of abstract harmonic structure in music. Journal of Experimental Psychology: Human Perception and Performance. GC conceptualised the study, implemented the experiment, analysed the data, and wrote the original manuscript with supervision by SAH and MAR.

5.1 Introduction

Most music theoretical and analytical approaches rely on the assumption that the musical surface as heard is the observable manifestation of some unobserved abstract underlying entities, such as chords or tonal functions, which stand in some relation to one another (Cecchetti et al., 2020; Dubiel, 2017; Finkensiep, 2023; Salzer, 1962). The interpretation of the musical surface in terms of these latent entities and relations – in short, structures – is often made explicit in the form of annotations at different levels of abstraction, such as Roman numerals (Piston, 1948) or Riemannian functions (Riemann, 1893), and in graphical analyses such as Schenkerian analyses (Schenker, 1935), hierarchical reductions (Lerdahl and Jackendoff, 1983a), derivation trees (Rohrmeier, 2011, 2020b), or voice-leading graphs (Finkensiep, 2023; Finkensiep and Rohrmeier, 2021; Yust, 2018). These models are meant to capture the "syntactic competence" (Chomsky, 1965) of a given musical idiom. In other words, they aim at characterising the structural representations that may be formed in the mind of an ideal "native speaker" of the given musical idiom, e.g., an ideal listener or composer, while listening to or reasoning about music (Cecchetti et al., 2020; Jackendoff and Lerdahl, 2006; Lerdahl and Jackendoff, 1983a; Rohrmeier, 2011). In this study, we investigate the psychological reality of such music-theoretically relevant representations in the idiom of Western tonality.

5.1.1 Abstract structural representations in music perception

An important commonality shared by the aforementioned approaches is the idea that some aspect of the underlying structure is invariant to transformations of the musical surface: for example, in Western tonality, different chords may be substituted to one another while preserving the same structural function, as empirically supported by corpus-based (Jacoby et al., 2015; Rohrmeier and Cross, 2008; White and Quinn, 2018) and psychological (Bigand et al., 1996; Cecchetti, Herff, Finkensiep, et al., 2023; Popescu et al., 2022) studies. As a consequence of this kind of invariance, if listeners form representations of the music-theoretically predicted structures, these representations need to be distinct from representations of sensory information, which is tied to a specific musical surface. As an example, in Figure 5.1, excerpts (a) and (b) differ substantially in terms of pitch content, yet they share a common structure in terms of how events relate to one another: the first chord is "prolonged" all the way to the final chord, whereas the second chord "leads to" or "prepares" the third chord which in turn prepares the last chord. A representation of the sensory content (reflecting the score notation in Figure 5.1) would then look quite different for the two examples, whereas a "structural" representation of the harmonic relations (encoding the arrows in Figure 5.1) would look comparatively similar in the two cases. Vice versa, the first two events in excerpts (b) and (c) share identical pitch content, yet the way they relate to each other and to the rest of the respective chord progression is different: in particular, differently from (b), the first chord in (c) is interpreted as preparing the second.

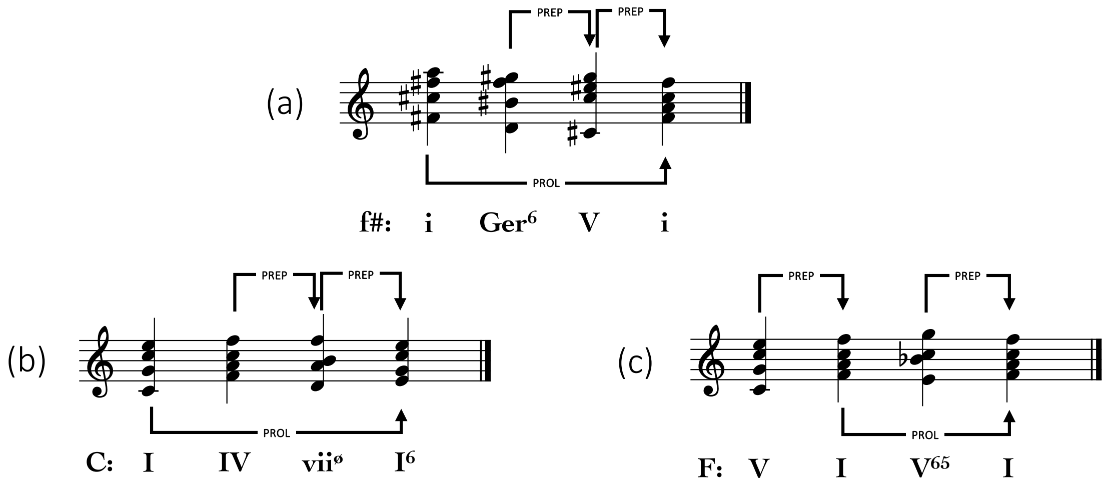


Figure 5.1 – Three tonal chord progressions exemplifying the differentiation of structural and sensory information. Progressions in **(a)** and **(b)** share analogous structural relations yet different sensory appearance. Vice versa, progressions **(b)** and **(c)** are comparatively similar in their sensory makeup (in particular, the first two chords are identical), but individual chords exhibit different structural relations with each other and with the remaining chords in the respective sequences.

In summary, representations of such structural interpretations are abstractions of the sensory information contained in the stimuli: stimuli that differ in their sensory makeup can share the same structural interpretation, and vice versa. Representations of structural relations, such as the ones exemplified in Figure 5.1, are assumed by most music-theoretical frameworks, as it is a primary concern of music theory to characterise what types of structural relationships are relevant for composition or analysis in a given musical idiom. However, while there is no question that representations of sensory information are formed during (music) listening, the cognitive reality of structural representations is harder to support with empirical evidence, as the interpretation of a musical passage is strongly linked to its sensory makeup (cf. Bigand et al., 2003).

Evidence from implicit-learning of artificial musical grammars (Jonaitis and Saffran, 2009; Loui, 2012; Rohrmeier, 2010; Rohrmeier and Rebuschat, 2012; Rohrmeier and Widdess, 2017; Tillmann et al., 2000) supports that listeners are sensitive to structural features of a (novel) musical idiom. Specifically, listeners can acquire implicit knowledge about the regularities or rules in a musical idiom and exploit this implicit knowledge to discriminate between stimuli that follow the rules of the idiom and those that do not. While these results demonstrate a general capacity to form idiom-specific representations of structural information, they do not investigate the particular representations formed for individual stimuli within a specific idiom, similarly to how representations for individual sentences are investigated in linguistics (cf. Branigan and Pickering, 2017). In this vein, Bigand (1990) showed that both musicians and non-musicians were able to distinguish families of idiomatic tonal musical stimuli that shared a specific common structure while differing in their pitch content. The absence of an effect of formal musical training and the anecdotal inability of participants to explain the thought process leading to above-chance performance suggest that the effect relies on subconscious, possibly automatically-processed representations and implicit knowledge. However, since participants were asked to report the family-membership judgement explicitly, it is possible that any notion of structural (dis)similarity only emerged intentionally and/or explicitly in the process of tackling the experimental task through reasoning, rather than being encoded in automatically processed mental representations induced by listening irrespectively of the experimental task. In this paper, we specifically investigate whether structural representations that encode structural relations (abstracted from their sensory instantiations) are formed automatically and implicitly during music listening, and what kind of information these representations encode. We address this issue by testing the existence of implicit and task-irrelevant manifestations of such representations. Specifically, we exploit a possible structural priming effect associated with two different levels of structural abstraction, that we introduce in the following section.

5.1.2 Two levels of structural abstraction: shallow and deep structure

An important aspect of structural relationships in Western tonal music is that they are hierarchically nested (cf. Asano et al., 2021; Lerdahl, 2015; Rohrmeier and Pearce, 2018 for theoretical arguments; Dibben, 1994; Harasim, 2020; Herff, Bonetti, et al., 2023; Herff, Harasim, et al., 2021; Koelsch et al., 2013; Lerdahl and Krumhansl, 2007; Serafine et al., 1989 for empirical support). In particular, generative accounts of musical structure model the observed musical events as the result of the elaboration of simpler templates through the recursive application of transformations. The different transformations, in turn, encode the possible relations linking events with one another: e.g., whether an event is interpreted as a prolongation of another, or as its preparation (cf. Figure 5.1). From this perspective, the structural interpretation of each event is modelled as the trace of the sequence of transformations that "justify" the presence of that particular event in the musical surface. For example, the tree structures in Figure 5.2 visualise the generative derivation of two chord progressions consisting of a common stem followed by a cadential closure. Note how the initial stem of the two stimuli, which is identical, has a different structural interpretation in the two cases, i.e., it carries different labels and is embedded differently into the tree structure.

The hierarchical nature of structural relations entails that representations of structure are possible at different levels of abstraction. For example, we could focus on lower hierarchical levels, closest to the surface. Such a "shallow" structural description encodes information that music theorists may refer to as the functional role of the individual chords in the given key (as in, e.g., Piston, 1948) – irrespectively of how the individual chords themselves are embedded in a specific derivation of the entire musical passage. This characterisation of structure is often encoded, for example, in Roman-numeral labelling or similar annotations and constitutes the leaves of a tree analysis (Figure 5.2, bottom). Vice versa, we could rather focus on a "deep" level of the hierarchy, which encodes a simple underlying harmonic template from which the entire surface is derived recursively. In particular, with "deep" structural level, we refer here to

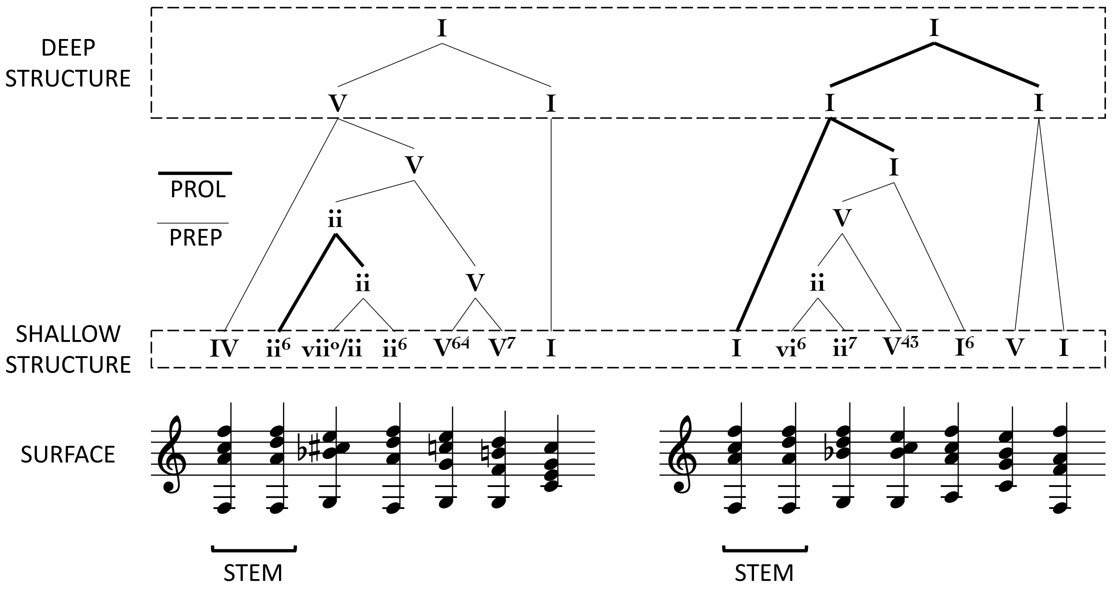


Figure 5.2 – Structural representation of two chord progressions after the music-theoretical formalism by (Rohrmeier, 2011, 2020b). Two different levels of structural description are highlighted in the dashed boxes: the functional role of individual chords, marked as "shallow structure", is represented as Roman-numeral labels at the bottom of the hierarchical representation; a more abstract level of representation comprising the top of the tree is marked as "deep structure". Thick lines represent the harmonic relationship of "prolongation" (PROL), whereas thin lines represent "preparation" (PREP). Note how the initial stem is identical in the two stimuli, yet both its shallow-structural representation and its embedding into the deep-structural level are different.

the very first step in the derivation of a musical passage starting from the root of a tree analysis. In the example displayed in Figure 5.2 (top), the two chord progressions differ at this deep structural level: in one case, the beginning of the progression encompasses a dominant region (V) that leads towards (prepares) the cadential tonic (resulting in a V-I template), whereas in the other case, the progression begins with a tonic region (I) that prolongs the cadential tonic (I-I).

Existing literature provides conflicting evidence regarding to what extent listeners are sensitive to fine-grained structural features and, in particular, the relative perceptual importance of local (or shallow) and global (or deep) structural descriptions is debated (cf. Tillmann and Bigand, 2004). However, it is not necessarily the case that such fine-grained structural features are perceived immediately during first-pass listening. It is rather plausible that subtle structural features are implicitly "discovered" as exposure increases over multiple hearings. Indeed, subtle effects of musical structure such as non-local harmonic dependencies have been demonstrated in experiments involving repeated exposure to the musical stimuli (e.g., Koelsch et al., 2013), and not in experiments only allowing for limited exposure (e.g., Cook, 1987; cf. discussion in Koelsch et al., 2013). Overall, the goal of this paper is to investigate the psychological reality of both shallow and deep structural representations during music listening, and, to this end, we also take into account the possible effect of exposure over the course of the experiment. In the following, we introduce the experimental approach that we adopted.

5.1.3 Structural priming as an empirical window on representations

The issue of identifying cognitively relevant representations is not unique to music: linguistic models of syntax characterise representations of the relationships linking words within a sentence (Chomsky, 1957), and it is a challenge for empirical research to test how these representations relate to those formed in language production and comprehension. A prominent framework for investigating linguistic representations is structural priming (Branigan and Pickering, 2017; Branigan et al., 1995), which is based on the assumption that "if processing one stimulus [the "prime"] affects the subsequent processing of another stimulus [the "target"], then these stimuli share some aspect of their representation" (Branigan and Pickering, 2017, p. 2). If such an influence is observed between a prime and a target that share some aspect of their linguistic structure but are otherwise unrelated, then the observed effect is evidence for the existence of a representation that encodes that particular aspect of linguistic structure. For example, the observation that sentences with shared syntactic structure yet unrelated semantic (e.g., Bock and Loebell, 1990) or phonological (e.g., Bock, 1989; Pickering and Branigan, 1998) content can prime each other suggests the existence of a level of linguistic representation that encodes syntactic relations abstracted from specific lexical or auditory instantiations. Priming paradigms constitute then a promising avenue for investigating representations at different levels of abstraction.

In music, priming paradigms are often adopted where a behavioural response to a target "probe tone" is influenced by the prior exposure to a musical context, which works as a prime (e.g., Bharucha and Stoeckig, 1986; Krumhansl and Kessler, 1982). These results can be interpreted as a form of structural priming between a (shallow-)structural representation of the context alone and one including the probe-tone. With respect to these results, though, encodings that are not independent of a specific sensory instantiation are in principle sufficient to account for the empirical observations. For example, connectionist models formalising the activation of key-specific representations based on the concrete pitches sounding in the musical surface successfully account for many such priming effects (e.g., Bharucha, 1987; Bharucha and Stoeckig, 1986). Even in studies that manipulate the order or the structural relations in the prime, so that they cannot be easily explained by spreading-activation accounts (e.g., Koelsch et al., 2013; Sears et al., 2023), the "target" tone (i.e., where the behavioural or neural measure is acquired) is typically meant to be heard in relation to the key established by the priming chord sequence. In all these cases, any effect on the target may be due to key-specific representation of the prime as opposed to a key-independent representation of its internal structural relations. Overall, evidence from these studies does not directly support that representations of structural relations are encoded in isolation from the concrete sensory instantiation.

Cross-modal priming is a possible avenue to overcome this concern. Pitch sequences obtained by drawing tones from two distinct sets A and B have recently been shown to prime low- or high-attachment in linguistic sentences depending on their arrangement in an ABB or ABA pattern, respectively (Van de Cavey and Hartsuiker, 2016). This result suggests that a representation of the prolongational dependency linking the two repeated instances of the sequence (A...A or ...BB) is formed during listening and can interfere with the formation of analogous structural representations across domains. However, due to the non-idiomaticity of the repetition structure adopted in the stimuli, it remains unclear whether implicitly-acquired idiomatic structuring principles as observed in actual musical practices (such as the harmonic relations of preparation and prolongation exemplified in Figures 5.1 and 5.2) afford the formation of structural representations, and whether these can prime each other within the musical domain. Cecchetti et al. (2021) recently showed evidence that memory representations of tonal melodies encode information about idiomatic structural interpretations. Such information could influence the perception of subsequent presentations of the same melody as expected in the context of structural priming. Based on these preliminary observations, we implemented here a novel structural-priming paradigm targeting common-practice tonal harmony as discussed in the following.

5.1.4 Implicit manifestations of structural priming

As it is typical in priming studies, the experiment was based on the presentation of a pair of stimuli in each trial under the assumption that exposure to the first stimulus, the Prime, may have influenced the perception of the second, the Target, provided that they shared some aspect of the (shallow- or deep-structural) representations they elicited. To this end, we distinguished between two categories of trials. On one hand, those where the Prime and the Target stimuli were related by the sharedness (Congruent priming condition) or non-sharedness (Incongruent priming condition) of aspects of both shallow and deep structure (for brevity, Shallow-Structure trials), and, on the other hand, those where Prime and Targets were related by the sharedness or non-sharedness of aspects of deep structure only (Deep-Structure trials).

Priming paradigms in language comprehension can rely on participants explicitly reporting their understanding of a sentence (e.g., Branigan et al., 2005). This is not generally possible in music-to-music priming, as there is no standard way (other than highly specialised analytical techniques) to report structural understanding of music in the absence of an associated semantics. Nevertheless, priming may facilitate the activation of some structural representations over others, which may in turn provide implicit processing advantages towards other tasks.

First, priming may impact reactions to extraneous stimuli presented while listening to music. For example, if a visual flash occurs jointly with a chord with a given shallow-structural interpretation, listeners may find a second occurrence of the flash more expected if it coincides with a chord with the same shallow-structural interpretation as in the first occurrence. As strong evidence supports that reaction times are generally shorter when stimuli are expected (in music, cf. Bharucha and Stoeckig, 1986; Bigand and Pineau, 1997; Politimou et al., 2021; Sears et al., 2019; Wall et al., 2020), listeners may be faster to react to the flash in this case, compared to a case where the second occurrence of the flash coincides with a chord with a different interpretation. This facilitating effect of shallow-structural congruency in a Flash Reaction Task may be understood as an implicit manifestation of priming.

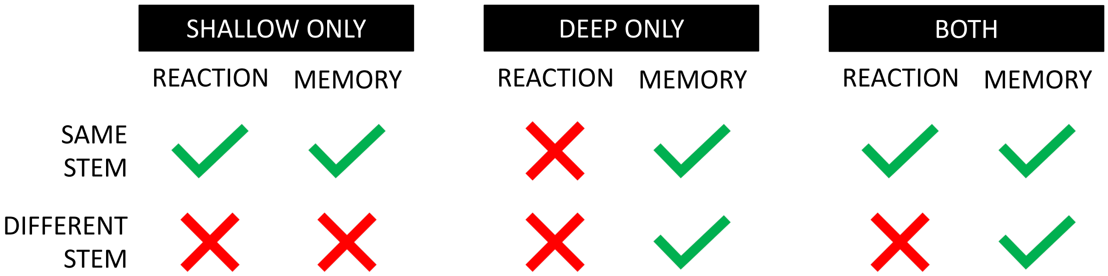
Second, listeners may find it easier to process a target musical stimulus if it affords the same structural interpretation as a preceding musical stimulus (the prime). This facilitating effect on processing is a hallmark of structural priming in language (Kaschak, 2006). In turn, increased processing demands when the target and the prime do not share the same structural interpretation may impede performance of other competing cognitive tasks, such as memory encoding and retrieval. Accordingly, the present experiment relies on performance in a Flash Reaction Task and a Memory Task as implicit manifestations of structural priming effects. The two tasks were used to disambiguate between the two different types of structural abstraction (shallow and deep). In particular, as detailed in Section 5.2.4, the Flash Reaction Task was designed to be influenced by priming of shallow-structural representations but not deep-structural representations, whereas the latter could in principle be influenced by priming of either type of representation.

5.1.5 Aims, summary of the paradigm, and hypotheses

This study aims at investigating the cognitive reality of two forms of representation of musical structure (deep and shallow). In each trial, participants performed two different tasks, a Flash Reaction Task and a Memory Task. We hypothesised that structural priming would result in improved performance in both tasks. Namely, we expected shorter reaction times in the Flash Reaction task in the Congruent compared to the Incongruent priming condition; in the Memory task, consistently with prior evidence showing a positive correlation of structural encoding and memory performance (Cecchetti et al., 2021), we expected higher accuracy and more confident responses, as reflected by faster response times. The two tasks were implemented so as to disambiguate between priming effects due to shallow and deep structure, as detailed in Section 5.2.4. In particular, as summarised in Figure 5.3,

- any performance improvement in the Flash Reaction Task in Shallow-Structure trials as a function of the priming condition (Congruent vs. Incongruent) would be evidence for the psychological reality of shallow-structural representations;
- any performance improvement in the Memory Task in both Shallow-Structure and Deep-Structure trials as a function of the priming condition (Congruent vs. Incongruent) would be evidence for the psychological reality of deep-structural representations.

Note that any observable effect of structural priming is assumed to be contingent on the activation of structural representations. This, in turn, may be influenced by increased ex-



	SHALLOW ONLY		DEEP ONLY		BOTH	
	REACTION	MEMORY	REACTION	MEMORY	REACTION	MEMORY
SAME STEM	✓	✓	✗	✓	✓	✓
DIFFERENT STEM	✗	✗	✗	✓	✗	✓

Figure 5.3 – Schematic visualisation of the hypothesised effects and their interpretation. The existence of shallow-structural representations alone would be reflected in priming effects in both tasks, but in Shallow-Structure trials only (left). The existence of representations encoding deep-structural information alone would be reflected in priming effects in the Memory Task only (middle). Priming effects in both tasks would indicate the existence of representations encoding both shallow- and deep-structural information (right). Note that the flash reaction task, by design, only had the potential to be influenced by shallow-structural priming in Shallow-Structure trials: the absence of an effect of priming in the Reaction Task in Deep-Structure trials is a sanity-check for the interpretability of the results.

posure over the course of the experiment, as discussed in Section 5.1.2. As a consequence, the hypothesised difference between priming conditions (Congruent vs. Incongruent) may manifest itself only in late trials – earlier trials serving as an exposure phase – or as a difference in the rate of improvement over the course of the experiment (exposure effect), rather than as a difference in average performance. For this reason, we analysed the participants' performance over the course of the entire experiment.

5.2 Methods

5.2.1 Participants

Ninety-nine participants (mean age 34.0, $SD = 11.0, \min = 19, max = 69$) with self-reported normal or corrected-to-normal hearing were recruited through Prolific Academic to take part in an online experimental session. Participants were reimbursed with 12CHF for their participation. The study was approved by the IRB of the École Polytechnique Fédérale de Lausanne (HREC 044-2022) and was conducted in accordance with the declaration of Helsinki. The average degree of musical training, quantified by the corresponding subscale (ranging from 7 to 49) of the Goldsmiths Musical Sophistication Index (GoldMSI; Müllensiefen et al., 2014), was 28.3 ($SD = 8.4, \min = 10, \max = 48$). For comparison, Müllensiefen et al. (2014) reported an average score of 26.5 ($SD = 11.4$) over a large sample of the general population from English-speaking countries.

5.2.2 Stimuli

Stimuli comprised 24 isochronous tonal chord progressions arranged in six quadruples (labelled A-F). Summary information about the stimuli is displayed in Table 5.1. Scores as well as Roman-numeral (after Hentschel, Neuwirth, et al., 2021) and tree analyses (after Rohrmeier and Neuwirth, 2015) for all stimuli are available as Supplementary Material S1.

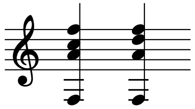
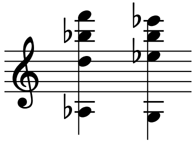
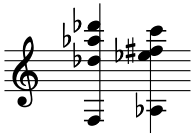
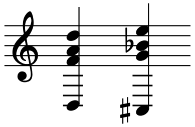
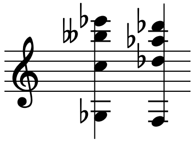
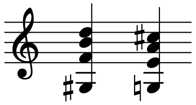
Each stimulus comprised an initial 2-chord stem followed by a 5-chord cadential continuation and a post-cadential ending of varying length (Figure 5.4, left), so that the overall length of a stimulus varied between 9 and 12 chords. All stimuli within each quadruple had identical initial stems as well as an identical final chord, but otherwise different cadential and post-cadential progressions. Furthermore, stimuli within each quadruple were divided into two types, depending on the deep structure of the segment comprised between the initial stem and the cadence: half of the stimuli in each quadruple exhibited deep structure of type V-I (cf. Figure 5.2, left), while the other half exhibited a deep structure of type I-I (cf. Figure 5.2, right). Stimuli in each type exhibited cadential progressions in the same key, which differed across the two types. As a consequence of the stimulus construction, the initial stem and the final chord had the same shallow-structural interpretation within each type in a given quadruple, yet different structural interpretation across the two stimulus types of the same quadruple (Table 5.1).

The keys associated with the two types were related by a different interval (from 1 to 6 semitones) in each quadruple, and the lengths of the four stimuli in each quadruple were counterbalanced across types and quadruples. Finally, stimuli were synthesised at 130bpm in the default piano timbre of the software MuseScore 3, resulting in an average nominal duration of 4.85s per stimulus (min = 4.15s, max = 5.54s). Audio files were normalised in loudness using the *pyloudnorm* package (implementing ITU-R BS.1770; Steinmetz and Reiss, 2021). After trimming the audio files at 1s after the nominal offset of the last chord, 10ms fade in and 500ms fade out were applied to smooth audio onset and offset.

5.2.3 Design

Quadruples were arranged in three pairs (A and F, B and E, C and D) and each participant was presented with all pairings of stimuli from two such pairs of quadruples. Specifically, participants were divided into three non-overlapping groups of 33 individuals each, and two pairs of quadruples were assigned to each group ($\{\{A, F\},\{B, E\}\}$, $\{\{A, F\},\{C, D\}\}$, or $\{\{B, E\},\{C, D\}\}$). In each trial of the experiment, the Prime and Target stimuli were drawn as one of the $\left|(X \cup Y)^2 \cup (W \cup Z)^2\right| = 128$ combinations of stimuli from either of the assigned pair of quadruples $\{X, Y\}$ and $\{W, Z\}$.

Table 5.1 – Summary of the stimuli adopted in the experiment. Each quadruple A-F comprised 4 stimuli of different lengths, further divided in two types depending on the shallow-structural interpretation of the stem and the final chord, as well as the deep-structural interpretation of the stimulus until the cadence.

Quadruple	Stem	Type	Deep Structure	Key	Shallow Structure (Stem)	Shallow Structure (Final Chord)	Length (#chords)
A		1	V – I	C	IV – ii^6	ii	9
							10
		2	I – I	F	I – vi^6	vi	11
							12
B		1	V – I	c	VII$^4_2$ – III^6	iv	9
							11
		2	I – I	E♭	V$^4_2$ – I^6	ii	10
							12
C		1	V – I	c	N^6 – Ger^6	VI	9
							12
		2	I – I	D♭	I^6 – V^7	V	10
							11
D		1	V – I	C	ii – vii° /ii	vii° /V	10
							11
		2	I – I	d	i – vii°	vii° /iv	9
							12
E		1	V – I	c	vii° /N^6 – N^6	Ger^6/iv	10
							12
		2	I – I	f♯	vii° /V – V^6	V^7	9
							11
F		1	V – I	C	vii° /V/ii – V^2/ii	IV^6	11
							12
		2	I – I	E	vii° /IV – IV^2	N^6	9
							10

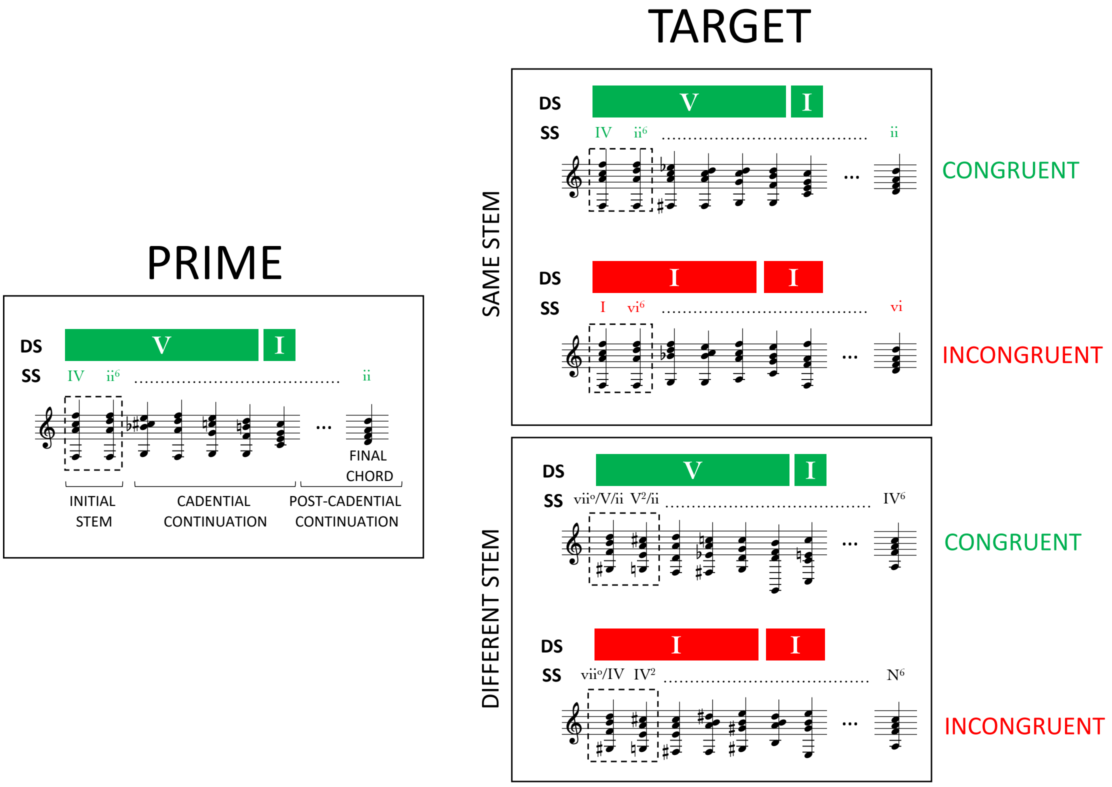


Figure 5.4 – (Left) Example of a stimulus comprising an initial stem, a cadential continuation, and a post-cadential continuation ending with a final chord. Shallow- (SS) and deep-structural (DS) interpretations are displayed above the stimulus. (Right) Example of four possible Targets for a given Prime. Some Targets (Shallow-Structure trials) share the same initial stem and final chord as the Prime, others do not (Deep-Structure trials). Among the Shallow-Structure Targets, some also share with the Prime the same shallow-structural interpretation (SS) of the initial stem and the final chord, as well as the same deep-structural interpretation (Congruent condition, green). Deep-Structure Targets differ from the Prime in terms of chord content and shallow-structural interpretation, yet some of them share the same deep-structural interpretation (DS) as the Prime (Congruent condition, green).

Over the course of the experiment, each Prime-Target combination was presented once, in random order. Possible pairings of Primes and Targets are exemplified in Figure 5.4 (right). In half of the trials (Shallow-Structure trials), Prime and Target belonged to the same quadruple, i.e., they shared the same initial stem as well as the same final chord. All other pairs of stimuli, the Deep-Structure trials, had different initial stem, different final chords, and different shallow-structural interpretation of both the initial stem and the final chord. Both the Shallow-Structure and the Deep-Structure trials were divided in two categories: those where Prime and Target shared some aspect of their structural representation (the "Congruent" priming condition) and those where they did not ("Incongruent"). In particular, in Shallow-Structure trials, the Congruent condition differed from the Incongruent condition because congruently-primed Prime and Target stimuli shared part of their deep-structural representation (either I-I or V-I) as well as the functional role, i.e., the shallow-structural representation, of both the initial stem and the final chord. Vice versa, in Deep-Structure trials, the Congruent condition

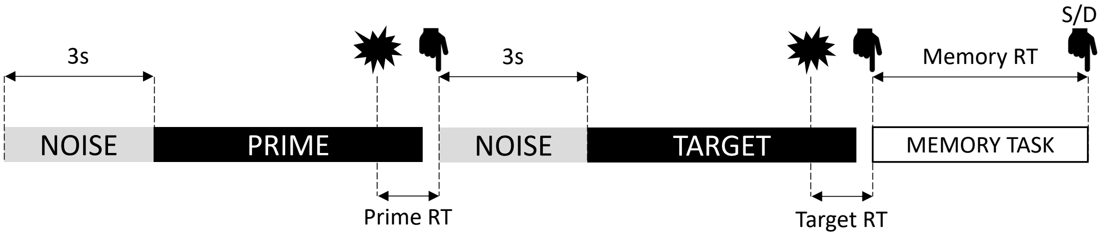


Figure 5.5 – Schematic unfolding of a trial. Four behavioural measures were collected: reaction times to the flash in the Prime and Target stimuli (Prime RT and Target RT), as well as a (S)ame/(D)ifferent memory response alongside its Response Time (Memory RT) at the end of the trial.

differed from the Incongruent condition because congruently matched Primes and Targets shared a deep-structural representation only. Both Shallow-Structure and Deep-Structure trials were equally divided between the Congruent and the Incongruent priming conditions.

Every combination of stimulus lengths occurred as many times with two stimuli of the same type as with two stimuli of different type, both in Shallow-Structure and in Deep-Structure trials. Note that in one half of the Congruently-primed Shallow-Structure trials (i.e., in one eight of the total trials) the Prime and Target stimuli were identical (except for a transposition): these trials served as a benchmark for assessing performance in a trivial case.

5.2.4 Experimental tasks

In any given trial, participants were presented with two stimuli, a Prime and a Target, interleaved by 3s of white noise (including 500ms fade in and 750ms fade out to reduce clipping and masking of the following stimulus). The Prime stimulus was presented in a random chromatic transposition (from −5 to +7 semitones relative to the template displayed in Table 1), and the Target stimulus was further transposed by an additional tritone, minimising the tonal relatedness of Prime and Target. This additional transposition, alongside the white-noise break, was meant to ensure that any priming would not be due to the effect of the Prime's key on the interpretation of the Target (similarly to previous probe-tone studies), but rather to the similarity of the structural relations within the Prime with those within the Target irrespectively of the tonal relatedness between Prime and Target. Participants were instructed to perform two tasks, a Flash Reaction task and a Memory Task, as illustrated in Figure 5.5.

Flash Reaction In the Flash Reaction task (Figure 5.6), participants were asked to react to a visual flash associated with the end of each chord progression. Participants were informed that the last chord of the Prime stimulus and of the Target stimulus in each trial would be marked by a visual flash, and they were instructed to react as quickly as possible to the occurrence of such flash by pressing the space bar with their right hand. The flash was presented as a white disk displayed in the centre of the screen for a duration of 100ms, starting 100ms before the nominal offset of the last chord of the progression (i.e., ((60s/130bpm) – 0.1s) after the

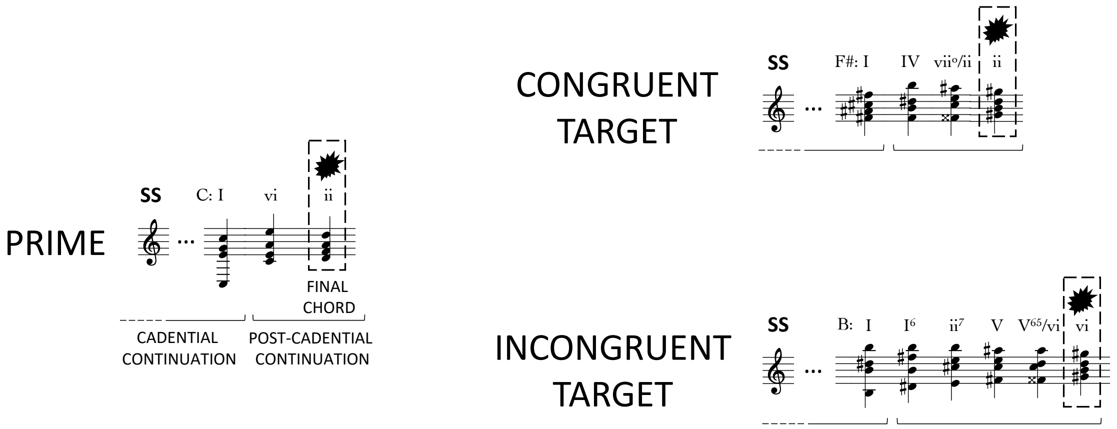


Figure 5.6 – Priming of shallow-structural interpretation of the final chord in Shallow-Structure trials. In the Congruent condition, the final chord of both the Prime and the Target stimulus shares the same functional role: in the example, both are supertonic (ii) in the key established by the preceding cadence. Vice versa, in the Incongruent condition, the final chord has a different shallow-structural interpretation in the Target (vi) relative to the Prime (ii). Once the Prime has primed listeners to a stimulus ending on chord ii, listeners may then find it easier to identify the final chord as such in the Congruent condition, where the stimulus also ends on chord ii, compared to the Incongruent condition, where it ends on a vi.

last chord's onset). The temporal placement of the flash was meant to give listeners some exposure to the last chord of a stimulus before being prompted to react, possibly increasing the capacity of the experimental manipulation to exert any influence on the performance of the reaction task. Reaction times were collected for both the Prime and the Target stimulus in each trial as the temporal distance of the key press from the onset of the flash.

Because of the counterbalancing of stimulus lengths (cf. Section 5.2.3), the location of the flash in the Prime stimulus did not allow participants to predict the location of the flash in the Target. In turn, any cues that would allow listeners to identify the final chord of the stimulus would also make the occurrence of the flash more expected, thus facilitating the reaction task. We hypothesised that, if the final chord of the Prime stimulus shared the same functional role as the final chord of the Target stimulus, the activation of the corresponding shallow-structural representation while listening to the Prime would facilitate the identification of the final chord in the Target, as listeners would be primed to hear a stimulus ending with that particular functional role. In all Shallow-Structure trials, the final chord of Prime and Target was identical under transposition, yet with either the same or different shallow-structural interpretation depending on the (Congruent or Incongruent) priming condition. Accordingly, we predicted shorter reaction times in the Congruent priming condition (where the final chord had the same functional role in both the Prime and Target stimuli) compared to the Incongruent priming condition. In Deep-Structure trials, instead, the shallow-structural interpretation of the last chord was different in both the Congruent and the Incongruent priming condition, so that no priming was expected.

Memory Task In each trial, after hearing both Prime and Target stimuli and performing the corresponding Flash Reaction tasks, participants were asked to remember whether the initial stem (i.e., the first two chords) of the Target stimulus was the same as or different from the initial stem of the Prime stimulus in that trial, irrespectively of any transposition. Responses were prompted by a textual instruction and participants responded with the left hand by pressing key ‘s’ for Same or ‘d’ for Different. Both the participant’s response (‘s’ or ‘d’) as well as the Response Time from the onset of the prompt were collected in each trial. Note that, in all Shallow-Structure trials, the correct response was Same (‘s’) irrespectively of the (Congruent or Incongruent) priming condition: in other words, the priming condition was irrelevant to the task. The same holds (*mutatis mutandis*) for Deep-Structure trials.

Performance in the Memory Task could be influenced by the structural similarity of Prime and Target. Specifically, the structural interpretation of the initial stem on its own was, in principle, ambiguous (Figure 5.7a): for instance, the two excerpts exemplified in Figure 5.2 differ both in terms of the shallow-structural interpretation of the stem as well as how this is embedded into a deep-structural interpretation. While the initial stem was ambiguous, the cadential continuation favoured one of the alternative interpretations, so that by the end of the cadence listeners would form a mental representation of the initial stem in accordance with such preferred interpretation (Figure 5.7b). If two such chord progressions are used as Prime and Target in a priming trial, the prior activation of one of the two plausible representations for the Prime’s stem may facilitate the activation of the same representation for the Target’s stem (which is identical – under transposition – to the Prime stem; Figure 5.7c).

Note that, because of the transposition, sensory information alone may not prime the preference for a structural representation over the other in the Target stimulus. Specifically, a representation of the Prime stimulus that is specific to a given key (i.e., tied to a tonic as a concrete pitch class) would not be a viable representation of the transposed Target stimulus, nor would the Target stimulus be plausibly interpretable in the key established by the Prime stimulus. It is rather the functional relationship between chords in the Prime stimulus – encoded in its abstract structural representation – that is preserved in the Target stem and may prime a certain hearing of the structural relationships between chords in the Target.

Crucially, the cadential continuation of the Target stimulus may induce an observable manifestation of such a priming effect by removing the structural ambiguity of the Target’s stem. Specifically, the Target’s cadential continuation may either confirm (in the Congruent priming condition) or conflict (in the Incongruent condition) with the interpretation favoured by the Prime stimulus (Figure 5.7d). In the Incongruent condition, then, listeners would need to suppress the primed representation and revise the structural representation of the Target, including its stem, according to a different interpretation. Such a phenomenon of retrospective revision, which has been proposed theoretically (Jackendoff, 1991) and recently observed experimentally in music (Cecchetti et al., 2022), is qualitatively analogous to linguistic garden-path effects (Frazier, 1978). In the Incongruent condition, the necessity to revise the interpretation induces increased cognitive load as well as a representational overdetermi-

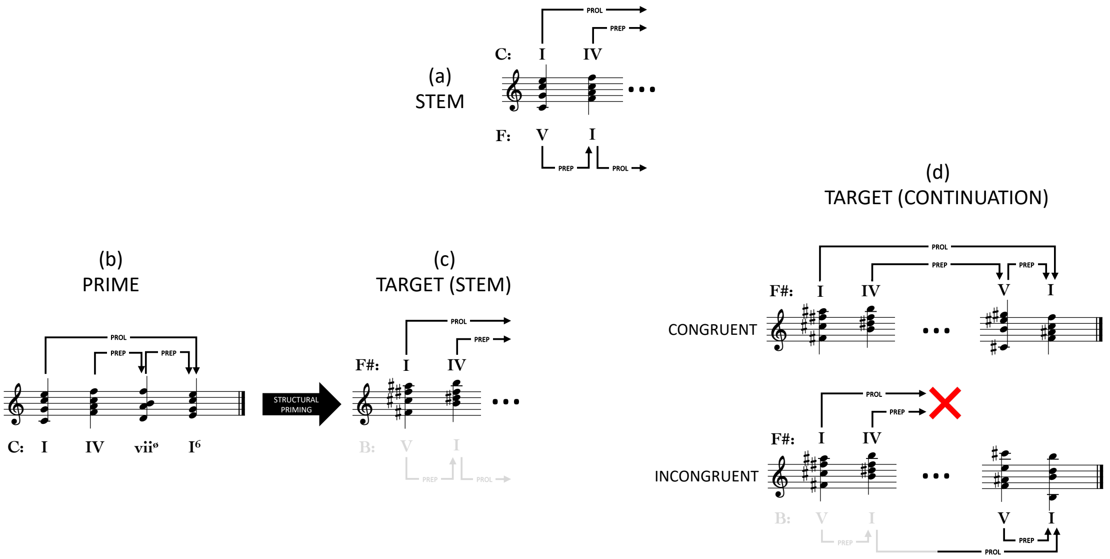


Figure 5.7 – Priming of the initial stem's interpretation. The initial stem (a) is ambiguous and affords (at least) two different interpretations, visualized above and below the score. When the initial stem is embedded in the Prime stimulus (b), one of the two interpretations is favoured over the other. After hearing the Prime stimulus, listeners may be more likely to favour the same interpretation of the stem when embedded at the beginning of the Target stimulus (c). However, the continuation of the Target stimulus (d) may either confirm (Congruent) or clash (Incongruent) with the primed interpretation. In particular, in the Incongruent condition, listeners may be forced to abandon the primed interpretation halfway through the Target stimulus, switching to the initially dispreferred alternative one. As part of the resulting garden-path effect, listeners may revise their interpretation of the initial stem retrospectively according to the new preferred interpretation.

nation associated with the temporary competition between two distinct representations of the ambiguous stem. Vice versa, the capacity to perform a concurrent cognitive task may be facilitated in the Congruent priming condition, where the representation of the Target stem is consistently the same as the one primed by the Prime stimulus, compared to Incongruent Targets, where the representation of the Target stem has to be revised. As a consequence, we predicted that, in the Incongruent condition, listeners would find it more difficult to identify the Target's stem as a transpositionally invariant replication of the previously presented Prime's stem, thus exhibiting worse performance in the Memory Task compared to the Congruent priming condition.

In Deep-Structure trials, no priming of shallow-structural representations was expected, as the interpretation of the initial stem was not shared in either the Congruent or the Incongruent priming conditions. Nevertheless, a priming effect due to deep-structural (dis)similarity may still be expected. Specifically, if listeners do form a representation of deep structure while listening to the Prime stimulus, activation of the same representation for the following Target stimulus may be facilitated compared to alternative ones. In the Incongruent priming conditions, where Prime and Target did not share the same deep-structural representation, listeners may have faced additional cognitive load for parsing the Target chord progression as a result of the necessity to suppress the primed representation in favour of an alternative one. In particular, the embedding of the initial stem into a deep-structural representation differs between the two types of deep structure we consider: the initial stem is either part of a tonic region (I) or of a dominant region (V) at the deep structural level. As a consequence, in the Incongruent priming condition, listeners primed to hear the initial stem as embedded into a tonic constituent would need to revise their interpretation in favour of a different one where the initial stem is embedded into a dominant constituent, and vice versa. Accordingly, we hypothesised that performance in the Memory Task would be facilitated in the Congruent priming condition compared to the Incongruent priming condition.

5.2.5 General Procedure

After granting informed consent, participants were shown detailed instructions for each experimental task, including examples to clarify the notion of identity under transpositional invariance. Specifically, participants were shown two families of chord progressions that were or were not transposed instances of one another. Before undertaking the actual experimental session, participants also familiarised themselves with the tasks with three training trials. The main experimental session comprised 128 trials, lasting approximately 50 minutes. The transition from one trial to the next was self-paced, and participants were encouraged to take a break at three equally-spaced moments during the experiment. After completing the main experimental session, participants filled in the Musical Training subscale of the Goldsmiths Music Sophistication Index (Müllensiefen et al., 2014). Overall, it took approximately 1 hour for participants to complete the entire experiment.

5.2.6 Analysis

Trials with identical Prime and Target stimuli (12.5% of the total trials) were used as a benchmark for the performance in the various tasks and excluded from the other analyses, as they represent a trivial manipulation were Prime and Target shared far more than an abstract structural representation. Data were then analysed with Bayesian mixed-effects models provided with weakly informative priors (t(3,0,1); Gelman et al., 2008) and implemented in the R package *brms* (Bürkner, 2018). We report evidence ratios (*Odds*) for the regression coefficients or some function of the regression coefficients to be strictly larger or smaller than zero. From a frequentist perspective, evidence ratios for one-sided hypotheses can be interpreted as significant (*) at a .05 confidence level when exceeding 19.

Models predicting each behavioural measure (Flash Reaction Times, Memory Accuracy, and Memory Response Times) were implemented as detailed in Sections 5.2.6. In each model, the main experimental manipulations were encoded as the interaction $Congruent \times Shallow \times TrialNumber$ where $x \times y := x + y + x : y$. For each trial, *Congruent* and *Shallow* are Boolean variables indicating whether the Prime and Target stimuli fell in the Congruent priming condition and/or constituted a Shallow-Structure trial, respectively (the reference level being

"False" for both variables). The interaction $Congruent \times Shallow$ characterised how behavioural performance was influenced by the congruency of the priming condition separately in Shallow-Structure and in Deep-Structure trials. The additional interaction with *TrialNumber* captured the possibility that such an effect may be moderated by the increased exposure to the stimulus materials and to the experimental task over the course of the experimental session, as hypothesised.

Based on our hypotheses (Section 5.1.5), we performed two kinds of hypothesis tests. First, we tested the performance difference between the Congruent and the Incongruent priming condition (Main Congruency Effect) across trials, quantifying the evidence for the hypothesis that performance in the Congruent condition would be better than in the Incongruent condition. Evidence supporting the main effect was notated as $Odds(Congruent \gtrless Incongruent|Shallow)$ for Shallow-Structure and Deep-Structure trials separately depending on the value of $Shallow$. Second, we tested the change of such differences over trials (Exposure Effect), quantifying evidence for the hypothesis that additional exposure would strengthen the priming effect and increase the difference between priming conditions. Such Exposure Effect was quantified as the interaction $Congruent : TrialNumber + Shallow : TrialNumber$ or $Congruent : TrialNumber$ in Shallow-Structure and Deep-Structure trials, respectively, with evidence

$$Odds(Congruent : TrialNumber \gtrless Incongruent : TrialNumber|Shallow).$$

Data and analyses are available as supplementary materials S2 at https://osf.io/yq8ku/?view_only=17d1dffd458f451480aa4671583ef430.

Flash Reaction Time Flash Reaction Times longer than 1500ms or shorter than 100ms were considered as outliers, consistently with studies were reaction times to visual stimuli were tested in conditions of high cognitive load or divided attention (e.g., Spence et al., 2001). Fixed outlier cut-offs tend to lead to conservative inferences by favouring Type II over Type I error (Berger and Kiefer, 2021). Only trials where both Flash Reaction Tasks were performed within the outlier thresholds were retained in the analysis.

The ratio of the reaction time in the Target to the reaction time in the Prime (*TargetToPrimeFlashRTRatio*) was computed for each trial, and was modelled as log-normally distributed (Ulrich and Miller, 1993) with mean

$$\begin{aligned}\text{TargetToPrimeFlashRTRatio} \sim\ & \text{Congruent} \times \text{Shallow} \times \text{TrialNumber}+\\ & + \text{PrimeLength} + \text{TargetLength}+\\ & + (1|\text{PrimeStimulus}) + \left(1\middle|\text{TargetStimulus}\right) +\\ & + \left(1 + \text{Congruent} \times \text{Shallow}\middle|\text{ParticipantID}\right).\end{aligned}$$

We hypothesised negative Main Congruency Effect (i.e., relatively faster reactions in Target stimuli due to the putative processing advantage associated with Congruent priming) and Exposure Effect (i.e., reactions in Congruent Shallow-Structure trials becoming faster over the course of the experiment at a higher rate compared to Incongruent Shallow-Structure trials).

The lengths of the Prime and Target stimuli were also included in the model to account for waiting time and foreperiod effects (Niemi and Näätänen, 1981). Random intercepts allow the model to account for intrinsic features of individual Prime or Target stimuli, as well as any response biases of individual participants. Additional random slopes by participant capture inter-participant variability in the sensitivity to the experimental manipulations (Congruency of the priming condition and sharedness of the same stem).

Memory Accuracy For the Memory Task, we excluded trials where the Response Time to the Memory Task exceeded 10s, approximately corresponding to the average joint duration of a Prime and a Target stimulus: this allowed for enough time for participants to ponder the answer (e.g., by mentally rehearsing the stimuli), while excluding extreme outliers.

We modelled the accuracy of responses without normalising for False Alarm Rate (i.e., wrong responses in the Deep-Structure trials), as participant-specific response biases were modelled by the corresponding random-effect terms separately for each experimental condition. We modelled the correctness of the memory response in any given trial (i.e., whether the response was "Same" in Shallow-Structure trials or "Different" in Deep-Structure trials) as a Bernoulli-distributed random variable with logit-transformed mean

$$\begin{aligned}\text{MemoryCorrect} \sim\ & \text{Congruent} \times \text{Shallow} \times \text{TrialNumber} + \text{PrimeLength} + \text{FlashRTPrime} + \\ & + \text{TargetLength} + \text{FlashRTTarget} + \\ & + (1|\text{PrimeStimulus}) + \left(1\middle|\text{TargetStimulus}\right) + \\ & + \left(1 + \text{Congruent} \times \text{Shallow}\middle|\text{ParticipantID}\right).\end{aligned}$$

Based on our hypotheses, we expected positive Main Congruency Effect (i.e., higher accuracy in the Congruent condition) and Exposure Effect (i.e., accuracy in Congruent trials improving over the course of the experiment at a higher rate compared to Incongruent trials). The random-effects structure was the same as in the previous model, with participant-specific response biases (e.g., increased false-alarm rate) being captured by the corresponding random intercept. The terms *PrimeLength*, *FlashRTPrime*, *TargetLength*, and *FlashRTTarget* quantified the timespan between the beginning of the Prime (i.e., the first exposure to the stem) up to the prompt of the Memory Task (cf. Figure 5.5).

Memory Response Time Response Times in the Memory Task were modelled as log-normally distributed with mean

$$\begin{aligned}\text{MemoryRT} \sim\ & \text{Congruent} \times \text{Shallow} \times \text{TrialNumber} + \text{PrimeLength} + \text{FlashRTPrime}+ \\ & + \text{TargetLength} + \text{FlashRTTarget}+ \\ & + (1|\text{PrimeStimulus}) + \left(1\middle|\text{TargetStimulus}\right) + \\ & + \left(1 + \text{Congruent} \times \text{Shallow}\middle|\text{ParticipantID}\right).\end{aligned}$$

Here, the hypothesised priming of shallow and deep structure would manifest itself as negative effects of the Congruent compared to the Incongruent condition in Shallow-Structure or Deep-Structure trials, respectively. Hypothesis tests were conducted as described for the Flash Reaction Time and the Memory Accuracy.

5.3 Results

5.3.1 Flash Reaction Time

A total of 342 trials (2.7%) were excluded from the analysis for containing reaction-time outliers in either the Prime or the Target stimulus, as described in Section 5.2.6. In the remaining trials, the median reaction time in Prime stimuli was 350ms (inter-quartile range 191ms), compared to 337ms (155ms) in the Target stimuli. Average reaction times to Target stimuli were faster in trials where the Prime and Target were identical (383ms) compared to all other trials (394ms), as supported by a paired t-test over participants ($t(98) = -2.597,\ p = .005$). This indicates that participants were not engaging randomly with the Flash Reaction Task, as it was influenced by the relatedness of Prime and Target.

In order to investigate the effect of the sharedness of shallow and deep structure between Prime and Target stimuli, and their dependency on the implicit discovery of structural features over trials, we analysed data with the model described in Section 5.2.6. Figure 5.8 shows posterior predictions for the Target-to-Prime ratio of reaction times across the different experimental conditions (Congruent and Incongruent priming, Shallow-Structure and Deep-Structure trials) as a function of the trial number.

Shallow-Structure trials. A significant difference between priming conditions was only found in very late trials ($\text{Odds}\left(Congruent < Incongruent|Shallow = True\right) > 19^*$ in the last 6% of the trials). Strong evidence was found that, over the course of the experiment, the difference in Target-to-Prime RT ratio was reduced (i.e., the Target's RT became faster in relation to its Prime's RT) at a greater rate in the Congruent priming condition compared to the Incongruent condition ($Odds\left(Congruent : TrialNumber < Incongruent : TrialNumber\middle| Shallow = True\right) = 65.67^*$). This is consistent with the hypothesis that increasing exposure progressively enabled the participant's performance to be influenced by the sharedness of

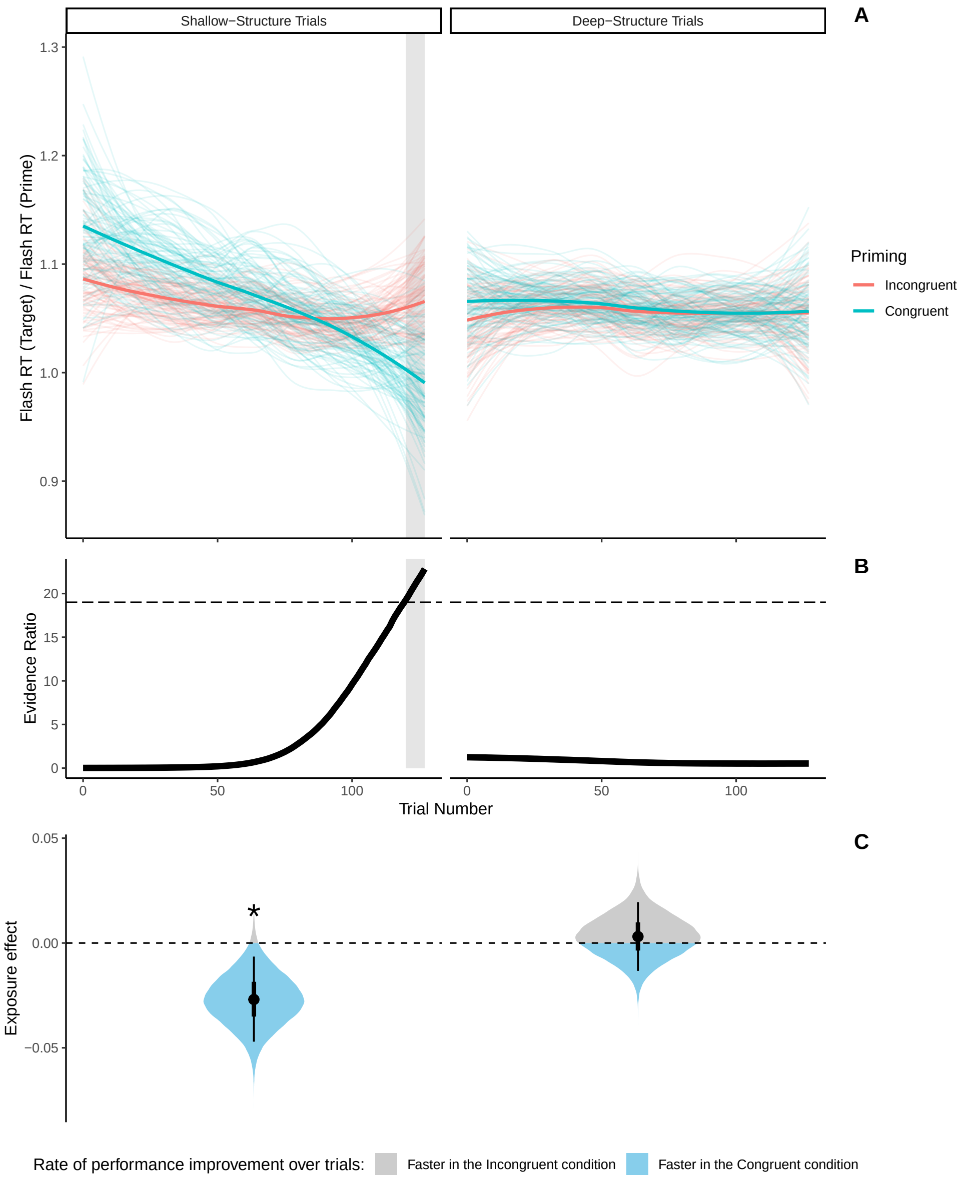


Figure 5.8 – (A) Posterior predictions for the flash reaction time in the Target stimulus in proportion to that in the Prime stimulus, expressed as a function of the priming condition and the trial number. For visualisation purposes, 100 sets of predictions drawn from the fitted posterior distribution (thin lines), as well as their average (thick line), are shown under locally estimated scatterplot smoothing (LOESS). The shaded areas (grey) indicate regions where RTs in the Congruent condition are significantly faster than in the Incongruent condition. In Shallow-Structure trials (left), reaction times in the Congruent condition (green) became faster over trials at a greater rate than in Incongruent trials (red). In Deep-Structure trials (right), no difference between priming conditions is apparent, as expected. (B) Evidence ratios in favour of the hypothesis that RTs in the Congruent condition are significantly faster than in the Incongruent condition. Grey shading indicates regions where the Odds exceed 19 (dashed line), corresponding to statistical significance at a .05 confidence level. (C) Posterior distribution of the Exposure Effect, quantified as the change over trials in the difference between the Congruent and the Incongruent condition. The significantly negative value in Shallow-Structure trials (*) indicates that the Flash Reaction Task was increasingly facilitated by Congruent priming over the course of the experiment.

shallow structural features between Shallow-Structure Prime and Target.

Since preliminary data exploration suggested that the two priming conditions may have not been indistinguishable in early trials (cf. Figure 5.8A), we also performed post-hoc test for the complementary hypothesis on the Main Congruency Effect. The Target-to-Prime RT ratio in the Congruent priming condition was found to be significantly larger than in the Incongruent condition in early trials (Odds$\left(Congruent > Incongruent \middle| Shallow = True\right) > 19^{*}$ in the first 21% of the trials). While this observation is consistent with the predicted trend, whereby Target RTs are favoured by congruent priming to an increasing extent in later trials compared to earlier ones, this also indicates that the expected Main Congruency Effect was reversed in early trials.

Deep-Structure trials. Since the final chords of Prime and Target stimuli from different Quadruples stood in no meaningful relationship relative to one another, no effect of priming condition was predicted in the Flash Reaction task in Deep-Structure trials. Consistently with the interpretation of the task, we found no evidence for a difference between priming conditions neither on average over trials nor as a different learning rate across trials (all $Odds < 2$).

5.3.2 Memory Accuracy

In analysing the Memory task, we excluded 24 responses (0.18%) that were given later than 10s after the prompt, as discussed in Section 5.2.6. The Memory task proved to be challenging for the participants, as evidenced by a high average false-alarm rate (49% of "Same" responses in trials where the correct response would have been "Different"). Nevertheless, the participants' accuracy was significantly higher than the false alarm rate, as supported by a paired t-test over participants ($t(98) = 8.091,\ p < .001$), indicating that the task was feasible with above-chance performance.

Figure 5.9 shows the average accuracy achieved in early and late trials as a function of the congruency of the priming condition. This exploratory visualisation suggests that the accuracy in the Memory Task was influenced by the priming condition. In order to evaluate statistical evidence for the effect of the sharedness of shallow and deep structure between Prime and Target stimuli, we analysed data with the model described in Section 5.2.6. Figure 5.10 shows posterior predictions for Memory Accuracy across the different experimental conditions (Congruent and Incongruent priming, Shallow-Structure and Deep-Structure trials) as a function of the trial number.

Shallow-Structure trials. We hypothesised that congruent priming of shallow structure would result in an improved memory performance, i.e., a positive effect of on Memory Accuracy. In the Congruent priming condition, Memory Accuracy improved over trials faster than

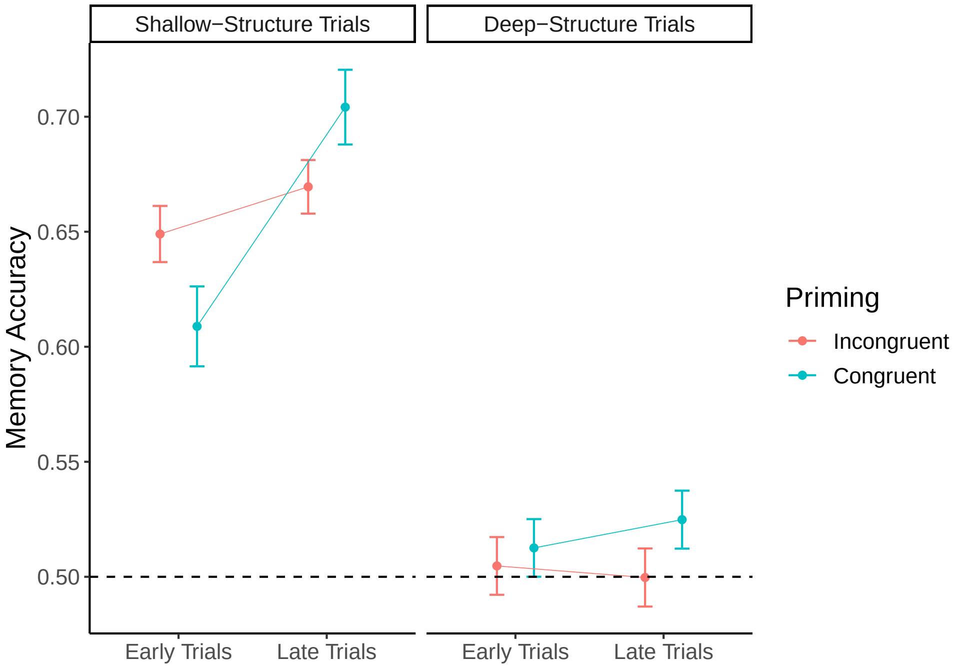


Figure 5.9 – Average accuracy (with 95% confidence intervals) in early (first half) and late (second half) trials, as a function of the experimental conditions. This summary visualisation of raw accuracy scores highlights improved performance in the Congruent priming condition relative to the Incongruent priming condition in later trials. Statistical support is discussed in the text and visualised in Figure 5.5.

in the Incongruent priming condition (Odds$\left(Congruent : TrialNumber > Incongruent : TrialNumber | Shallow = True\right) = 79.97^*$). In particular, participants were more accurate in the Congruent priming condition than in the Incongruent priming condition in late trials (Odds$\left(Congruent > Incongruent | Shallow = True\right) > 19^*$ in the last 14% of the trials), the opposite holding in early trials (Odds$\left(Congruent < Incongruent | Shallow = True\right) > 19^*$ in the first 20% of the trials). While this is consistent with the expected improvement of performance as exposure to the stimuli increased over the course of the experiment, the observation that performance in the Incongruent condition was higher than in the Congruent condition in early trials was unexpected.

Deep-Structure trials. We hypothesised that, if listeners formed deep-structural representations of the stimuli, structural priming in Deep-Structure trials would result in an improvement of Memory Accuracy. In Deep-Structure trials, Memory Accuracy was indeed significantly higher in the Congruent priming condition compared to the Incongruent priming condition in late trials (Odds$\left(Congruent > Incongruent | Shallow = False\right) > 19^*$ in the last 40% of the trials), although only little evidence for an overall linear divergence of learning trajectories across priming conditions was found (Odds$\left(Congruent : TrialNumber > Incongruent : TrialNumber | Shallow = False\right) = 7.78$).

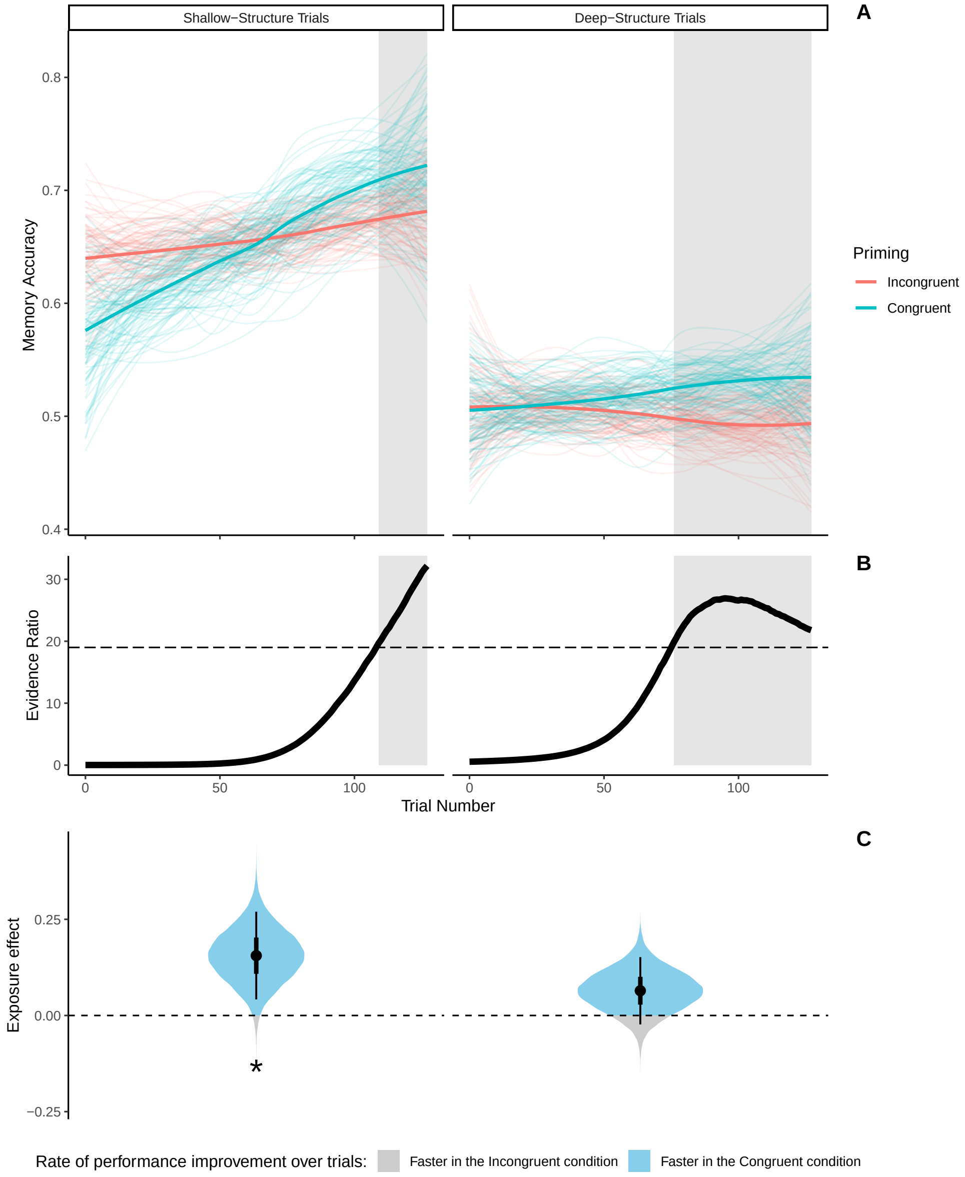


Figure 5.10 – (A) Posterior predictions for the accuracy in the Memory Task, expressed as a function of the priming condition and the trial number. For visualisation purposes, LOESS curves are shown for 100 sets of predictions drawn from the fitted posterior distribution (thin lines) as well as for their average (thick line). The shaded areas (grey) indicate trials where accuracy in the Congruent condition was significantly higher than in the Incongruent condition. In both Shallow-Structure (left) and Deep-Structure (right) trials, accuracy in the Congruent condition (green) became faster over trials at a greater rate than in Incongruent trials (red). This supports a priming effect due to both shallow- and deep-structural representations. (B) Evidence ratios in favour of the hypothesis that accuracy in the Congruent condition are significantly faster than in the Incongruent condition, as a function of the trial number. Grey shading indicates regions where the Odds exceed 19 (dashed line), corresponding to statistical significance at a .05 confidence level. (C) Posterior distribution of the Exposure Effect, quantified as the change over trials in the difference between the Congruent and the Incongruent condition. The significantly positive value in Shallow-Structure trials (*) indicates that accuracy in the Memory task was increasingly facilitated by Congruent priming over the course of the experiment.

5.3.3 Memory Response Time

Result for Memory Response Times are shown in Figure 5.11.

Shallow-Structure trials. We hypothesised that congruent priming of shallow structure would facilitate performance in the Memory Task in Shallow-Structure trials, resulting in faster responses at least as a function of increased exposure over trials. Supporting this hypothesis, response times were indeed shorter in the Congruent than in the Incongruent priming condition in late Shallow-Structure trials (Odds$\left(Congruent > Incongruent|Shallow = True\right) > 19^{*}$ in the last 54% of the trials). Little evidence for an overall linear divergence of learning trajectories across priming conditions was found (Odds$\left(Congruent : TrialNumber > Incongruent : TrialNumber|Shallow = True\right) = 6.98$).

Deep-Structure trials. We hypothesised that, if listeners formed deep-structural representations of the stimuli, structural priming in Deep-Structure trials would also result in facilitating the Memory Task (i.e., speeding up responses), at least in terms of an increased exposure effect over trials. Consistently with this hypothesis, we found evidence that responses became faster over trials to a greater rate in the Congruent condition than in the Incongruent condition (Odds$\left(Congruent : TrialNumber > Incongruent : TrialNumber\middle|Shallow = False\right) = 30.86^{*}$). However, unexpectedly, this did not correspond to response times being significantly faster in the Congruent condition than in the Incongruent condition in late Deep-Structure trials (Odds$\left(Congruent > Incongruent|Shallow = False\right) < 3$ for all trials), but rather to response times being faster in the Incongruent condition than in the Congruent condition in early Deep-Structure trials.

5.4 Discussion

In this study, we investigated the psychological reality of abstract representations of musical structure formed implicitly during listening. In particular, we focused on two levels of abstraction afforded by hierarchical accounts of harmonic structure: "shallow" structural representations, encoding the functional role of individual chords in the musical surface, and "deep" structural representations, encoding latent templates that may underlie the entire span of a chord progression. By analogy with linguistic structural priming, we hypothesised that the formation of a structural representation for a Prime stimulus may facilitate or impair the formation of a structural representation for a following Target stimulus depending on the (non-)sharedness of aspects of such representations. This, in turn, may impact performance on cognitive tasks such as the capacity to react to an overlapping stimulus or memory for the music itself, provided that listeners have enough exposure to the music for a structural representation to emerge.

Our results support that the sharedness of both shallow and deep structural representations

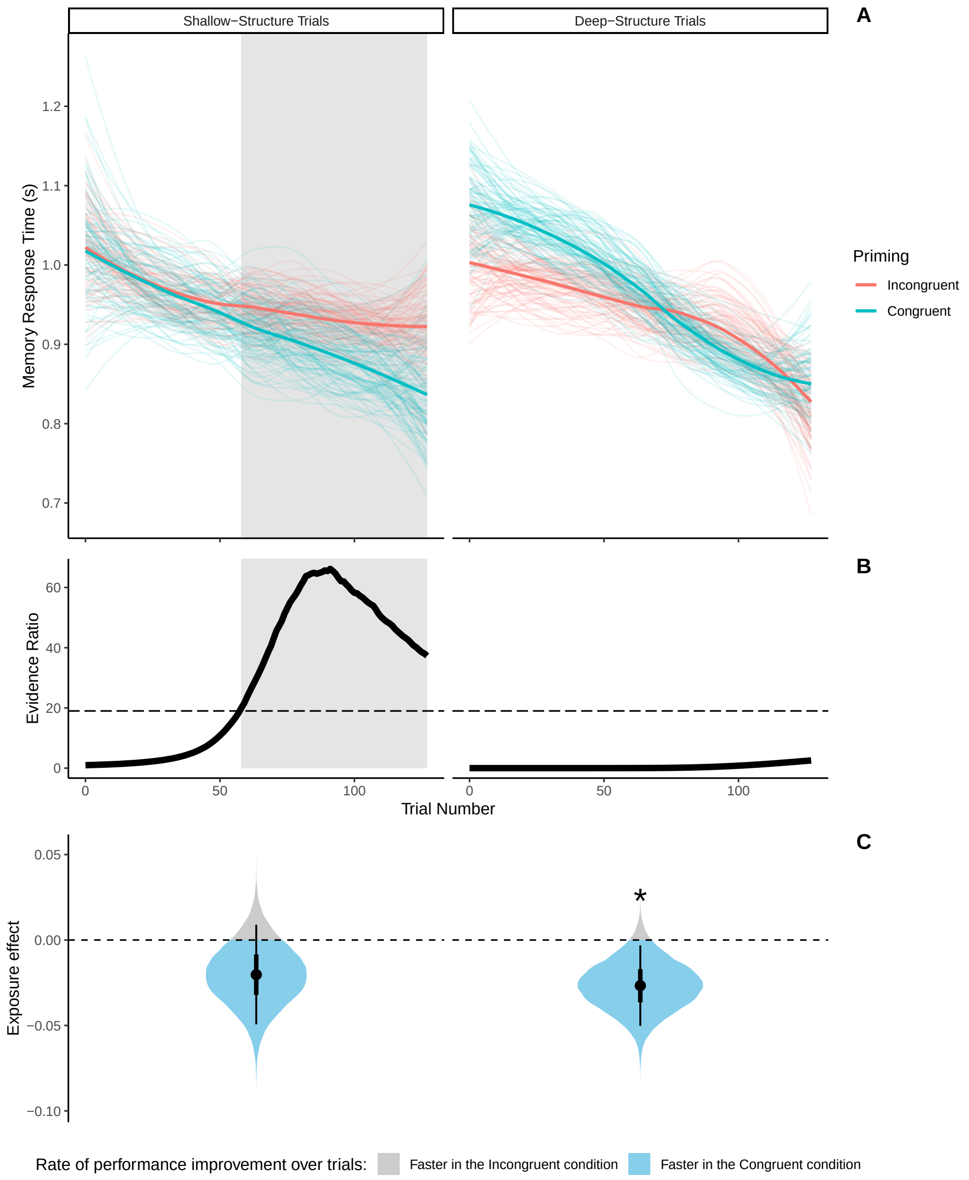


Figure 5.11 – (A) Posterior predictions for the response time in the Memory Task, expressed as a function of the priming condition and the trial number. For visualisation purposes, LOESS curves are shown for 100 sets of predictions drawn from the fitted posterior distribution (thin lines) as well as for their average (thick line). The shaded areas (grey) indicate trials where accuracy in the Congruent condition was significantly higher than in the Incongruent condition. In both Shallow-Structure (left) and Deep-Structure (right) trials, accuracy in the Congruent condition (green) became faster over trials at a greater rate than in Incongruent trials (red). This supports a priming effect due to both shallow- and deep-structural representations. (B) Evidence ratios in favour of the hypothesis that RTs in the Congruent condition were significantly faster than in the Incongruent condition, as a function of the trial number. Grey shading indicates regions where the Odds exceeded 19 (dashed line), corresponding to statistical significance at a .05 confidence level. (C) Posterior distribution of the Exposure Effect, quantified as the change over trials in the difference between the Congruent and the Incongruent condition. The significantly negative value in Deep-Structure trials (*) indicates that response times in the Memory task were increasingly facilitated by Congruent priming of deep structure over the course of the experiment.

impacted performance in two behavioural tasks, a Flash Reaction Task and a Memory Task. In particular, the observed differences between the Congruent and the Incongruent conditions in Shallow-Structure trials is consistent with the psychological reality of shallow-structural representations, while the observed differences in Deep-Structure trials are consistent with the psychological reality of deep-structural representations. The observation that both shallow and deep structural features are encoded in the listeners' representations is consistent with hierarchical accounts of harmonic structure, whereby structural relations are recursively nested (Rohrmeier, 2020b; Rohrmeier and Pearce, 2018; Steedman, 1984). These results thus support previous behavioural (Dibben, 1994; Herff, Harasim, et al., 2021; Serafine et al., 1989), neuroscientific (Cheung et al., 2018; Herff, Bonetti, et al., 2023; Koelsch et al., 2013; Ma et al., 2018a, 2018b), and computational (Harasim et al., 2020) studies in this direction with converging evidence. As a proof of existence, we focused on two extreme levels of the structural hierarchy, and future research may focus on targeting the entire recursive structure that is predicted by grammar-based models of music (Rohrmeier, 2020b). In this respect, results pertaining to the nature of structural representation, such as those acquired in priming paradigms (Branigan and Pickering, 2017), may complement those pertaining to the nature of grammar-based, cognitively-plausible parsing mechanisms (e.g., Cecchetti, Tomasini, et al., 2023) in refining our understanding of music processing.

It should be noted that an improvement of performance in the Congruent condition compared to the Incongruent condition was not observable in early trials. The difference between priming conditions rather manifested itself in late trials, or as an overall difference in the rate at which increased exposure to the stimuli resulted in an improved performance. Overall, our results suggest that the hypothesised priming effect due to the sharedness of structural representations requires listeners to become familiar with the musical stimuli. This may be interpreted as indicating that the formation or accessibility of structural representations does not systematically happen upon first-pass listening, and rather multiple exposures are needed for fine-grained structural representations to be fully activated in their entire complexity. As the mental representation of a given melody is refined to encode more and more structural information over multiple exposures, performance in Congruently-primed Targets increasingly benefits from the priming effect while performance in Incongruently-primed Targets remains comparatively unaffected. This may explain the different trends in Congruent and Incongruent trials from early to late trials.

Flash Reaction in Shallow-Structure trials, as well as Memory Accuracy in Shallow-Structure trials and Memory Response Times in Deep-Structure trials, did exhibit an observable difference between the Congruent and the Incongruent conditions even in early trials, yet in the opposite direction than predicted. Importantly, even if surprising with respect to the hypothesised directionality of the effect, these results still indicate that listeners were implicitly influenced by the sharedness of both shallow- and deep-structural features between Primes and Targets even in early trials, even though the hypothesised facilitating effect of structural priming towards the experimental tasks does not provide a mechanistic explanation for the totality of these observations.

As a speculative explanation for this finding, it is possible that two different mechanisms are involved in early and late trials. Specifically, in early trials – when familiarity with the stimuli are lower and representations are coarser and weakly activated – phenomena associated with online processing may have greater impact than those associated with representations themselves, as speculatively discussed in the following. We hypothesised that listeners would be primed to preserve aspects of the representation activated by the Prime stimulus when faced with parsing the Target stimulus. While, in the Congruent condition, the primed aspects of the representation were confirmed by the Target stimulus, in the Incongruent condition listeners would rather be forced to abandon the primed representation in favour of a different one at some point during the Target stimulus. The additional processing complexity introduced in Incongruent trials by this garden-path effect may have induced listeners to "pay more attention" to the Target stimulus, which in turn may have favoured performance in the behavioural tasks. On the contrary, the unproblematic processing of Targets in the Congruent condition may have not required the recruitment of additional attentional resources, resulting in comparatively worse performance relative to the Incongruent condition. In later trials, the (non-)sharedness of aspects of the final preferred representation for the Prime and Target stimuli may become more salient than the presence or absence of a processing disruption that leads to the activation of such representation. For example, it is possible that the parsing strategy changes altogether over the course of the experiment as listeners can increasingly exploit veridical memory of the individual stimuli as opposed to parsing them in real time. In particular, in later trials, listeners might delay the preference for or commitment to a particular representation until they can retrieve the "correct" representation upon recognising the specific Target stimulus. Since garden-path effects require the processor's commitment to a particular (misled) representation during online processing, the impact of such processing effects may then be reduced in late compared to early trials. Nevertheless, the final representation of the Target stimulus would still share or not share aspects of the Prime's representation, resulting in the priming effect as predicted.

A possible explanation[1] for the early-trial results in the flash-reaction task may also come from the observation that listeners tend to respond faster to locally unexpected rhythmic oddballs — thus reversing the standard predictive-attending facilitation effect (Large and Jones, 1999 — as a consequence of a reactive-attending listening strategy, whereby oddballs acquire greater saliency ("capture effect", Penel and Jones, 2005; cf. also Katz et al., 2015). Although this observation pertains to a different musical dimension, it is possible that our participants adopted a reactive-attending listening strategy at the beginning of the experiment, when stimuli where unfamiliar, while transitioning to a more predictive-attending strategy as the availability of a structural representation for each stimulus allowed for top-down predictions. Overall, the unexpected results in early trials may point to interesting processing phenomena and may warrant additional future research to investigate possible explanations such as the one proposed above.

[1] We thank an anonymous reviewer for this suggestion.

This study contributes to our understanding of musical structural representations, in a similar vein as linguistic structural-priming studies contribute to the understanding of syntactic representations in language (Branigan and Pickering, 2017). In particular, our results strongly support that structural information related to idiom-specific structural relations is abstracted from sensory information during listening. These results extend previous studies investigating explicit judgements of structural similarity or grammaticality, for example in implicit learning paradigms (Rohrmeier and Rebuschat, 2012). While evidence for implicit artificial-grammar learning supports the existence of a general capacity for acquiring the computational tools to process idiomatic structure through exposure, our study further supports the existence of representations with specific properties, as predicted by hierarchical music-theoretical accounts (e.g., Lerdahl and Jackendoff, 1983b; Rohrmeier, 2020b). Furthermore, the present study demonstrates that such representations can induce implicit effects on behaviour, besides explicit similarity, grammaticality, or liking judgements. Specifically, in our experimental design, all Shallow-Structure trials were associated with the same "correct" memory response irrespectively of the Congruent or Incongruent priming condition, as were all Deep-Structure trials. As a consequence, the priming condition was irrelevant towards the performance of the Memory Task. This supports the view that representations encoding both shallow- and deep-structural information abstracted from the musical surface are formed implicitly and spontaneously during listening as a result of implicit processing, rather than being the result of intentional analysis. This is consistent with evidence supporting structural accounts of harmonic processing, relying on representations of harmonic relations, over statistical or sensory accounts, relying on the frequency-based activation of pitch- or key-specific representations (Sears et al., 2023). This also complements recent evidence of cross-domain structural priming due to pitch structure in music-like tone sequences (Van de Cavey, 2016), by showing that conventional structuring principles typical of idiomatic music also give rise to representations that can induce priming within the musical domain.

5.5 Conclusion

The present study supports the psychological reality of structural representations for tonal harmony based on converging evidence from three behavioural measures. Such representations are shown to encode both (correlates of) the key-independent functional role of individual chords ("shallow" structure) as well as the way chords are embedded into more abstract templates ("deep" structure), while abstracting away merely sensory information. Specifically, our results show implicit effects of structural priming due to the sharedness of structural features, in analogy with the corresponding psycholinguistic phenomenon, and suggest that such representations are task-independent and formed spontaneously during listening. Building up on these results, future research may further investigate fine-grained, music-theoretically motivated features of idiomatic musical representations, similarly to how linguistic representations and their emergence through processing are investigated in psycholinguistics.

Processing Part III

6 Harmonic functions as syntactic categories

Abstract

Functional harmony is an integral part of many repertoires in the Western musical practices, including both diatonic and extended tonality. In the latter context, music-theoretical accounts suggest that the three octatonic equivalence classes (OECs) consisting of pitch-classes related by stacked minor-third intervals may be associated with tonic (**T**), dominant (**D**), and subdominant (**S**) functions. Whether this theoretical description of music is also relevant to the perception of music has not yet been tested empirically. In this study, 100 participants familiar with Western repertoires were presented with jazz chord progressions containing chord substitutions. When each stimulus had been played, participants predicted how many more chords they would have expected to hear before the progression reached a plausible conclusion. We computed the similarity of responses for pairs of stimuli containing different harmonic substitutions and modeled such similarity values based on different measures of harmonic relatedness between substitutions. Data show that the OEC membership of substitutions strongly predicts the similarity of participants' completion ratings. Bayesian mixed-effects modelling of similarity values further showed a categorical distinction between **D** and **S** as functional categories, on the one hand, and **T**, on the other. The data also appear to reflect the prevalent influence of rock and pop repertoires on the sample tested, encouraging further research into the influence of stylistic diversity and musical expertise. Overall, results contribute to the characterization of listeners' implicit knowledge of the principles of harmonic structure in extended tonality, and support the relevance of OECs not only as descriptors of extended-tonal compositional practices but also parsimonious predictors of perceived functionality.

This Chapter is published as: Cecchetti, G., Herff, S. A., Finkensiep, C., Harasim, D., & Rohrmeier, M. A. (2023). Hearing functional harmony in jazz: A perceptual study on music-theoretical accounts of extended tonality. *Musicae Scientiae*, **27**(3). GC, SAH, CF, and DH conceptualized the study. GC implemented the experiment, analysed the data, and wrote the manuscript with supervision by SAH and MAR. CF and DH contributed to the stimulus creation.

6.1 Introduction

6.1.1 A motivating example

A core feature of tonal harmony is that musicians may realize analogous harmonic contexts with different chords, that is, that different chords may *substitute* for one another. Figure 6.1A, for example, shows the normative cadential resolution of a chord rooted on the fifth scale degree (V) towards the tonic I. This is a common chord progression in tonal music, in which V serves the function of setting up the expectation of resolving towards I (Piston, 1948). Panels B, C, and D in Figure 6.1 show three examples in which chords rooted on different scale degrees other than the fifth (here III, ♭II, and ♭VII, respectively) are used by composers to prepare a resolution towards the local tonal center I; all these chords may be conceived as being mutually substitutable in expressing the same preparatory function towards the I as the V.

It is within the scope of music-theoretical accounts to identify the chords that are mutually substitutable in a certain musical style. In particular, theories of *extended* or *chromatic* tonality focus on characterizing several repertoires in the Western musical tradition that go beyond major/minor diatonicism, for example by exploring the entire chromatic space or by mixing different diatonic modes (Haas, 2004; Rohrmeier and Moss, 2021; Tymoczko, 2011). Such repertoires include subsets of classical, film, jazz, rock and pop music (Capuzzo, 2004; Heine, 2018; Rohrmeier, 2020b). In jazz, for example, substitutions are essential to improvisation, and musicians are explicitly trained to express their creativity by choosing different equivalent harmonizations (Levine, 1995).

These accounts represent music-theoretical insights into the way harmony is deployed in compositional practice. However, in the vein of understanding music cognition as a convergent research program bridging music-theoretical, computational and psychological approaches

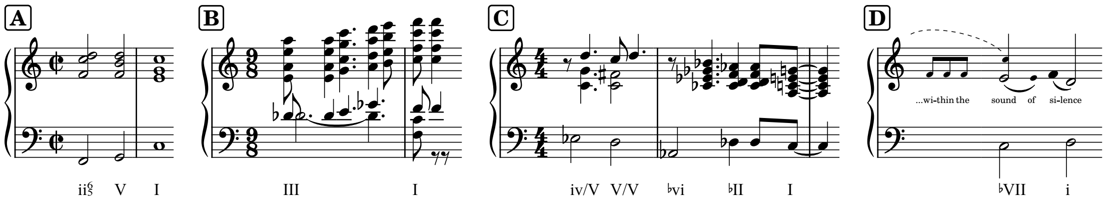


Figure 6.1 – Each panel shows the use of a different chord to elicit the expectation of a resolution towards a tonic (I) in extended-tonal repertoires. In panel **A**, a standard cadential progression is shown, where the final I is prepared by a dominant V. Instances of this progression are ubiquitous in Western classical music. In panel **B**, a harmonic sonority rooted on pitch-class A (locally, III) resolves towards the local tonic F in mm. 174ff. of Bartók's Divertimento Sz113 (Lendvai, 1971). Panel **C** shows the tritone substitution of the cadential ii-V progression, which then becomes ♭vi-♭II, from the jazz standard Satin Doll by Duke Ellington (Biamonte, 2008). Finally, the analytical reduction of the final bars of Simon and Garfunkel's *The Sound of Silence* in panel **D** (Everett, 2004) highlights the structural resolution of a harmony rooted on the Dorian (♭)VII towards the concluding tonic.

(Huron, 2006; Pearce and Rohrmeier, 2012), the extent to which music-theoretical insights also represent a parsimonious and accurate characterization of listeners' perception of harmonic functions and substitutions in extended tonality remains an open empirical question. In this article we address this question using an experimental design inspired by the idiomatic use of harmonic substitutions in jazz.

6.1.2 The functional syntax of tonal harmony

Theories of diatonic harmony in common-practice Western tonality typically share the view that harmonic entities, such as the degrees of the scale and the chords built on them, can be assigned some harmonic function (e.g., tonic, dominant, or subdominant) within a tonal context (Agmon, 1995; Lester, 1982; Meeus, 2000; Piston, 1948; Riemann, 1893). The temporal organization of functional harmonic progressions can then be described based on two principles (Agmon, 1995; Rohrmeier, 2011):

(1) On an abstract level, functions stand in some implication-realization relationship to one another; for example, the dominant function elicits expectations towards (or *prepares*; Rohrmeier, 2020b) the tonic function, which in turn can resolve (or *discharge*; Harrison, 1994; K. M. Smith, 2020) expectations from the dominant function (authentically) or the subdominant function (plagally). Such implication-realization relationships can be chained recursively, hierarchically, and cyclically, so that the target of an implication can simultaneously function as the source of a new implication with a different target. For example, the dominant function in a subdominant-dominant-tonic progression resolves an implication set up by the subdominant and, at the same time, establishes a new implication towards the tonic.

(2) On the musical surface, each function can be fulfilled by several different chords that collectively form an equivalence class, whose representatives can substitute for one another in compositional practice.

In other words, harmonic functionality is held here to be characterized by (1) patterns of directed expectations, and (2) a classification of reciprocally substitutable chords based on relationships of functional equivalence.

Empirical studies based on musical corpora have shown harmonic functions to be an accurate and parsimonious way of categorizing chords for the purpose of characterizing common-practice repertoires (Anzuoni et al., 2021; Jacoby et al., 2015; Rohrmeier and Cross, 2008; White and Quinn, 2018). Crucially, a functional understanding of harmony also allows for clear predictions to be made as to the patterns of expectations it elicits in listeners, which numerous studies have tested in the context of diatonic tonality (J. Brown et al., 2021; Janata et al., 2002; Leino et al., 2007; Sears et al., 2019; Wall et al., 2020). However, features proper to extended tonality may also contribute to listeners' perception (Bisesi, 2017; Krumhansl, 1998; Milne and Holland, 2016). In particular, music-theoretical accounts identify functional uses of harmony common to extended-tonal compositional practices to various degrees (Doll, 2017; Everett, 2004; Haas, 2004; Harrison, 1994; Lendvai, 1971; McGowan, 2010; K. M. Smith, 2010, 2020). In the present study, we focus on a specific music-theoretical formalization of harmonic functionality in extended-tonal repertoires, presented in the next section, and investigate its perceptual reality in a sample of Western-enculturated listeners.

6.1.3 Octatonic equivalence classes as syntactic categories in chromatic harmony

Music-theoretical approaches such as the Riemannian theory of diatonic tonality (Riemann, 1893), neo-Riemannian theory (R. Cohn, 2012), and *Tonfeld* theory (Haas, 2004; Polth, 2018) of extended tonality in classical music, as well as functional theories for the functional aspects of non-classical repertoires (Doll, 2017; Everett, 2004; McGowan, 2010), identify recurring patterns in the ways that chords are employed in particular repertoires to express harmonic functionality. Specifically, some degrees of the scale may be more likely than others to be considered by composers and musicians as expressions of a given function, which then affords predictions in the form "a chord rooted on scale degree X is a viable instantiation of function Y."

In the case of extended tonality, such predictions can be characterized as geometric regularities over representations of pitch-class space such as the *Tonnetz* (Euler, 1773; Rohrmeier and Moss, 2021; Figure 6.2a). In particular, three octatonic equivalence classes (OEC) can be obtained by partitioning the chromatic space into collections of pitch-classes related by minor-third transposition. Converging analytical insights highlight how each OEC constitutes a set whose elements, when interpreted as chord roots or generally as chord tones, tend to express the same function in classical (Haas, 2004; Lendvai, 1971; Polth, 2006, 2011; K. M. Smith, 2010), jazz (Rohrmeier, 2020b) and possibly other extended-tonal repertoires (Rohrmeier and Moss, 2021). Once a global key is fixed, the 12 chromatic scale degrees, relative to the global tonal center, can then be divided into three classes **T**, **S**, and **D**, as shown in Figure 6.2b.

The labeling reflects the fact that, relative to the global key, implication-realization relationships linking the tonic, subdominant and dominant functions can be generalized in terms of **T**, **S,** and **D** respectively. The OEC **D**, for example, contains the dominant of the relative minor key, III, as well as the so-called backdoor (♭VII) and tritone (♭II) substitutes of the global dominant V; as illustrated in Figure 6.1 and discussed above, these are all viable expressions of the dominant function as preparation for a tonic in extended-tonal repertoires. Listeners who are exposed to music organized according to these principles may then perceive harmonic movement across OECs as movement across functions, and movement within OECs as side-steps without functional change. In other words, different representatives of the same class may be understood as reciprocal substitutes expressing the same function. The similarity of probe-tone profiles for octatonic scalar contexts and triadic contexts rooted a fifth apart constitutes preliminary perceptual evidence in this direction (Krumhansl, 1990).

It should be noted that the functional logic inherited by diatonic tonality is not the only or

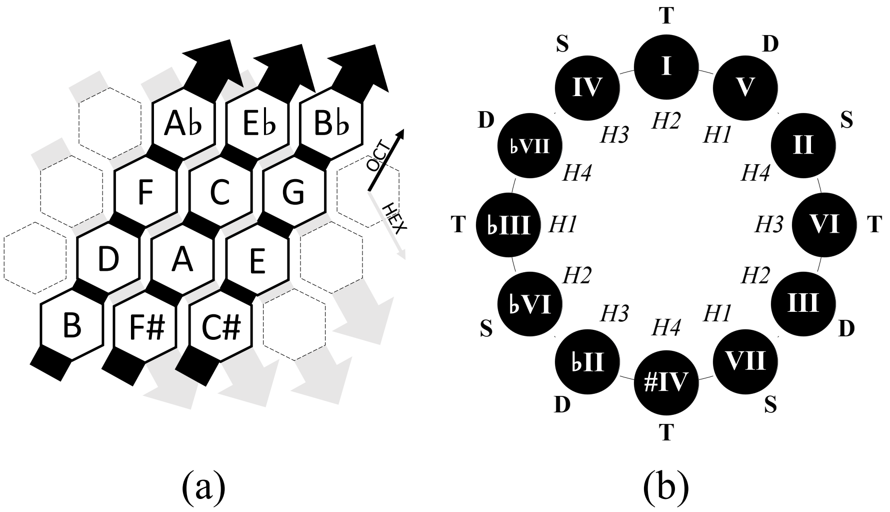


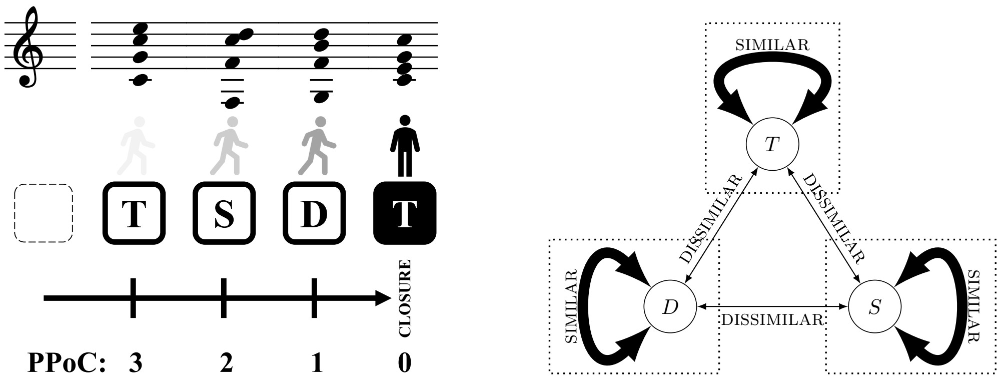
Figure 6.2 – Harmonic equivalence classes on Euler's *Tonnetz* Euler, 1773. (a) Different dimensions of Euler's Tonnetz correspond to different relationships of harmonic relatedness: highlighted in the figure are octatonic (dark arrows) and hexatonic (light arrows) relatedness. Motion along one of these dimensions can be interpreted as motion inside a single Octatonic Equivalence Class or Hexatonic Equivalence Class, respectively. (b) The twelve chromatic scale degrees (under enharmonic equivalence) arranged around the circle of fifths. On the outside of the circle, the membership of each scale degree to the T-, D-, or S-functioning Octatonic Equivalence Class is shown. On the inside of the circle, the membership of each scale degree to one of the four Hexatonic Equivalence Classes H1-4 is shown.

main structuring principle in extended-tonal idioms. In *Tonfeld* theory, for example, the above-mentioned OECs generated by stacking minor thirds and imbued with functional meaning coexist with two other types of tonal organization generated by stacking fifths or major thirds (Haas, 2004; Schiltknecht, 2011), the latter broadly related to the hexatonic collections characterized by neo-Riemannian theory (R. Cohn, 2012). Like OECs, Hexatonic Equivalence Classes (HEC, numbered here as *H1-H4* as per Figure 6.2b, inner circle) also constitute distinct classes of potentially substitutable chords (R. Cohn, 2012). However, while hexatonic collections have also been interpreted functionally (R. Cohn, 1999), movement across hexatonic collections typically conveys a sense of "the uncanny" (R. Cohn, 2007, p. 230) and contrast rather than the creation and resolution of goal-directed expectancy (R. Cohn, 2007; K. M. Smith, 2020). In other words, with respect to the definition of harmonic functionality given above, HECs are expected to satisfy condition (2), at least to some degree, but not condition (1); that is, substitutability but not the capacity to induce expectancy. While this study specifically targets the perceptual reality of harmonic functionality as modelled by octatonic equivalence, we also test HECs as a plausible and widely investigated alternative characterization of extended-tonal harmony. Our hypothesis is that HEC-membership is not a better predictor than OEC-membership of listeners' responses in a task that relies on the perception of harmonic function, such as the one described in the following section.

6.1.4 Aims and hypotheses

For the purpose of our experiment, we intended to target classes of functionally substitutable chords selectively, rather than other types of pitch-space structures. Based on the definition given at the outset, an experimental paradigm aiming to investigate harmonic functionality might exploit goal-directed expectations as proxies for functional hearing, and test for similarity in expectancy as a proxy for functional equivalence. In particular, we focused on the defining feature of functionality, which is the syntactic relatedness of functions in terms of their potential to set up patterns of harmonic expectancy (Huron, 2006; Rohrmeier, 2013), or, to put it another way, cadential resolution as captured in condition (1) above. Functions resolve into one another locally and, in particular, cadential resolution into the global tonic is a marker of global harmonic closure (Rohrmeier and Neuwirth, 2015). Thus, representatives of different OECs were hypothesized to differ with respect to the expectations for closure they elicit. For example, chords in **D** can resolve directly into chords in **T**, chords in **S** can resolve into chords in **D** or, plagally, into chords in **T**, while chords in **T** cannot resolve into other chords in **T**. If chords in **T** do not mark global closure themselves, they require additional progressions of subdominant- and dominant-functioning chords before closure can be achieved. Note that, due to the cyclic nature of functional relations, **T** then subsumes two distinct harmonic functions, as a marker of harmonic closure as well as preparation for **S.** Overall, functional syntactic organization allows listeners to orient themselves with respect to the perceived proximity of harmonic closure (PPoC; cf. Herff, Cecchetti, et al., 2021), as cadential closure may be predicted to occur nearer or farther in the future depending on the functional status of the current harmonic context (Figure 6.3a). As illustrated in Figure 6.3b, PPoC elicited by contexts sharing the same functional status may then be expected to be relatively more similar to each other than predictions elicited by contexts expressing different functional status.

Previous research has shown that listeners' PPoC is predicted by computational models of structural organization (Herff, Cecchetti, et al., 2021), suggesting that PPoC reflects an implicit knowledge of harmonic relationships (Rohrmeier et al., 2012; Tillmann, 2005). In this study, we replicated the experimental paradigm proposed by Herff, Cecchetti, et al. (2021) and investigated how such implicit knowledge, as reflected by PPoC, relates to a music-theoretical formalization of extended tonality. Specifically, we presented potentially interrupted chord progressions and tested listeners' predictions as to how imminent harmonic closure would be. We assumed that such predictions reflect the functional status of the harmonic context at the time the prediction is made. Drawing inspiration from the practice of chord substitutions, which are idiomatic in jazz as illustrated in Figure 6.2b (Levine, 1995), we manipulated harmonic context systematically with chromatic transpositions of penultimate and pre-penultimate events in chord progressions (cf. Figure 6.3a). Finally, we tested whether OECs, as music-theoretically motivated markers of functional harmonic status in extended tonality, represent parsimonious and accurate predictors of participants' expectations compared to other putatively non-functional characterizations of diatonic (i.e., distance on the circle of fifths) and chromatic harmonic relatedness (i.e., chroma distance, HEC membership).



(a)

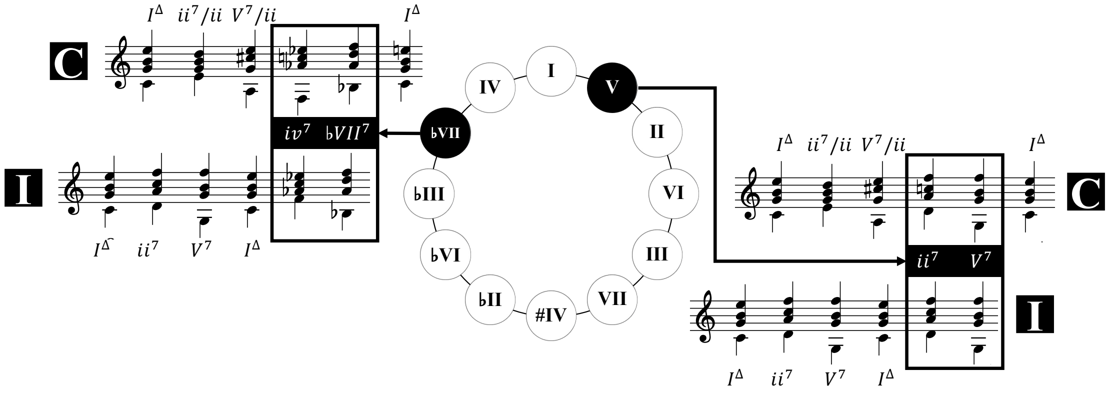
(b)

Figure 6.3 – **(a)** Example of cadential approach to harmonic closure. Each step preceding the achievement of closure instantiates a different function, and listeners may to some extent infer how distant harmonic closure is based on the function of the current harmonic context. The quantitative estimate of such distance, in terms of the number of events missing until closure is achieved, is termed here perceived proximity of closure (PPoC). **(b)** Schematic visualization of the expected similarity of PPoCs depending on the functional status of the harmonic context. Thick edges connecting a function with itself indicate that contexts expressing the same function are expected to induce mutually similar PPoCs. Thin edges across functions indicate that contexts expressing different functions are expected to elicit different PPoCs. Overall, observing this pattern of similarity and dissimilarity among members of the classes corresponds to the theoretical classification into three mutually distinct classes, as highlighted by the dotted boxes.

6.2 Methods

6.2.1 Participants

One hundred participants (mean age 27.33, $SD = 7.63$, $min = 18$, $\max = 61$) took part in an online experimental session. The mean musical training score across the sample, assessed using the Musical Training subscale of the Goldsmiths Music Sophistication Index (Gold-MSI) (Müllensiefen et al., 2014), was 16.96 (*SD*=7.63, range 7-45). Comparison with the mean musical training score (26.52, *SD*=11.44) reported by Müllensiefen et al. for a large sample of a population recruited online across mainly Western-enculturated, English-speaking countries suggests that the participants in the current study were mostly non-musicians, having received little to no explicit tuition in music theory. Since this study set out to investigate music-theoretical constructs in Western musical idioms, it is also relevant to mention that 79 participants reported spending their formative years in Europe or North America, and all participants reported musical preferences for styles linked to some degree to Western musical practices (e.g., rock, pop, jazz, blues, hip-hop, classical). Participants were reimbursed with CHF 7.50 for their participation. The study was granted ethics approval by the IRB of the École Polythechnique Fédérale de Lausanne (HREC 049-2021) and was conducted in accordance with the Declaration of Helsinki.



C
I
$I^\triangle$ ii^7/ii V^7/ii $I^\triangle$
iv^7 $\flat VII^7$
$I^\triangle$ ii^7 V^7 $I^\triangle$
I
IV
V
♭VII
II
♭III
VI
♭VI
III
♭II
#IV
VII
$I^\triangle$ ii^7/ii V^7/ii $I^\triangle$
C
ii^7 V^7
I
$I^\triangle$ ii^7 V^7 $I^\triangle$

Figure 6.4 – Examples of Complete (C) and Incomplete (I) stimulus cores for two different substitutions. On the right, the original 6-chord core progressions are shown, where the ii^7 – V^7 block is transposed by zero semitones. On the left, the ii^7 – V^7 block is transposed by 3 semitones, so that the dominant V is replaced by its backdoor substitute ♭VII.

6.2.2 Stimuli

Stimuli for this experiment comprised 24 jazz chord progressions. The jazz style was adopted because explicit chord substitutions are idiomatic in jazz improvisational practice (Levine, 1995), offering a natural template for our experimental manipulation as detailed below. Furthermore, jazz harmony shares characteristics with both classical as well as rock/pop traditions (McGowan, 2010; Rohrmeier, 2020b), potentially resonating with the implicit harmonic familiarity of a variety of Western listeners. In order to clarify the stylistic context, chords were realized in 4-part voicings, as is idiomatic in jazz (Levine, 1989; McGowan, 2011). Accordingly, in the following, the triangle $\triangle$ indicates a root-position triad with an additional major seventh, the apex 7 indicates a triad with an additional minor seventh, and the notation X/Y indicates scale degree X relative to the key of scale degree Y (e.g., the applied dominant of ii is V/ii).

Each chord progression comprised an initial 8-chord introduction sequence and a 6-chord core, the latter falling into either the Complete or Incomplete category as illustrated in Figure 6.4 and explained in detail below. The introductory sequence comprised a repeated $I^\triangle - V^7$ oscillation introduced by a fade-in, which served the purpose of firmly establishing a major global key while blurring metricality.

Complete stimuli. The core of each Complete chord progression was obtained starting with a I$^\triangle$ – ii^7/ii – V^7/ii – ii^7 – V^7 – I$^\triangle$ progression, which achieves satisfactory harmonic closure according to Western music theory, and replacing the ii^7 – V^7 constituent with one of its 12 chromatic transpositions. We replaced two chords, rather than just one, to ensure that listeners would not rule out the substitution as a harmonic oddity, as opposed to integrating the transposed material, locally at least, into an incrementally constructed interpretation. Note that, because of these replacements, stimuli in the Complete category were not necessarily all expected to be perceived as complete, even though they all ended on the global tonic. This is only the case if the entire progression that precedes the final tonic, and in particular the penultimate chord, is interpretable as a viable preparation for harmonic closure (Rohrmeier and Neuwirth, 2015).

Incomplete stimuli. The cores of Incomplete chord progressions were obtained starting with a $\mathrm{I}^{\triangle} - \mathrm{ii}^7 - \mathrm{V}^7 - \mathrm{I} - \mathrm{ii}^7 - \mathrm{V}^7$ progression, which does not achieve harmonic closure due to the missing global tonic at the end, in which the final $\mathrm{ii}^7 - \mathrm{V}^7$ constituent was replaced by one of its 12 chromatic transpositions. Thus all Incomplete stimuli ended with a dominant-seventh chord. Since such a chordal form is typically associated with inducing rather than resolving expectations, these stimuli were not expected to be perceived as complete. Nevertheless, different substitutions elicited expectations towards different targets, resulting in different PPoCs depending on the functional status of the targets.

The voicing of the transposed block in Complete and Incomplete stimuli was adjusted for better fit to the chords that preceded and followed it. We term the transposed $\mathrm{ii}^7 - \mathrm{V}^7$ constituent in a chord progression the *substitution* for that stimulus, and label each substitution with the scale degree (relative to the global key) of its second chord (the dominant-seventh chord); for example, the identity substitution transposed by zero semitones is labeled V because its dominant-seventh chord is V^7, while the substitution corresponding to a transposition by three semitones upwards is labeled ♭VII because its dominant-seventh chord is $\flat\mathrm{VII}^7$. Each OEC is represented by four such substitutions, hence by four Complete and four Incomplete stimuli. For convenience, we also label each Complete or Incomplete progression with the label of the substitution it contains. Stimuli were rendered in MuseScore 3 with piano timbre at 80 bpm, after which the loudness of the audio files was normalized using the *pyloudnorm* package (Steinmetz and Reiss, 2021), which provides an implementation of the ITU-R BS.1770 recommendation for assessing the perceptual loudness of audio signals. Finally, the fade-in was applied using a custom Python script. All stimuli are available in Supplementary Material S1.

6.2.3 Experimental task

In each trial of the main experimental task, one of the Complete or Incomplete chord progressions was presented in a random transposition from −4 to +7 semitones relative to C major. In a replication of the task described by Herff, Harasim, et al. (2021), participants were told that each stimulus represented the potentially interrupted concluding section of a song, and asked to estimate how many more chords they would have expected to come before the end of the song. We interpreted this estimate as a measure of the PPoC (cf. Figure 6.3a). Participants clicked a mouse to select an integer value between 0 (meaning that they perceived the chord progression to be complete as presented) and 3 (meaning that they expected three chords to be missing for the chord progression to be complete), presented on screen as a horizontal array of labeled buttons. After recording the PPoC, participants were asked to report how confident they were about their estimate by selecting a value between 0 (not confident) to 100 (fully confident) on a quasi-continuous horizontal rating scale.

6.2.4 Procedure

The experiment was administered online. The user interface for the main experimental task was implemented in PsychoPy 3 (Gallant and Libben, 2019; Peirce et al., 2019) and hosted on the online repository Pavlovia.org (Bridges et al., 2020). At the beginning of the experimental session, participants were shown an informed-consent form and proceeded to the instructions after confirming consent. Instructions for the main experimental task included a tutorial trial using a stimulus that was not part of the materials described in Section 6.2.2. Over the course of the session, four presentations of each chord progression were arranged in random order, resulting in a total of 96 trials interleaved by 3s of white noise to mitigate any carry-over effects of key across trials. After completing the main behavioral task (lasting ~40 min), participants filled in the Gold-MSI questionnaire (~8 min).

6.2.5 Analyses

Individual similarity and joint entropy.

Since we expected substitutions sharing the same function to elicit similar PPoCs, we quantified their similarity by defining the individual similarity between two stimuli (IND) for each participant as the proportion of identical[1] PPoC values estimated by each participant across the different presentations of the two stimuli. Specifically, if $\mathbf{a}^{(i)}$ and $\mathbf{b}^{(i)}$ are the multisets containing the four PPoC values estimated by participant i across the four presentations of stimuli a and b, respectively, then

$$\mathrm{IND}_i\,(a,b) = \frac{1}{16}\sum_{(\alpha,\beta)\in\mathbf{a}^{(i)}\times\mathbf{b}^{(i)}} \delta_{\alpha\beta}, \tag{6.1}$$

where δ is Kronecker's delta ($\delta_{\alpha\beta} = 1 \iff \alpha = \beta, \delta_{\alpha\beta} = 0 \iff \alpha \neq \beta$). The IND score quantifies the extent to which a given listener perceived two stimuli as being similar for the purpose of estimating a continuation towards global closure. For example, a pair of stimuli with the same completeness status (either Complete or Incomplete) and substitutions belonging to the same functional class may be expected to elicit high IND values. By contrast, a pair

[1] We adopted a conservative criterion to quantify similarity as a categorical binary variable, whereby non-identical PPoCs are considered maximally dissimilar to one another irrespective of their numerical difference. While alternative criteria may be adopted, for example a discrete or continuous similarity metric, it is unclear what properties would be required of such metrics in this context. In particular, it is questionable whether such a numerical metric should be linear (e.g., the dissimilarity between PPoC 0 and 1 may be much larger than between 2 and 3 because of a ceiling effect) or even monotonic (e.g., PPoC 3 and 1 may be more similar to one another than 2 and 1 because of metricality). As a consequence, identity is the only similarity metric that we found justifiable *a priori*.

of stimuli differing by completeness status and/or substitutions' functional classes may be expected to elicit low IND values yet still be characterized by consistent patterns. For example, listeners perceiving the substitution in an Incomplete stimulus as having a **D** function and the substitution in another Incomplete stimulus as having the **S** function may be expected to respond to the first stimulus with a PPoC of 1, and the second with PPoC of 2, resulting in dissimilar yet highly predictable response patterns. To account for such cases, we quantified the uncertainty on the joint occurrence of responses in the pair of stimuli *a* and *b*. Specifically, we computed for each participant *i* the (normalized) joint entropy

$$H_i\left(a,b\right) = -\frac{1}{\log 16} \sum_{(\alpha,\ \beta)\in \mathbf{a}^{(i)}\times \mathbf{b}^{(i)}} p\left(\alpha,\beta\right)\log p\left(\alpha,\beta\right), \tag{6.2}$$

where $p\left(\cdot,\ \cdot\right)$ is the joint distribution of responses for a pair of stimuli. We expected pairs of stimuli that both elicited clear, yet possibly different, PPoCs to exhibit lower values of joint entropy.

Modelling the effect of substitution classes on IND.

Our aim was to investigate whether consistency in PPoC, as quantified by IND, reflects music-theoretically motivated equivalence classes between substitutions. We therefore aimed to define models predicting the values of IND based on features that encode the music-theoretical relatedness of harmonic substitutions. For each pair of stimuli, we characterized the relationship between the substitutions they contain in five different ways as follows:

The *SemitoneDistance* variable quantifies the shortest distance in semitones, under octave equivalence, between the two substitutions (e.g., the ♭VII substitution is three ascending semitones away from the *V* substitution), while the *FifthDistance* quantifies the shortest distance on the circle of fifths between the two substitutions (e.g., the ♭VII substitution is three steps away, moving counter-clockwise, from the *V* substitution in the circle of fifths);

The *OctPair* categorical variable encodes the functional status of the dominant-seventh chords contained in the two given substitutions (e.g., the ♭VII and the *V* substitutions both belong to **D**, thus engender a **DD** pair, while ♭VII and II engender a **DS** pair);

Similarly, the *HexPair* categorical variable encodes the substitutions' membership of hexatonic classes (e.g., the ♭VII and the *V* substitutions form a *H4H1* pair);

Finally, the *SubstitutionPair* categorical variable explicitly encodes the pair of substitutions contained in the two given stimuli (e.g., the pair containing substitutions ♭VII and *V* would be encoded simply as ♭VII-*V*). On the one hand, this represents an unbiased encoding that does not incorporate any music-theoretical interpretation of the relationship between the two substitutions beyond their mere identity. On the other hand, it captures in full detail the information we have about the nature of the substitutions; any other encoding that groups substitutions, as the other variables do, would, in principle, result in a loss of information about the possible relationships substitutions may have relative to one another. However, this comes at the cost of a greater number of free parameters, one for each pairing of stimuli. As a consequence, if any of the music-theoretically motivated models captures a large proportion of the variance in the data despite having fewer parameters, we may argue the corresponding measure of relatedness to be a useful characterization of the perceived relationships between substitutions, as it would encode data parsimoniously while also affording a music-theoretically meaningful interpretation.

Each pair of stimuli is further characterized by its *CompletenessStatusPair*, a categorical variable encoding whether the two stimuli in the pair are both Complete (*CC*), both Incomplete (*II*), or with opposite completeness status (*IC*).

Bayesian mixed-effects models.

After standardizing all non-categorical variables to null mean and unit standard deviation, data were analyzed with Bayesian mixed-effects models provided with weakly informative priors (t(3,0,1); Gelman et al., 2008) and implemented in the R package *brms* (Bürkner, 2018). For each one of the main predictors

$$x \in \{\mathit{SubstitutionPair}, \mathit{SemitoneDistance}, \mathit{FifthDistance}, \mathit{OctPair}, \mathit{HexPair}\}$$

we defined the compound predictor

$$\begin{aligned}\hat{x} := x \times (1 + &\mathit{CompletenessStatusPair} + \mathit{MusicalTraining} + \\ &+ \mathit{MusicalTraining} \times \mathit{CompletenessStatusPair}),\end{aligned}$$

where *MusicalTraining* is quantified through the corresponding subscale of the Gold-MSI. Each such compound predictor models the main effect of predictor x, as well as how this effect is modulated by the Completeness of the stimuli in the pair, by the musical training of the participants, as well as the interaction between them. We first adopted a model comparison approach to determine whether these compound predictors capture incremental information relative to one another. Given the set of predictors

$$\Pi = \{\mathit{SemitoneDistance}, \mathit{FifthDistance}, \mathit{OctPair}, \mathit{HexPair}\},$$

we considered all subsets of Π as well as all sets $\{\mathit{SubstitutionPair},\ x\}$ for $x \in \Pi$. In other words, for each set $X \in \mathscr{P}(\Pi) \cup \bigcup_{x \in \Pi} \{\{SubstitutionPair,\ x\}\}$, where $\mathscr{P}(\cdot)$ is the power set (i.e., the set of all subsets) of its argument, we trained model

$$\begin{aligned}\mathrm{IND}(X) := \sum_{x \in X} \hat{x} + CompletenessStatusPair + MusicalTraining + \\ + MusicalTraining \times CompletenessStatusPair + (1 \mid Participant),\end{aligned}$$

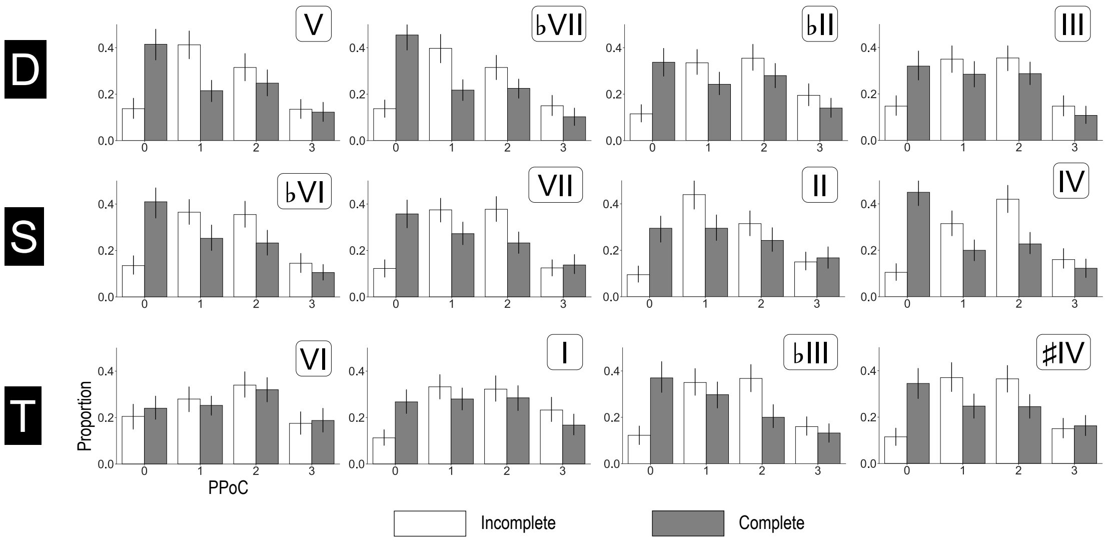


Figure 6.5 – Distributions of PPoC responses across all participants for Complete (dark bars) and Incomplete (light bars) stimuli. Each panel reports data for a given substitution, as indicated by the corresponding label. Panels are arranged so that each row comprises substitutions from the same OEC, indicated by the label on the left. In all stimuli the proportion of PPoC 0 is greater in Complete than in Incomplete stimuli, indicating that ending on a global-tonic chord increased the likelihood of null PPoC irrespective of the substitutions.

allowing for a participant-specific random intercept. Such a model predicts the values of IND for pairs of stimuli based on all compound predictors in X, while also accounting for the main effects of the Completeness of the pair of stimuli, the main effect of the participants' musical training, and the interaction between them. We then compare these models' performance under leave-one-out cross-validation with Pareto-smoothed importance sampling (PSIS-LOO; Vehtari et al., 2017). Differences in the estimated out-of-sample predictive fit (expected log pointwise predictive density, *elpd*) quantify the extent to which adding or removing a predictor results in the capture of a greater proportion of the data's variance beyond that which is simply justified by the sheer number of parameters. Results are reported in Section 6.3.2. In the following section, we report our investigation of the influence of music-theoretically motivated predictors, as reflected by their estimated coefficients when they are combined to predict IND. Data and code are available in Supplementary Material S2.

6.3 Results

6.3.1 Distributions of PPoC

Average PPoC for all Complete and Incomplete stimuli are shown in Figure 6.5 separately for each substitution. The minimal expected effect was that Complete stimuli (i.e., stimuli ending on the stable global-tonic chord) would have a significantly higher probability of being perceived as requiring no more chords than Incomplete stimuli, which end on a dominant-seventh chord. A paired t-test comparing the proportion of zeros among the PPoC responses to Complete vs. Incomplete stimuli ($t = 21.44,\ \ p < .001^*$) confirmed that listeners were indeed sensitive to the difference between Complete and Incomplete stimuli in a music-theoretically meaningful way. The difference between Complete and Incomplete stimuli with respect to the likelihood of PPoC 0 was smaller for substitutions in OEC **T** compared to substitutions from both **D** ($t = -3.76,\ \ p < .001^*$) and **S** ($t = -4.43,\ \ p < .001^*$), with no such significant difference between **S** and **D** ($t = 0.87,\ \ p = .383$), suggesting that some substitutions elicited a clearer expectation for closure than others. Additional exploratory analyses of the response distributions are available as Supplementary Material S3, including a 2-dimensional representation of the relative distances between response distributions of different substitutions, obtained through multidimensional scaling, and a statistical evaluation of within-OEC and across-OEC distances. However, note that the analysis of response distributions pooled across participants may fail to capture aspects of the participants' perception; for example, response distributions for two stimuli may be identical with no single participant reporting the same PPoC in both stimuli. We therefore analyzed the data further by quantifying similarity and uncertainty among stimuli on an individual basis, as discussed above and reported below.

6.3.2 Model comparison

Table 6.1 shows the results of the comparisons between all the models, each predicting IND based on a different set of predictors. The best performing model had $\widehat{\text{OctPair}}$ as the only compound predictor towards IND values. Furthermore, no predictive advantage was gained when combining $\widehat{\text{OctPair}}$ with any other compound predictor, as shown by the negative *elpd* values for all other models. Note that all music-theoretically motivated predictors also outperformed the neutral *SubstitutionPair* predictor, although the latter encoded common information with each of the former. Specifically, this further suggests that structuring chromatic space according to music-theoretically motivated criteria, particularly OECs, offers a more parsimonious explanation of the data than simply encoding each substitution individually.

6.3.3 Effect of music-theoretical relatedness on perceived functional equivalence

The results presented above show that model IND ($\{OctPair\}$) was the most parsimonious, and we now present the observed effect of its parameters. We report coefficient estimates (β) with their estimated error (EE) as well as evidence ratios (*Odds*) for the coefficients or some function of the coefficients being larger or smaller than zero. From a frequentist perspective, evidence ratios can be interpreted as significant(*) at a .05 confidence level when exceeding 19 (Milne & Herff, 2020). A table showing the results in full is available in Supplementary Material S4.

Effects of completeness and musical training

As shown in Figure 6.6, the *IND* values strongly depended on the *CompletenessStatusPair* of the pairs of stimuli (i.e., whether the pair comprised two complete stimuli [CC], two incomplete stimuli [II], or a complete and an incomplete stimulus [IC]). Specifically, taking IC as the reference level for factor *CompletenessStatusPair*, PPoCs for pairs of Complete ($\beta = 0.35$, $EE = 0.03$, $Odds(\beta > 0) > 9999^{*}$) or Incomplete stimuli ($\beta = 0.30$, $EE = 0.04$, $Odds(\beta > 0) > 9999^{*}$) were more similar than for pairs with opposite completeness status (IC). This effect was further strengthened by the interaction between the participants' musical training scores for pairs of Complete stimuli ($\beta = 0.11$, $EE = 0.04$, $Odds(\beta > 0) = 403.50^{*}$) and pairs of Incomplete stimuli ($\beta = 0.29$, $EE = 0.04$, $Odds(\beta > 0) > 9999^{*}$), the latter significantly exceeding the former ($Odds\ (MusicalTraining \times CompletenessStatusPairII > MusicalTraining \times CompletenessStatusPairCC) > 9999^{*}$). In other words, participants perceived a categorical difference between Complete and Incomplete stimuli, and increasing musical experience favored the salience of the similarity between stimuli with matching completeness status to a greater extent for Incomplete than Complete stimuli.

Table 6.1 – Comparison among IND(X) models predicting IND values based on different subsets X of compound predictors (see Section 6.2.5). Differences in expected log pointwise predictive density ($\Delta elpd$) relative to the best performing model IND(OctPair) are reported, together with the standard error (SE) of such estimates.

Predictors (X)	**Δelpd**	**SE**
{OctPair}	0.0	0.0
{OctPair, HexPair}	−1.9	11.4
{FifthDistance, OctPair}	−3.9	2.6
{SemitoneDistance, OctPair}	−5.6	1.3
{SemitoneDistance, OctPair, HexPair}	−7.5	11.4
{FifthDistance, SemitoneDistance, OctPair}	−9.2	2.9
{FifthDistance, OctPair, HexPair}	−9.2	2.9
{FifthDistance, SemitoneDistance, OctPair, HexPair}	−10.9	11.7
{HexPair}	−18.0	15.7
{FifthDistance}	−20.6	10.8
{SemitoneDistance}	−21.1	10.5
{FifthDistance, HexPair}	−22.6	15.9
{SemitoneDistance, HexPair}	−23.3	15.7
{FifthDistance, SemitoneDistance}	−25.8	10.8
{FifthDistance, SemitoneDistance, HexPair}	−28.3	15.9
{SubstitutionPair}	−154.0	24.4
{SubstitutionPair, FifthDistance}	−154.7	24.5
{SubstitutionPair, SemitoneDistance}	−154.9	24.5
{SubstitutionPair, OctPair}	−155.3	24.5
{SubstitutionPair, HexPair}	−156.2	24.5

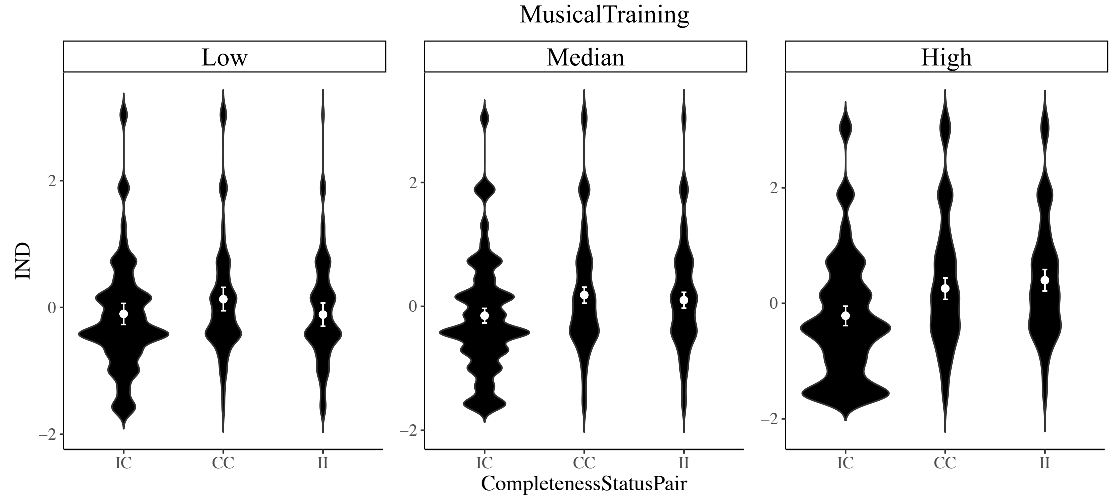


Figure 6.6 – Distribution of IND values by CompletenessStatusPair for three ranges of musical training: bottom quartile (left panel), inter-quartile range (middle panel), and top quartile (right panel). White marks (with 95% CI) report the completeness status's conditional effect on IND at the midpoints of each musical-training range. CC and II pairs have higher IND than IC pairs for increasing musical expertise.

Membership in OECs

To find out if octatonic equivalence classes (OECs) were perceived as correlates of harmonic functionality, we tested whether *IND* scores among OECs reflected the expected similarity relations illustrated in Figure 6.3b. Specifically, we wished to test whether IND scores were higher within than across OECs. We therefore let $F = (F_1,\ F_2),\ G = (G_1,\ G_2)$ and $F_1, F_2, G_1, G_2 \in \{\mathbf{T},\ \mathbf{D}, \mathbf{S}\}$ denote two pairs of OECs, and $Z \in \{\text{CC, II}\}$. We then computed the evidence ratios as follows:

$$\begin{aligned}\text{Odds}\,(F > G)_Z := & Odds(OctPairF + CompletenessStatusPairZ \times OctPairF > OctPairG+ \\ & + CompletenessStatusPairZ \times OctPairG).\end{aligned}$$

This value quantifies the evidence in favor of the hypothesis that the IND scores between two (In)Complete stimuli with substitutions in OECs F_1 and F_2 would be higher than IND scores between two (In)Complete stimuli with substitutions in OECs G_1 and G_2.

Complete stimuli with substitutions in the **D**-functioning OEC had significantly higher *IND* scores when compared with one another than with substitutions in the **T**-functioning OEC (Odds $(\mathbf{DD} > \ \mathbf{DT})_{\text{CC}} = 196.80^{*}$), as did Complete stimuli with substitutions in the **S**-functioning OEC (Odds $(\mathbf{SS} > \mathbf{ST})_{\text{CC}} > 9999^{*}$). However, Complete stimuli from **D**- and **S**-functioning OECs did not elicit higher *IND* scores when compared within the same OEC than across OECs (Odds $(\mathbf{DD} > \mathbf{DS})_{\text{CC}} = 0.30$; $Odds\,(\mathbf{SS} > \ \mathbf{DS})_{\text{CC}} = 0.74$). These results, illustrated in Figure 6.7a, suggest that **D**- and **S**-functioning OECs constitute a single equivalence class in perception with respect to the task of estimating PPoC. By contrast, PPoC judgements elicited by pairs

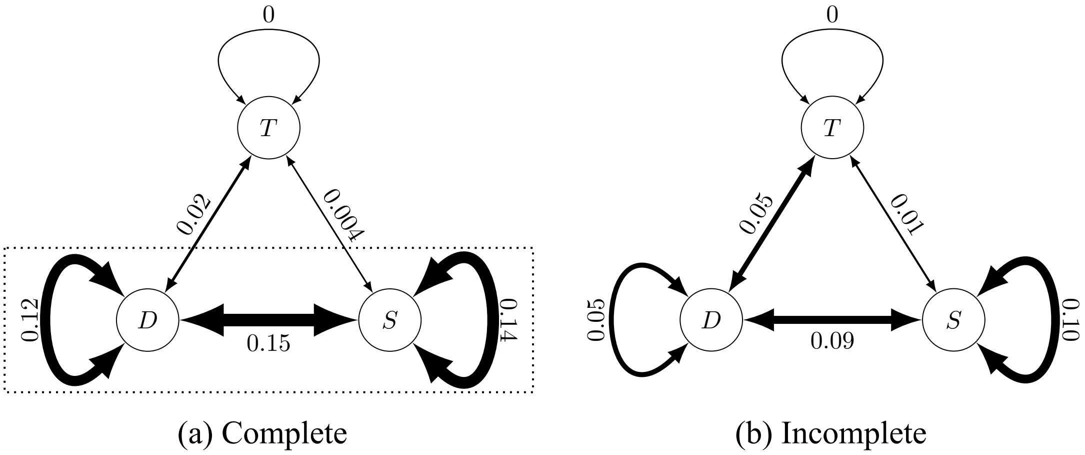


Figure 6.7 – Graphs show the observed relationships of similarity between pairs of OECs for Complete (a) and Incomplete (b) stimuli. The thickness of the edge connecting two nodes F_1, F_2 is proportional to the estimated value for the combination of coefficients OctPairF+CompletenessStatusPairZ×OctPairF, with $Z \in \{CC, II\}$. Values are expressed relative to the reference level TT of the OctPair variable. For Complete stimuli, D- and S-functioning OECs appeared to constitute a single equivalence class in perception (dotted box), while members of the T-functioning OEC were not perceived as an equivalence class according to our PPoC metric.

of complete substitutions from the **T**-functioning OEC were not significantly more similar than those elicited by mixed **DT** pairs of Complete stimuli (Odds(**TT** > **DT**)$_{CC}$ = 0.35) and **ST** stimuli (Odds(**TT** > **ST**)$_{CC}$ = 0.85). Furthermore, we found strong evidence that Complete versions of members of the **T**-functioning OEC were perceived as less similar to one another than members of the **D**- (Odds(**DD** > **TT**)$_{CC}$ = 184.57*) and **S**-functioning OECs (Odds(**SS** > **TT**)$_{CC}$ = 922.08*). Overall, unlike members of **D** and of **S**, members of the **T**-functioning OEC in their Complete form were not perceived to form mutually coherent expectations for harmonic closure, as captured by our measure of PPoC.

In Incomplete stimuli, no evidence was found for any OEC or group of OECs to constitute a separate class in perception. Specifically, for no OEC were IND scores higher among its members than they were across its members and members of a different OEC (for {**T**, **D**, **S**} $\ni X_1 = X_2 = Y_1 \neq Y_2$, at least one Odds$(X > Y)_{II} < 2$). Figure 6.7b shows the resulting patterns, which do not support the perceptual relevance of OECs as functional equivalence classes in Incomplete stimuli.

6.3.4 Post-hoc analysis of joint entropy

The results reported above show that, for Complete stimuli, members of **T** elicit more dissimilar PPoCs than members of **D** and **S**. This observation may be due to members of **T** failing to elicit clear PPoCs, thus resulting in dissimilar responses across stimuli. However, it is also possible that members of **T** nevertheless elicited clear and unambiguous PPoC that happened

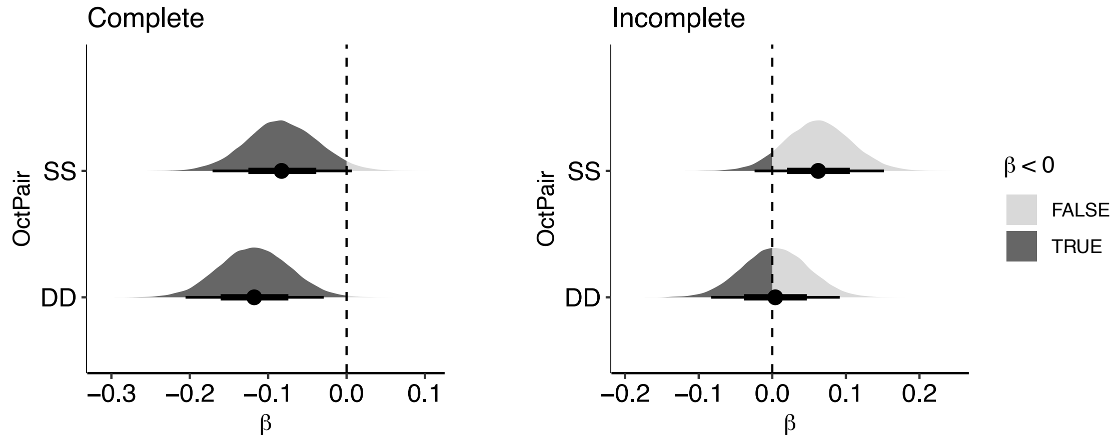


Figure 6.8 – Posterior distribution of the incremental effect on the joint entropy, relative to the reference category TT, for OctPair categories DD and SS in CompletenessStatusPair categories CC (Complete, left) and II (Incomplete, right). Responses to Complete stimuli in T exhibit significantly greater uncertainty (i.e., β<0) than responses to Complete stimuli in both D and S, while no such difference is observed with Incomplete stimuli.

to be different across different members of **T**. To disambiguate these two explanations, we analyzed the joint entropy of the responses to pairs of stimuli by replacing IND_i with H_i as the dependent variable in model IND({OctPair}). As shown in Figure 6.8, convincing evidence was found that joint entropy between Complete members of **T** was higher than between Complete members of **D** (Odds (TT > DD)$_{\text{CC}}$ = 226.85*) or **S** (Odds (TT > SS)$_{\text{CC}}$ = 27.28*). This suggests that Complete stimuli in **T** elicited less consistent expectations of PPoC compared to members of other OECs. No such evidence was found for Incomplete stimuli, where **D** and **S** were not significantly different from **T** (Odds (DD > TT)$_{\text{II}}$ = 1.17, Odds (SS > TT)$_{\text{II}}$ = 11.57).

6.4 Discussion

In this study, we set out to test the extent to which perceived functional equivalence, as quantified by similarity in the perceived proximity of harmonic closure (PPoC), reflects music-theoretical accounts of harmonic function, as drawn from theories of extended tonality. Our results show that music-theoretical accounts attributing functional meaning to OECs in the extended-tonal harmonic idiom are not only appropriate characterizations of both historical and current compositional practices but also provide a parsimonious model of perceived functional harmonic relations for a sample of listeners who are familiar with different instantiations of compositional practices in the idiom of extended-tonal practice. We found that OECs differ in terms of the clarity of the expectation for closure their members elicit, as well as in terms of their coherence as equivalence classes. In particular, our results indicate a comparatively lower relatedness among members of **T** than among members of other OECs. In the following, we interpret these findings in light of their music-theoretical framing and highlight limitations and prospects for further research.

While our results are consistent with the hypothesis that OECs as functional categories may constitute a cognitively relevant representation of pitch-space structure, the present results should not be read as conclusive evidence that listeners' perception is guided in some sense by representations of such structuring principles. Other classifications of harmonic sonorities may also characterize perception similarly or even more accurately, and more complex models with maximal random-effects structure (Barr et al., 2013) may further identify sources of inter-participant variability potentially underlying the reported effects. Overall, while we cannot fully conclude that listeners' cognitive representation of pitch space employs OECs or an isomorphic representation, the present results offer empirical evidence that OECs may be and indeed have been adopted as a way of formalizing, modelling, and expressing listeners' functional hearing. Such parsimonious descriptive adequacy of OECs in perception may have represented a stable equilibrium towards which compositional practices and music-theory have converged, contributing to the feedback-loop between the introspections of composers, musicians, and theorists (folk psychology, as described by Cross, 1998), on the one hand, and idiomatic musical practices and theoretical formalizations on the other.

According to our analysis, Complete and Incomplete stimuli elicited different behavioral responses. Specifically, PPoC between a Complete and an Incomplete stimulus was less similar than PPoC between two Complete or two Incomplete stimuli. Furthermore, increased musical training seemed to favor the likelihood for listeners to report the same PPoC for two Incomplete stimuli to a greater degree than for two Complete stimuli. Recall that the core chord progressions underlying Complete and Incomplete stimuli were different, so that their lengths could be matched. However, it is unclear why musical training would have different effects on the two types of progressions. A possibly more salient common feature of Complete stimuli is that they all ended with a global tonic, whereas each Incomplete stimulus ended with a different chord. Such surface similarities alone may explain the finding that pairs of Complete stimuli elicited similar PPoCs irrespective of listeners' expertise, while experienced listeners would, in their responses, reflect the less salient structural similarity of Incomplete stimuli to a greater extent.

Complete and Incomplete stimuli also differed in terms of how substitutions produced harmonic functionality. For Incomplete stimuli, substitutions in both **D** and **T** elicited less similar PPoC than those in **S**, with no single equivalence class consistent with the criteria set out in Figure 6.3 emerging from the data. For Complete stimuli, in turn, a clear picture emerged that is partially consistent with music theoretical accounts, whereby substitutions in **D** and **S**, on the one hand, and **T**, on the other, can be distinguished from one another. The observation that the extended-tonal music-theoretical perspective is more closely matched by perceptual data in response to Complete rather than Incomplete stimuli could reflect the possibility that harmonic functionality may still not be fully established at the time an expectancy-inducing chord or its corresponding scale degree is presented, and yet be determined *a posteriori* once a viable resolution of the expectancy is achieved, for example, by transitioning into a globally stable sonority such as the tonal center. Listeners may not necessarily attribute global dominant function to a certain chord at the time of its occurrence but may be willing to accept resolution towards the global tonic as marking a satisfactory harmonic closure. Only then may the chord be understood (retrospectively) as expressing functional meaning in some generalized sense which allows for retrospective reinterpretation (Cecchetti et al., 2022). In this respect, our data shed light on how establishing a global tonal context determines which of the 12 chromatic degrees of the scale, relative to the tonal center, share such (potentially revisable) functional behavior.

In light of the low average degree of musical sophistication of our participants, and of the scarce evidence for an effect of musical training beyond Complete/Incomplete discriminability, it should be expected that the correspondence between perception and the structural principles guiding composition are a result of musical acquisition processes and implicit learning from exposure to particular musical repertoires (Pearce, Ruiz, et al., 2010; Reber, 1989; Rebuschat, 2022; Rohrmeier and Rebuschat, 2012). This leaves open the question as to which aspects of compositional practice are actually acquired by listeners. In particular, substitutability is a core feature of functional harmony, and listeners may learn through exposure to identify chords belonging to the same OEC as substitutes by observing how frequently they occur in analogous contexts (Jacoby et al., 2015; Rohrmeier and Cross, 2008; White and Quinn, 2018). If this were the case, we would expect high similarity of PPoC within OECs, and low similarity across OECs, as illustrated in Figure 6.3. However, our results do not show that members of the same functional class behave as mutual substitutes by systematically eliciting similar response patterns from our participants, nor that the three OECs form three distinct families of substitutes. On the contrary, functional classes can be distinguished based on the degree of similarity among their members, at least insofar as **T** behaves differently from **S** and **D** in our Complete stimuli. Specifically, representatives of the tonic function appear to elicit maximally dissimilar (as quantified by *IND*) and uncertain responses (as quantified by *H*) when compared with one another, relative to **D** and **S**.

The commonality of **D** and **S** in Complete stimuli, which appear to form a single class, as well as the comparatively coherent perception of pairs of Incomplete stimuli belonging to **S** as opposed to **D** and **T**, may result from the prevalent exposure to blues, rock and pop repertoires among our pool of participants. In such repertoires, the extensive use of plagal closure (Everett, 2004; Temperley, 2011; cf. Harrison, 1994 for a more general dualist view of chromatic harmony) endows the subdominant with a cadential role that is exclusively attributed, rather, to the dominant in common-practice harmony (Caplin, 1998). More generally, in most Western tonal-compositional practices, **D** and **S** share the purpose of eliciting expectations towards a particular goal, whereas members of the tonic function are less likely to be used with the purpose of eliciting clear expectations towards any harmonic goal; relative to the global tonal context, and unlike **S** and **D**, **T** also subsumes the tonic function, which is defined in negative terms as the absence of a goal-directed drive (Doll, 2017; K. M. Smith, 2020). This finding is also supported by computational clustering analyses of chord-transition probabilities in classical repertoires (Rohrmeier and Cross, 2008). Finally, it is possible that, in tonal music, the tonic function as a marker of global closure may only be instantiated by a single harmony, the global tonic I as determined by the harmonic context, rather than by a set of substitutable chords.

Overall, we can conclude that representatives of the tonic function elicit incoherent patterns of expectations when employed as preparations for some harmonic goal within a tonal context. This observation may be interpreted as indicating the difficulty of parsing, or interpreting, chord progressions in which members of **T** are employed as non-goal elements of implication-realization pairs in the proximity of global harmonic closure. Furthermore, our results are consistent with a categorical perceptual distinction between tonic-functioning harmonies (gathered in **T**) and expectation-inducing harmonies (**S** and **D**). The latter may be thought of as equivalent and possibly substitutable in their quality of being perceived as preparations (Rohrmeier, 2020b; Rohrmeier and Moss, 2021) for future harmonic events and harmonic closure.

The perceptual biases we identified may be thought of as part of a competence (Chomsky, 1965) for harmonic syntax in music, forming the basis of the capacity for processing and interpreting idiomatic extended-tonal music (Cecchetti et al., 2020; Lerdahl and Jackendoff, 1983a; Steedman, 1996). Our results may thus inform theoretical and computational models attempting to formalize such implicit knowledge (e.g., Finkensiep and Rohrmeier, 2021; Steedman, 1984), as well as the human capacity to learn (Harasim, 2020) and process musical harmonic structure (Granroth-Wilding and Steedman, 2014; Harasim et al., 2018; Jackendoff, 1991). In particular, experimental evidence supports the view that listeners construe mental representations of hierarchical musical structure (Cecchetti et al., 2021; Herff, Harasim, et al., 2021; Koelsch et al., 2013; Leino et al., 2007; Serafine et al., 1989). In modelling such representations, the present results offer empirical support for the choice of syntactic dependencies that reflect observed harmonic relationships, as suggested, for example, in syntactic accounts of extended tonality (Rohrmeier and Moss, 2021). Nevertheless, further investigation into the role of stylistic familiarity and musical expertise is necessary, as the present results are only representative of some so-called average listener with generically Western musical enculturation, while the relationship between music-theoretical formalization and perception is likely to be strongly dependent on musical idiom and individuals' exposure and training.

In this study, stimuli were designed to be particularly evocative of jazz voicings, with the purpose of providing listeners with a deliberately chosen, ecologically valid stylistic context in which extended-tonal harmony and functional substitutions are idiomatic (Levine, 1995; Rohrmeier, 2020b). However, while there are global principles of extended tonality that persist over its entire historical span (Haas, 2004; Rohrmeier and Moss, 2021; Tymoczko, 2011), jazz harmony is a specific instantiation of certain stylistic preferences within the possible range of musical relations. For example, tritone substitution is particularly prominent in jazz harmony (Biamonte, 2008; Levine, 1995), while backdoor substitutions and plagal closure are typical in pop and rock (de Clercq and Temperley, 2011; Doll, 2017; Everett, 2004; A. Moore, 1995; Temperley, 2011). As a consequence, it is likely that prevalent individual familiarity with pop and rock music and its stylistic preferences may have influenced perceived harmonic relatedness as quantified in this study, as suggested by previous evidence for the stylistic priming of harmonic expectancy (Vuvan and Hughes, 2019). Future research may also investigate how patterns of harmonic relatedness are influenced by metricality, which was not manipulated in our experimental design. Specifically, metrical weight may interact with harmonic expectancy, hence with perceived harmonic functionality.

Finally, hexatonic relatedness did not capture any additional variance in our data compared to octatonic relatedness. Considering that our task was based on goal-directed expectancy, a characteristic aspect of functional harmony, this observation suggests that HECs were not perceived by listeners as carrying this type of functional meaning, consistently with music-theoretical literature. It should be noted that previous empirical approaches to tonal relatedness based on probe-tone profiles also failed to find evidence for the perceptual reality of hexatonic relatedness (Krumhansl, 1990). Nevertheless, these results do not exclude the possibility that hexatonic relatedness may constitute a cognitively relevant representation with non-functional meaning, for example by expressing manipulations and contrasts of harmonic color. While tasks leveraging goal-directed expectancy as a proxy of harmonic relatedness have already offered an accessible gateway into the perception of functional tonal harmony, it will be a challenge for future research to identify appropriate experimental paradigms to investigate notions of harmonic relatedness in non-functional harmony, including transformational and non-functional aspects of extended-tonal musical practices (see, for example, Guichaoua et al., 2021).

6.5 Conclusion

This study highlights similarities and differences between music-theoretical accounts of functional harmony in extended tonality on the one hand, and perceptual manifestations of harmonic functionality in the perceived proximity of harmonic closure (PPoC) on the other. We found evidence that octatonic equivalence classes (OEC), as defined music-theoretically, parsimoniously predict similarity in a behavioral response such as PPoC. However, while such theoretical accounts hypothesize three distinct OECs, characterized by similar PPoCs within classes and dissimilar PPoCs across classes, this is not directly reflected in our results. In fact, we rather observed members of class **D** to behave similarly to each other and to members of class **S,** and vice versa. This could possibly be the result of listeners having been primed by such similarities in pop and rock music. By contrast, **T** elicited distinctly different behavioral responses compared to **S** and **D**, and exhibited lower coherence as a class in the sense that members of class T did not elicit more similar or mutually predictive responses with other members of class T than with members of S or D. Specifically, we also found evidence that tonic function may be defined in negative terms as gathering harmonies that fail to elicit consistent expectations for closure, possibly because they are not employed in this way in these repertoires. As a consequence, we interpret these findings as reflecting a music-theoretically meaningful distinction between, on one hand, substitutions that are meant to induce expectancy (**S**, **D**) and, on the other hand, those that are not meant to do so (**T**). This may represent a cognitive correlate of the implicit knowledge of abstract structuring principles underlying the capacity of Western-enculturated listeners to perceive extended-tonal music as structured (Jackendoff, 1991; Rohrmeier, 2020b). Overall, this study complements music-theoretical accounts and contributes to an understanding of shared perceptual structural templates in Western music beyond common-practice tonality.

7 Incremental parsing: a study on rhythm

Abstract

Music can be interpreted by attributing syntactic relationships to sequential musical events and, computationally, such musical interpretation represents an analogous combinatorial task to syntactic processing in language. While this perspective has been primarily addressed in the domain of harmony, we focus here on rhythm in the Western tonal idiom and we propose for the first time a framework for modelling the moment-by-moment execution of processing operations involved in the interpretation of music. Our approach is based on (1) a music-theoretically motivated grammar formalising the competence of rhythmic interpretation in terms of three basic types of dependency (preparation, syncopation, and split; Rohrmeier, 2020a), and (2) psychologically plausible predictions about the complexity of structural integration and memory storage operations, necessary for parsing hierarchical dependencies, derived from the Dependency Locality Theory (Gibson, 2000). With a behavioural experiment, we exemplify an empirical implementation of the proposed theoretical framework. One-hundred listeners were asked to reproduce the location of a visual flash presented while listening to three rhythmic excerpts, each exemplifying a different interpretation under the formal grammar. The hypothesised execution of syntactic-processing operations was found to be a significant predictor of the observed displacement between the reported and the objective location of the flashes. Overall, this study presents a theoretical approach and a first empirical proof-of-concept for modelling the cognitive process resulting in such interpretation as a form of syntactic parsing with algorithmic similarities to its linguistic counterpart. Results from the present small-scale experiment should not be read as a final test of the theory, but they are consistent with the theoretical predictions after controlling for several possible confounding factors and may form the basis for further large-scale and ecological testing.

This Chapter is published as: Cecchetti G., Tomasini C. A., Herff S. A., & Rohrmeier M. A. (2023). Interpreting rhythm as parsing: Syntactic-processing operations predict the migration of visual flashes as perceived during listening to musical rhythms. *Cognitive Science*, **47**(12). GC conceptualised the study with SAH, analysed the data, and wrote the original manuscript. CAT implemented the paradigm and collected data under supervision by GC and SAH. All authors revised the manuscript.

7.1 Introduction

The idea that musical structure is interpretable in terms of hierarchically nested patterns of syntactic relationships among events is widespread in music theory (Baroni et al., 1983; Bernstein, 1976; Keiler, 1978; Lerdahl and Jackendoff, 1983a; Rohrmeier, 2011, 2020b; Schenker, 1935). For example, hearing an event *as* a preparation of another event, or *as* being temporally displaced relative to some reference location, putatively corresponds to different attributes of the events as perceived (cf. Dubiel, 2017). Formally, this notion of interpretation can be modelled through rule-based analysis-by-synthesis, sharing formal analogies with linguistic syntax (Chomsky, 1957) and, more generally, generative Bayesian modelling (Tenenbaum et al., 2006; Ullman and Tenenbaum, 2020). Such models of abstract syntactic knowledge, or "competence" (Chomsky, 1965), can be turned into models of processing as incremental syntactic parsing (Jackendoff, 2002b; Steedman, 2000), which have been extensively investigated empirically in the linguistic domain (see, e.g., Frazier, 1978; Pickering and van Gompel, 2006). From a computational perspective, the processing of syntactically organised music has been proposed to represent an analogous combinatorial problem to linguistic sentence processing (Asano, 2021; Asano and Boeckx, 2015; Jackendoff, 2009; Katz and Pesetsky, 2011; Tillmann, 2012). From this perspective, idiom-specific music-theoretical frameworks can be interpreted as explicit hypotheses about the nature of a syntactic competence (Cecchetti et al., 2020), which both expert and untrained listeners may acquire through implicit learning (Rohrmeier, 2010; Tillmann, 2005) and grammar induction from the exposure to repertoires (Harasim, 2020; Rohrmeier and Cross, 2009, 2013). Research on this topic has predominantly focused on the domain of harmony, showing that listeners are sensitive to hierarchical structures and to their complexity (Herff, Harasim, et al., 2021; Koelsch et al., 2013; Ma et al., 2018b), and that harmonic syntactic structures are processed during listening resulting in the detection of violations (Maess et al., 2001; Patel et al., 1998; Slevc et al., 2009) as well as retrospective revision (Cecchetti et al., 2022). At least part of the experience of music may then originate, similarly to language, as the cognitive implementation of an incremental parser (Cecchetti et al., 2022; Jackendoff, 1991). However, it remains to be demonstrated empirically whether specific cognitive operations occurring during music listening can be modelled at the algorithmic level of description (Marr, 1982) as the operations of a parser.

In this study, we present a framework for testing hypotheses about the moment-by-moment execution of cognitive operations involved in parsing music into an interpretation. We also extend research on the syntactic interpretability of music beyond the dimension of pitch, focusing on musical rhythm in the idiom of Western tonality. Note that different musical features – such as pitch, harmony, and rhythm – are organised according to different structuring principles that interact in complex ways (Harasim et al., 2019; Prince, 2011; Prince et al., 2009; Yust, 2018). However, while pitch-related musical dimensions can hardly be abstracted from the time dimension in idiomatic music, rhythms can be presented – and investigated – in isolation from the pitch dimension through the use of non-pitched or single-pitched instruments. Rhythm is then a perfect candidate to investigate whether and how an interpretation of a given musical dimension is formed incrementally during listening. In particular, we test here whether the processing leading to such an interpretation is compatible with a parsing model entailing operations of structural integration and memory storage, as it is also predicted in linguistic sentence comprehension (Gibson, 1998, 2000).

7.1.1 Musical rhythm and its interpretation

Rhythm is minimally characterised in terms of the onsets and durations of musical events: this information identifies the temporal location of events as encoded in a musical score, and allows performers to reproduce the notated temporal patterns. However, the experience of musical rhythm is not fully identified by such properties of events in isolation (Honing, 2008; Levitin et al., 2018). For example, temporal regularities in the patterns of sounded events as a whole result in the perception of relative relationships of metrical strength across different time positions. The resulting "metrical" grid of alternating strong and weak metrical positions, conceived as hierarchical layers of (typically regular) pulses (Large and Jones, 1999; Lerdahl and Jackendoff, 1983a), manifests itself in neural and behavioural entrainment as well as peaks of heightened expectancy and attention (Fitzroy and Sanders, 2015; Large and Snyder, 2009; Mathias et al., 2020; Nozaradan et al., 2012), even in the absence of actual sounded events (e.g., during imagination; Herff et al., 2020). Furthermore, based on Gestalt principles, events are grouped with one another (Deliege, 1987; Deutsch, 1999; W. J. Dowling, 1973) resulting in a hierarchy of nested groupings (Lerdahl and Jackendoff, 1983a; Zhang et al., 2016). Both meter and grouping are well-studied in music psychology and contribute to characterising how listening to rhythm feels like for listeners, as demonstrated by their predictive value towards behavioural and neural responses.

Rhythmic interpretation, as conceived in music theory, is related to the interaction of metrical and grouping structure (Honing, 2008; Lerdahl and Jackendoff, 1983a; Rohrmeier, 2020a; Rothstein, 1989). Specifically, depending on their alignment with the underlying metrical grid, events and groups of events are interpreted as being in specific types of functional relationships with one another. For example, the passage exemplified in Figure 7.1 (bottom) may be interpreted as an elaborated version of a simple rhythmic template (top). In particular, events 2 and 3 may be heard as displaced instances of their metrically stronger counterparts in the template, engendering a *syncopation*. Events 4 to 6, that group forwards with the metrically stronger event 7, are instead introduced as an *upbeat*, i.e. a preparation leading towards the latter event, the downbeat. In this theoretical framework, interpreting rhythm refers to the process of making sense of the functional relationship of each rhythmic event relative to the other events and to the underlying metrical grid – for example, the relationship between a downbeat and the upbeat that leads towards it.

Music-theoretical discourse offers a rich characterisation of rhythmic interpretation (Caplin, 2002; Mirka, 2009; Morgan, 1978; Rothstein, 1989; Schenker, 1935; Yust, 2018), and empirical research has investigated aspects of the associated phenomenology – in particular, the musical

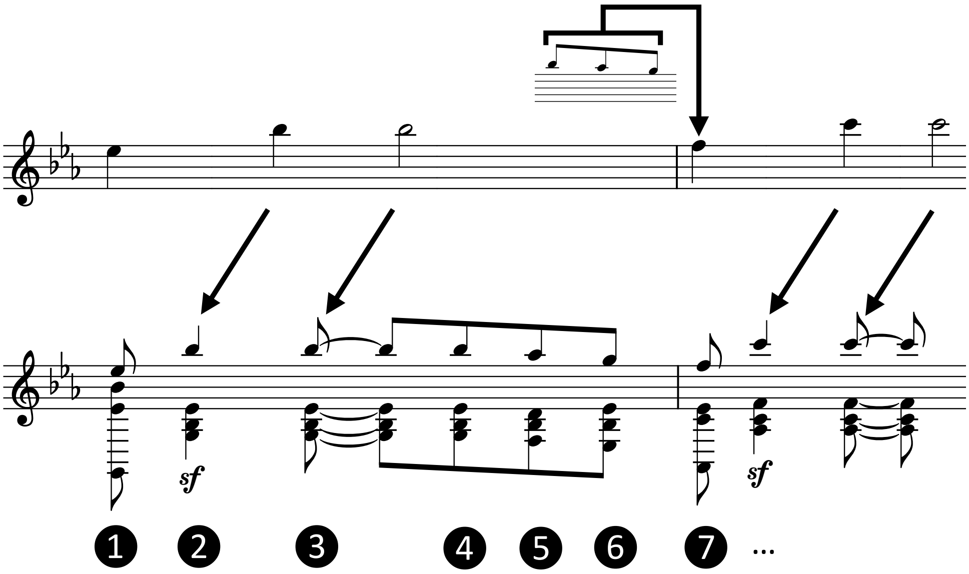


Figure 7.1 – An excerpt from L. van Beethoven's op. 18 n. 4, i, mm. 60f. The rhythm of this passage (bottom) may be understood as an elaborated version of a simple template (top). Specifically, events 2 and 3 may be interpreted as displaced versions of their counterparts in the template, whereas events 4 to 6 may be understood as an insertion that "leads towards" event 7.

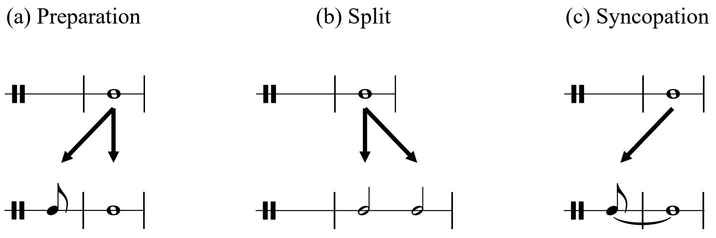
and psychological factors that influence whether and how strongly listeners perceive a given rhythm as being locally misaligned with meter by interpreting certain events as upbeats (London et al., 2009) or syncopations (Ladinig et al., 2009; Witek, Clarke, Kringelbach, et al., 2014). While these findings shed light on the relationship between rhythmic events and meter, it remains unclear whether and how the relationships linking events with one another within a given rhythm, that are also implied by an interpretation, contribute to the listeners' experience. Here, we build on the understanding of rhythmic interpretation as mapping the musical surface into a coherent network of functional relationships, and we present a first step in the direction of supporting such a cognitive model with empirical evidence. In the following, we first introduce a formalisation of the syntactic competence (Chomsky, 1965) that underlies rhythmic interpretation from a music-theoretical perspective, whereby a rhythm's interpretation is modelled as its derivation under a generative grammar (Section 7.1.2; Rohrmeier, 2020a). Based on this account at the competence level, we then sketch a theory of processing (Section 7.1.4) inspired by existing psycholinguistic models (Section 7.1.3; Gibson, 2000) and we introduce flash reproduction as a behavioural task to test the predictions of such a model (Section 7.1.5).

7.1.2 Formalising the competence of rhythmic interpretation: split, preparation, and syncopation

Generative grammars provide a natural framework for modelling interpretations arising through the recursive composition of elementary syntactic relationships (Chomsky, 1957). In the musical domain, generative models have been proposed to account for structural interpretations over different parameters such as harmony (Harasim et al., 2018; Rohrmeier, 2011, 2020b; Steedman, 1984), melody (Baroni, 1999; Boltz and Jones, 1986; Finkensiep and Rohrmeier, 2021; Finkensiep et al., 2019), grouping (Lerdahl and Jackendoff, 1983a), metrical structure (Longuet-Higgins, 1978), as well as rhythmic structure in the idiom of Western tonality (Foscarin et al., 2019; Rohrmeier, 2020a; Sioros et al., 2018). In this context, three fundamental types of functional relationships have been proposed to model the interpretation of sequences of event durations in the context of Western tonality by Rohrmeier (2020a) and Sioros et al. (2018): preparation, split, and syncopation. The two models share many similarities, and while the present approach is rooted in the former formalism, many arguments would also apply to analyses based on the latter. In particular, rhythmic interpretation as modelled under Sioros et al. (2018) has been shown to correlate with listeners' similarity judgements (Bruford et al., 2020), supporting the perceptual relevance of the three fundamental rhythmic relations.

In the Abstract Context-Free Grammar (Harasim et al., 2018) proposed by Rohrmeier (2020a), these are implemented as three different families of generative rules operating on non-terminals that correspond to intervals of time (Figure 7.2). We loosely overview here the relevant aspects of the formalism, together with their musical motivation, and we refer the reader to the original paper for full details. Some formal properties of the grammar's rules, such as their headedness and arity, will then be leveraged to generate predictions about real-time processing.

Preparation. A metrically weak event may be grouped together with a following metrically stronger event, with the former being understood as an upbeat, or *preparation*, of the latter (Figure 7.2a). This is qualitatively associated with a sense of directionality, as upbeats are always "upbeats to..." (e.g., Morgan, 1978, p. 446): according to the Grove's definition, upbeats entail a "forward rhythmic impulse [...] towards the accent" (Doğantan-Dack, 2001). While the grouping and the metrical weights favour the interpretation as a preparation, the qualitative characterisation hints at an additional aspect of how such a rhythm may be experienced – an aspect that is not reducible to simply identifying the grouping and the metrical weights. Specifically, since the functional role of the event that is perceived as a preparation is to induce an "impulse" towards the accented downbeat, the former is understood as being subordinate to the latter on purely rhythmic grounds. Formally, the application of a preparation rule maps the parent time interval into two "children" events: a metrically weak one – the upbeat – on the left, and a metrically stronger part – the downbeat – on the right. Extending the original formulation, we introduce here a notion of headedness whereby preparation rules are



(a) Preparation (b) Split (c) Syncopation

	(a) Preparation	(b) Split	(c) Syncopation
Interpretation	goal-directed expectancy	“rebound” of a metrically strong event	temporal displacement
Arity	binary	(typically) binary	unary
Headedness	right	left	/

Figure 7.2 – Three kinds of interpretation of rhythmic events are reflected in the properties (arity and headedness) of the generative rules in the grammar proposed by Rohrmeier (2020a). The application of a preparation rule (a) introduces a metrically weak event leading towards a following, stronger event. The application of a split rule (b) introduces a metrically weak event as a “rebound” of a metrically stronger one. Syncopation (c) displaces the onset of an event.

right-headed, reflecting the distribution of metrical weight between the children nodes.

Split. A group of events may be understood as subdividing, or *splitting*, a metrical timespan (Figure 7.2b). In such cases, the first event of the group coincides with the onset of the timespan and is metrically stronger than the other events. As a consequence, the latter are understood as subordinate “afterbeats” (Lerdahl and Jackendoff, 1983a) of the former, and may be characterised metaphorically as “rebounds” or “echoes” (Hauptmann, 1853, p.191) of the metrically stronger event. The resulting grouping would coincide with a “metrical grouping” as *per* Longuet-Higgins and Lee (1982), as opposed to “phrasal groupings” (for example, the grouping linking an upbeat with its downbeat) which do not necessarily start with their metrically strongest event. Formally, the application of a split rule divides the time interval corresponding to the parent non-terminal into (typically) two children parts, whereof the first one – the left child – is metrically the strongest and is understood as the head of the resulting constituent.

Syncopation. Finally, events may be interpreted as being displaced from a strong metrical position to a weaker one, originating the phenomenon of *syncopation* (Figure 7.2c). Syncopations typically manifest themselves as accented events in relatively weak metrical positions followed by (or preceded by) a metrically stronger metrical position – the *lacuna* – carrying a non-accented event or no event at all (Huron and Ommen, 2006). This characterisation suffices to describe how a syncopation looks like on the surface, and this type of rhythmic structure is associated with pleasurable groove (Sioros et al., 2014; Witek, Clarke, Wallentin, et al., 2014) as well as with behavioural and neural correlates of perceived incongruency between rhythm and meter (Vuust et al., 2018; Vuust and Witek, 2014). However, in the Western tonal idiom, interpreting an event as being syncopated may entail acknowledging an additional *latent* attribute: namely, that the syncopated event *is* a (displaced) instance of the missing accented event in the *lacuna* (I. Tan et al., 2019; Temperley, 1999; cf. also Schenker, 1935); as such, it inherits the relationships that an accented event in the *lacuna* would have formed towards other events. Formally, syncopations are modelled here as unary rules that displace the onset of an event to a metrically weaker position, either as an anticipation or a delay.

It should be noted that the aforementioned relationships do not only link individual events with one another, but also groups of events with one another. For example, entire groups of events can collectively be understood as preparations of a subsequent event, as in the case of upbeat phrases (Beach, 1995; R. McClelland, 2006). This indicates that interpretive relationships compose hierarchically, accounting for the functional role of individual notes as well as, through their composition, for the global interpretation of an entire segment of music.

Such a formalism captures the computational tools that allow to generate idiomatic rhythms through rule applications and, vice versa, to infer plausible derivations of a given rhythm (for a detailed account of probabilistic learning and parsing of Abstract Context Free Grammars, cf. Harasim, 2020; Harasim et al., 2018). For example, a possible derivation of the rhythm proposed in Figure 7.1, highlighting the functional role of all events as they are generated through splits, preparations, or syncopations, is illustrated in Figure 7.3. Note that a derivation of the rhythmic surface represents an encoding of a possible interpretation of the surface itself. The mapping of a rhythmic surface into a derivation is then a characterisation at the computational level Marr, 1982 of the cognitive process of *interpreting* rhythm. Here, we propose, and offer a preliminary test for, a model of how such competence could implicitly underlie online processing as listeners develop an interpretation of rhythmic stimuli during listening. To this end, we introduce a framework to quantify the processing costs associated with parsing the structures implied by the music-theoretical formalism, and investigate whether such costs have observable manifestations in a behavioural task.

7.1.3 Cognitive operations involved in online sentence processing

Cognitively plausible models of sentence processing suggest that different sources of cognitive load (e.g., the implementation of different processing operations such as Fork and Join in a left-corner parser; Resnik, 1992; Shain et al., 2016; van Schijndel et al., 2013), contribute to processing complexity. In particular, predictions stemming from the Dependency Locality Theory (DLT; Gibson, 2000) are in good agreement with reading times as a behavioural manifestation of processing complexity (Gibson, 1998, 2000; Shain et al., 2016). Here, we take the DLT as a psychologically plausible characterisation of the cognitive operations involved in the syntactic processing of sequential inputs in the linguistic domain.

Let the sequence $\{x_{t_n}\}_{n<N}$ of N terminal symbols occurring at times $t_0 < t_n < t_{N-1}$ be a

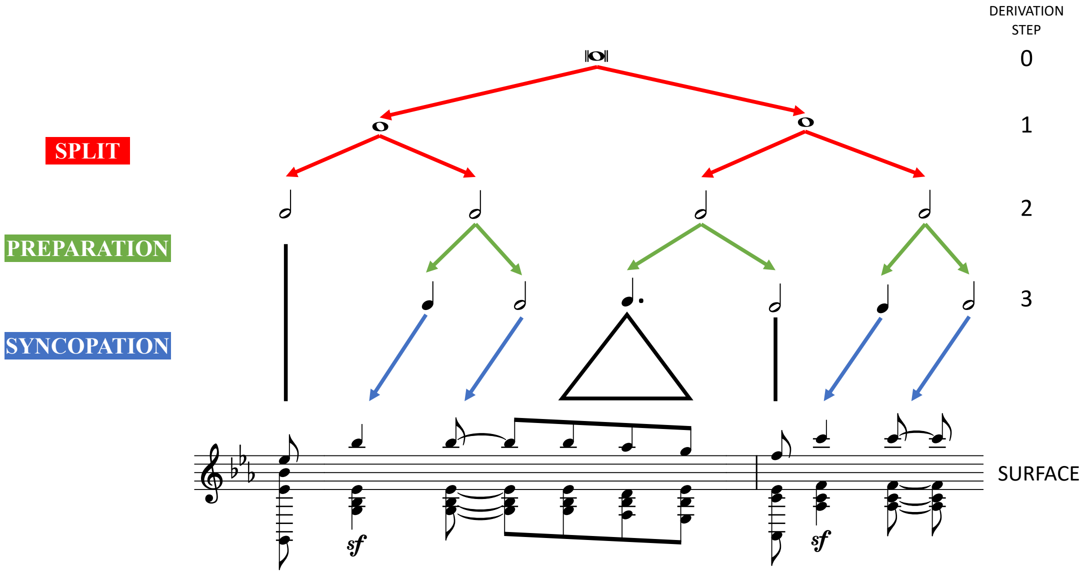


Figure 7.3 – Schematic visualization of a plausible rhythmic derivation of the excerpt introduced in Figure 7.1. The events sounding in the musical surface are generated by the application of splits, preparations and syncopations. The displacements due to syncopations, and the insertion of preparations, shape the duration of the events as they appear in the surface: for example, the first event, which is represented by a half note in derivation step 2, is reduced to an eight-note duration by the insertion of a syncopated quarter-note preparation in derivation step 3, which is further anticipated by an eight-note to become the second event in the surface. Full details of the formalism can be found in Rohrmeier (2020a), and a complete derivation in the original notation is available as Supplementary Material S1. Note that alternative interpretations for the same musical surface, corresponding to different derivations, are also possible.

linguistic input. As the input is processed incrementally (Marslen-Wilson and Tyler, 1980), at every time t_n the parser has already processed terminals $x_{t_0}, \dots, x_{t_{n-1}}$ into an incomplete derivation and moves on to read the current terminal x_{t_n}. The DLT hypothesises three sources of processing cost, corresponding to different computations to be performed during parsing: discourse processing, structural integration and memory storage. These are described below, followed by a discussion of how such costs are predicted to occur in the context of musical rhythm.

Discourse processing. Discourse processing refers to the cognitive cost of introducing a new referent in the discourse, constructing the corresponding discourse structure, and is incurred every time nouns that indicate discourse objects or tensed verbs that indicate discourse events are encountered (Gibson, 1998; Webber, 1988).

Structural integration. Structural integration refers to the processing operation of attaching a constituent head x_{t_n} to a constituent whose head $x_{t_{n-k}}$ (with $k > 0$), the "target" of the attachment, belongs in the past (Figure 7.4a). As such attachment requires that, at time t_n,

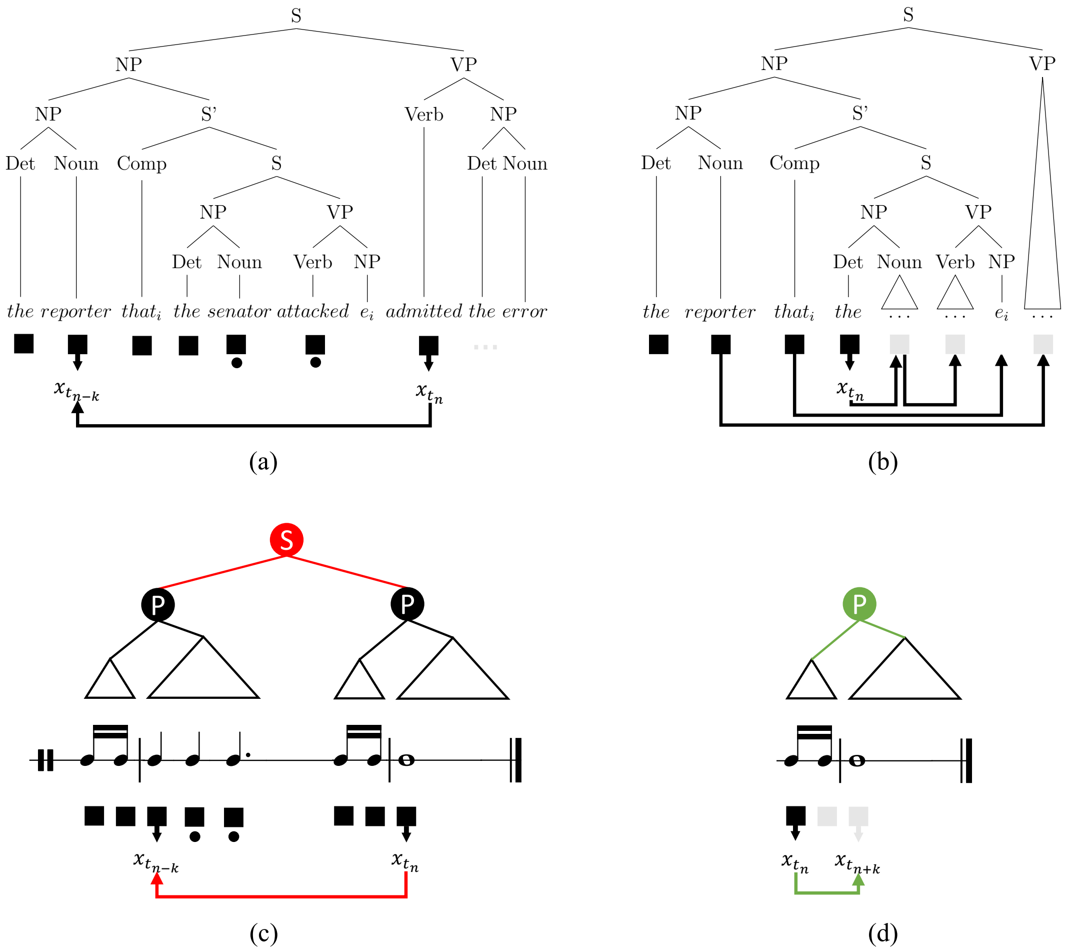


Figure 7.4 – Examples of structural integration and memory storage as hypothesised in language (a,b; from Gibson, 1998) and musical rhythm (c,d). In (a), a non-local attachment occurs at time t_n as the verb phrase (VP) headed by admitted is integrated with the noun phrase (NP) headed by reporter. Such structural-integration operation results in cognitive costs quantified by the number of discourse referents (marked by black dots) intervening in the attachment region (arrow). In (b), as the second the is encountered at time t_n, several constituent heads are expected to occur in the future in order to complete a grammatical sentence: a noun that completes the NP started by the, a verb and an empty-category NP to complete the relative clause, and a verb with subject reporter (arrows). Holding such incomplete dependencies in memory results in a memory storage cost. In (c), structural integration is predicted to occur at time t_n as the head of the right child of the split-rule application (S, red) is attached to its left sibling encountered at an earlier time t_{n-k}. The cost of such integration is estimated as the number of beat-level events (black dots) intervening in the attachment region (arrow). In (d), a constituent head is required to occur at some later time t_{n+k} as the left child of a preparation-rule application (P, green) is encountered at time t_n. In the time span separating t_n from t_{n+k} (arrow), cognitive resources are engaged for implementing such memory storage operation, resulting in increased memory storage cost.

comprehenders access a representation of $x_{t_{n-k}}$ from memory, each intervening word in the attachment region $(t_{n-k},\ t_n)$ is potentially a source of interference that weakens the accessibility of $x_{t_{n-k}}$ and makes the operation more demanding. In particular, the DLT assumes that only words introducing new discourse referents contribute to such interference, as they require additional dedicated cognitive effort compared to words that do not introduce discourse referents (Gibson, 1998). Accordingly, structural integration cost $I(t_n)$ for the attachment of head x_{t_n} is incurred at time t_n and is proportional to the number of intervening discourse referents introduced within the attachment region, i.e., between $x_{t_{n-k}}$ and x_{t_n}.

Memory storage. Finally, memory storage refers to the processing operation of holding incomplete constituents in working memory until their head is encountered in the input string. In the DLT, memory storage cost $S(t)$ at time t is proportional to the number of heads that are expected to come in the future (i.e., among the $x_{t_{n+k}}$ with $k > 0$) in order to form a grammatical sentence (Figure 7.4b).

7.1.4 Computing structural integration and memory storage costs for musical rhythm

Since the generative framework proposed by Rohrmeier (2020a) and overviewed in Section 7.1.2 formalises notions of constituency and headedness in the context of musical rhythm, predictions from the DLT have a natural correspondence in the latter domain. In particular, although the formalism is presented as a phrase-structure grammar, dependencies may be inferred based on the headedness and arity of the grammar rules.[1] The sequence of rhythmic events is assumed to be processed incrementally, with each terminal x_{t_n} being either the head of a constituent, to be integrated into the pre-existing incomplete structural representation, or a non-head child introducing a dependency towards (and, possibly, the expectation for) a constituent head to come in the future.

Discourse processing. First, the discourse processing component is excluded from the model as, differently from linguistic words, rhythmic events are not expected to map to discourse referents through semantic associations.

Structural integration. A structural integration cost is expected to be incurred at time t_n if terminal x_{t_n} is the head of a constituent that is generated as the span of the right child of a split rule application, to be attached to the head $x_{t_{n-k}}$ of its left sibling (Figure 7.4c). Similarly to the linguistic case, we assume that the accessibility of $x_{t_{n-k}}$ – which is required for executing the attachment – decays for $t > t_{n-k}$ due to the interference of intervening rhythmic events. While all sounded events are likely to contribute to such interference to different degrees,

[1] It may have been possible to introduce the entire formalism directly in the form of a dependency grammar: we rather adhered to the original formalism to facilitate comparisons and references to Rohrmeier (2020a).

depending on their salience, as a first approximation we only consider the effect of events that lie on metrical beats. Such events are likely to be structurally more important, more expected, and are aligned with peaks of heightened attention and behavioural entrainment (Fitzroy and Sanders, 2015; Large and Jones, 1999). Accordingly, we quantify the effect of the non-locality of the attachment as the number of events occurring on metrical beats within the attachment region.[2] In other words, let x_{t_n} and $x_{t_{n-k}}$ be the heads of the constituents generated as the spans of the right child and of the left child of a single split-rule application, respectively. Let b be the periodicity of the metrical beat (e.g., in 4/4 meter, the duration of a quarter note), so that $B = \left\{x_{t_n} \middle| \, t_n = \hat{t} + jb \text{ for some } j \in \mathbb{N}\right\}$, where $\hat{t}$ is the temporal location of the first downbeat of the first bar of the rhythmic excerpt, is the subset of sounded events that coincide with metrical beats. We then compute

$$I(t_n) = \left|\left\{x_{t_{n'}} \in B \,|\, n-k < n' < n\right\}\right|,$$

where $|\cdot|$ indicates set cardinality.

Memory storage. A non-zero memory storage cost is incurred whenever constituent heads are predicted to occur as the right child of a rule application. At the time the head of a split is encountered, there is no syntactic cue that makes the occurrence of the right child required or expected. Accordingly, the expectation for a constituent head to come as the right child of a rule application is only introduced by preparation rules. In order to quantify memory storage costs, given a derivation tree, let Π be the set of non-terminal nodes in the derivation tree where a preparation rule is applied. For every such node $v \in \Pi$, let $x_{t_{n_v}}$ be the leftmost terminal of the span of the preparation's left child and $x_{t_{n_v+k_v}}$ be the head of the span of the preparation's right child. Let also $H_t = \left\{x_{t_{n_v+k_v}} | v \in \Pi, \; t \in \left(t_{n_v}, \; t_{n_v+k_v}\right)\right\}$ be the set of the heads of all preparation dependencies that are open at time t. The memory storage cost at time t is then estimated as $S(t) = |H_t|$: in other words, we look at all right-headed (i.e., preparation) dependencies that are open at time t and simply count the number of their distinct heads (Figure 7.4d).

In summary, a structural integration cost is expected at the time the right child of a split is encountered; increased memory storage cost is expected in the time span separating an upbeat to its downbeat; in this framework, syncopations do not introduce additional integration or storage costs but displace the location in time where such costs are incurred.

Based on this framework, it is possible to formulate moment-by-moment predictions about the time-course of construing rhythmic interpretation, in terms of the involvement of cognitive resources pertaining to structural integration and memory storage. These predictions can then be tested in terms of their agreement with behavioural responses that rely on cognitive processes that may interfere with rhythmic interpretation. In this study, we make use of an

[2] Additional criteria may be needed for dependencies linking events below the beat level, but this is not going to play a role in the following.

adaptation of click-detection paradigms (J. A. Fodor and Bever, 1965), as detailed in the next section.

7.1.5 Clicks and parsing in language and music

Click-reproduction and click-reaction responses have been fruitfully employed in psycholinguistic research to investigate the cognitive reality of sentence structure in language processing. In a typical task, participants are aurally presented with a click while they listen to a sentence, and are either asked to immediately react to the click (click-reaction response) or instructed to memorise and subsequently report the location of the click relative to the sentence (click-reproduction response). The clicks' reported location and the reaction times have been shown to be consistently influenced by the constituent structure of the underlying sentence. In particular, reaction times were faster in the breaks between clauses than within clauses (Bever et al., 1969; Holmes and Forster, 1970), suggesting that competition for cognitive resources while parsing clauses impedes fast and accurate perceptual encoding of clicks (however, cf. Abrams and Bever, 1969). Furthermore, the location of clicks that objectively occurred inside constituents was reported as shifted towards constituent boundaries (J. A. Fodor and Bever, 1965; Holmes and Forster, 1972) irrespectively of language-specific prosodic features (van Ooyen et al., 1993), suggesting that constituents are perceptually robust to interruption (Ladefoged and Broadbent, 1960). A general tendency to report clicks earlier than their objective location was also observed: this may indicate that the perceptual encoding of linguistic materials is to some extent delayed while it is integrated into a complete constituent, whereas non-linguistic stimuli such as clicks are encoded faster – hence, appearing to have occurred earlier – relative to the surrounding linguistic materials (Holmes and Forster, 1972).

Both types of click-detection paradigms have also been employed to investigate structural features of music, both in terms of perceptual segmentation as well as online processing. Listeners have been shown to displace clicks towards phrase boundaries in melodies (Sloboda and Gregory, 1980; Stoffer, 1985) and to react faster to clicks between (rather than within) prolongational units in ecological musical stimuli (Martínez, 2018). Furthermore, musicians' reaction times to clicks are significantly slowed down by the co-occurrence of an unprepared change of tonality, or modulation (Berent and Perfetti, 1993). Such modulations are hypothesised to introduce an increased difficulty in terms of structural integration, as the function of the chords following the modulation has to be interpreted under the new tonal context rather than the old one. Accordingly, click reaction has proven to be a sensitive proxy for the occurrence of cognitive operations implementing complex syntactic processing in music.

Overall, the effects associated with click detection in both language and music support that syntactic processing interacts with other perceptual and cognitive operations, such as the detection and perceptual encoding of simultaneous stimuli, and that syntactic constituents are cognitively relevant as both *loci* of syntactic processing as well as representational units. In this study, we adopt a variation of the click-reproduction paradigm where auditory clicks are replaced by short visual flashes. This aims at avoiding that clicks, which may be mistaken for percussive sounds, would be integrated in the underlying musical rhythm rather than processed as extraneous stimuli.

Our working hypothesis is that the interpretation of rhythm forms part of how rhythms in the Western tonal idiom are cognitively represented, and that such interpretation emerges through the process of parsing the musical surface. Note that, differently from a reaction task, in a (flash-)reproduction task listeners are forced to construe a representation of the underlying rhythm and to assess the position of the flash relative to such representation, in order to be able to reproduce the location of the flash. Hence, based on prior evidence from language and music, we hypothesise that the reported location of extraneous stimuli – such as flashes – presented while listening to musical rhythms would reflect the characteristics of the processing operations occurring in the proximity of the objective location of the flashes themselves as listeners construe a representation of the rhythms.

7.1.6 Aims and hypotheses

In this study, based on the theoretical apparatus outlined in Sections 7.1.2-7.1.4, we demonstrate an empirical approach for investigating whether rhythms are processed by listeners into interpretations consistently with a rule-based parsing model. To this end, we compared flash-reproduction performance in different rhythmic stimuli, each suggesting a different rhythmic interpretation. Based on the prior evidence reviewed in the previous Section 7.1.5, we expect a *syntactic* flash-migration effect due to the different syntactic interpretations of the rhythmic stimuli. Specifically, we hypothesise that the reported location of flashes overlapping with a rhythmic event x_{t_n} would be influenced by the processing operations pertaining to the structural integration of that event with the pre-existing incomplete parse. Similarly, if terminal $x_{t_{n+k}}$ is the head of a constituent beginning with terminal x_{t_n}, we hypothesise that the reported location of flashes occurring at $t_n \leq t < t_{n+k}$ would be influenced by the cognitive operations pertaining to memory storage. We assume the magnitude of such effects to be monotonically related to the structural integration cost $I(t_n)$ and to the memory storage cost $S(t)$ computed at the objective flash locations.

However, it is likely that reported flash position is also influenced by non-structural factors. Specifically, we expect the reported flash position to be influenced by the proximity to sounded events in the underlying rhythms, as well as by the position relative to the metrical grid. The proximity of sounded events, on one hand, may influence the participants responses because of foreperiod (Niemi and Näätänen, 1981) or cross-modal attentional blink effects (Arnell, 2006; Arnell and Jolicœur, 1999; however, cf. Soto-Faraco and Spence, 2002). On the other hand, the encoding of flashes may be influenced by heightened or reduced attention due to meter (Large and Jones, 1999), as well as by the attraction to metrically strong locations. Furthermore, both sounded events and metrical beats may provide helpful temporal reference points for the perceptual encoding of each flash. Any systematic bias in the reported flash location merely due to the relative proximity to sounded events and metrical beats would then represent a *non-syntactic* flash-migration effect. The predictive value of a syntactic flash-migration effect is expected to be incremental relative to its non-syntactic counterparts: in other words, we expect syntactic predictors encoding processing costs to explain a proportion of the variance that is not captured by non-syntactic predictors such as the proximity to sounded events or metrical beats. This would be consistent with the cognitive relevance of the assumed processing model, and may form the basis for future work in this direction.

7.2 Methods

7.2.1 Participants

One hundred participants (mean age 28.5, $SD = 8.4, \text{min} = 18, max = 55$) with self-reported normal or corrected-to-normal hearing were recruited through Prolific Academic to take part in an online experimental session. Participants were reimbursed with 12CHF for their participation. The study was approved by the Human Research Ethics Committee of the École Polytechnique Fédérale de Lausanne (HREC 078-2021) and was conducted in accordance with the declaration of Helsinki. Data from 11 participants were excluded from the analysis due to technical issues with the user interface, and data were recollected after recruiting new participants to reach the final number of 100 participants. The average degree of musical training, quantified by the corresponding subscale (ranging from 1 to 7) of the Goldsmith Music Sophistication Index (Müllensiefen et al., 2014), was 2.59 ($SD = 1.41, \text{min} = 1, \text{max} = 6$). A comparison with the sample of the general population from English-speaking countries tested by Müllensiefen et al. (2014), who reported an average score of 3.79 ($SD = 1.63$), suggests a relatively low degree of explicit musical expertise among our participants.

7.2.2 Stimuli

Three rhythmic excerpts spanning 2 bars in 4/4 meter were designed by the authors with the goal of exemplifying the three elementary types of rhythmic interpretation. All three stimuli were structurally composed of two major constituents, A and B, generated by a split-rule application to the root node and roughly spanning the first and the second bar respectively (Figure 7.5). All stimuli also shared the positioning of one specific event, that we term the *critical event*, on the second sixteenth-note subdivision of the last beat of the first bar (Figure 7.5, dashed line).

Crucially, the interpretation of this event in terms of its syntactic relatedness to A and B was different in the three stimuli. In the Split stimulus, the critical event represented the right child of a (left-headed) split rule application, and further marked the end of constituent A (Figure 7.5a). In both the Syncopation and Preparation stimuli, instead, the critical event rather grouped forwards as the first event of constituent B. However, the interpretation of the critical event was also different in the Syncopation and Preparation stimuli. In the former case,

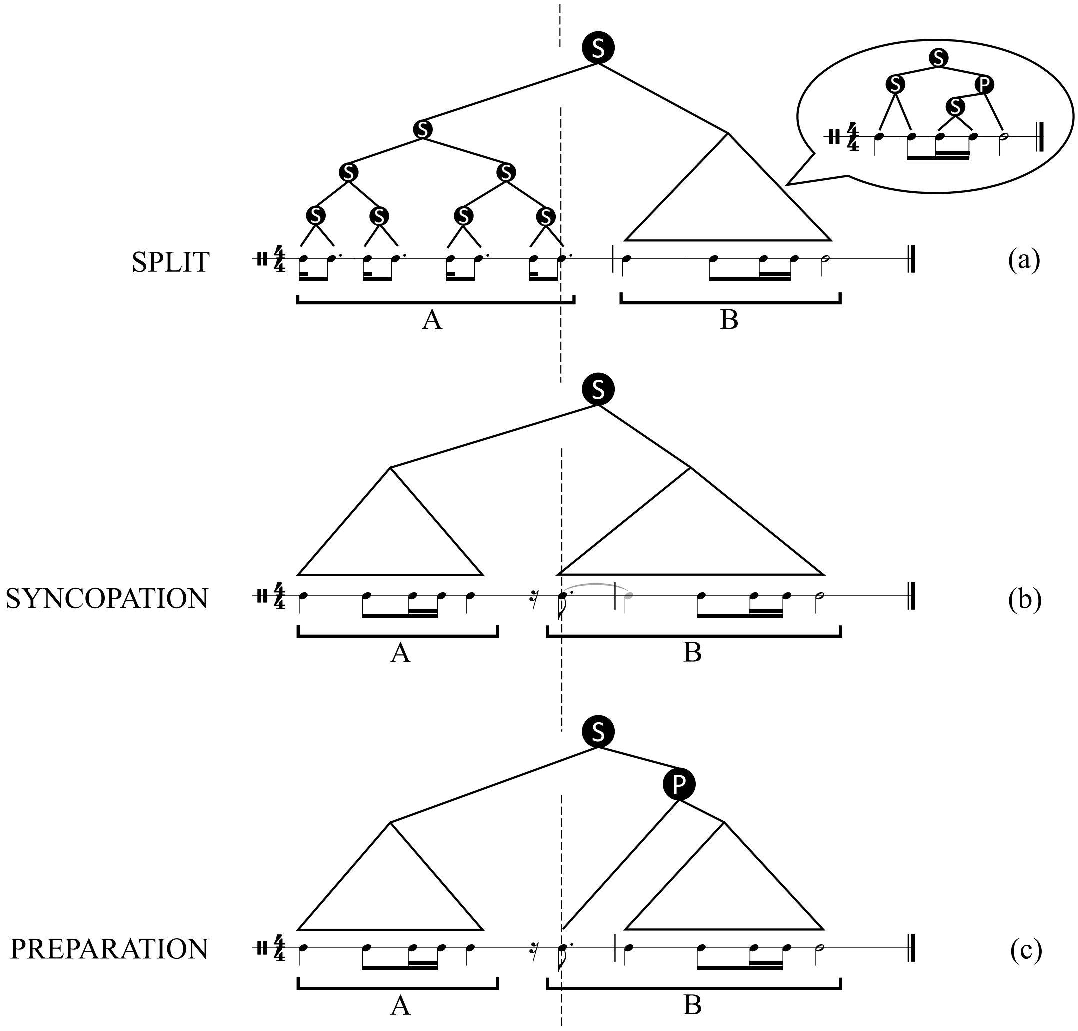


Figure 7.5 – Schematic derivations for each one of the three stimuli (full-fledged derivations after Rohrmeier (2020a) can be found in Supplementary Material S1 as Figures S2-S4). Each stimulus comprises two main constituents, A and B, and is based on a 1-bar template. The inner structure of the template is replaced here by a triangle and is exemplified separately in the balloon. In the Split stimulus (a), the template is used as constituent B. In the Syncopation stimulus (b), the template is used as both constituent A and constituent B, but the first event of constituent B is anticipated (the note on the downbeat of the second bar, grayed out, is tied to the preceding note and is not sounded). In the Preparation stimulus (c), the template is used as both constituent A and constituent B, but an upbeat is added through the application of a preparation rule. Note that in all three stimuli a critical event occurs in the same metrical position (dashed line), and that all dotted eight-notes exhibit a syncopation (i.e., they are anticipated by the duration of a sixteenth-note; to avoid cluttering the figure, this low-level syncopation is not marked explicitly in the derivations).

the critical event was a syncopated (i.e., displaced) instance of the downbeat of the second bar, thus acting itself as the head of constituent B (Figure 7.5b). In the latter case, the critical event acted as the left-branching upbeat of the downbeat of the second bar, the head of constituent B (Figure 7.5c).

The position of the critical event was then determined so as to make it possible to manipulate the surrounding context in a way that would lead listeners to attribute to it the intended interpretation in the three stimuli. In particular, the critical event was placed early enough to make it plausible for listeners, given the appropriate context, to hear it as grouping unambiguously backwards with constituent A (as in the Split stimulus), yet late enough to be possibly perceived as a displaced instance of the following downbeat (as in the Syncopation stimulus).

The regions of the stimuli preceding and following the critical event were then designed to increase the plausibility of the proposed interpretation. A rhythmic template was adopted as constituent B in all stimuli, with the downbeat being omitted (and replaced by the critical event as discussed above) in the Syncopation stimulus. The same template was adopted as constituent A in both the Preparation and the Syncopation stimulus. The parallelism between constituent A and constituent B was meant to support an analogous interpretation for the two constituents (Lerdahl and Jackendoff, 1983a), with the critical event being understood as a rhythmic elaboration as an upbeat or a syncopation as discussed above. In order to strengthen the interpretation of the critical event as a backwards-grouping event, constituent A in the Split stimulus was instead designed as a sequence of four Lombard-rhythm cells (a metrically accented short note followed by a metrically weak longer note, ∪́–), one per beat. The critical event, as the last event of the sequence, was then closest to the immediately preceding, metrically accented event, thus being likely grouped with it due to the *Gestalt* principle of proximity (Deutsch, 1999).

A one-bar metronome count-in consisting of 4 quarter notes was prepended to each rhythmic excerpt, allowing listeners to prepare for the task and to entrain to the tempo and meter. In particular, this ensured that strong and weak metrical positions were inferred consistently in each stimulus, contrasting the possible perceptual accents associated with long durations relative to short ones in the Lombard rhythm (Povel and Okkerman, 1981). All three stimuli were finally synthesised using the MuseScore 3 MIDI library. In order to mark the distinction between the count-in and the actual rhythmic excerpt, the General MIDI High Woodblock timbre was adopted for the former and the Low Woodblock timbre for the latter. Each complete stimulus comprised then three bars played at 100bpm, for a total duration of 7.2s each.

7.2.3 Flash-reproduction task

The different interpretations suggested by the three stimuli predict different estimated processing costs in the proximity of the critical event and the following downbeat, as detailed in Section 7.2.5. This region of interest was covered by five evenly spaced flash positions (numbered 0 to 4) separated by 150ms, i.e., the duration of a sixteenth-note at the given

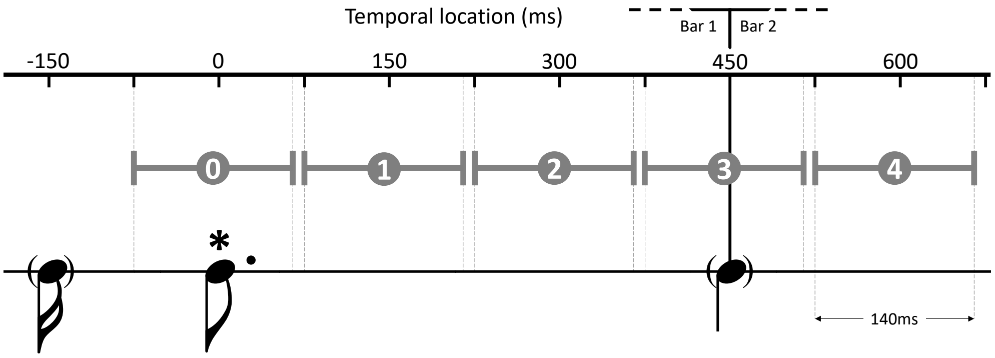


Figure 7.6 – Temporal placement of the five visual flashes (numbered 0 to 4) relative to the critical event (marked by *).

tempo (Figure 7.6). Flash onsets thus occurred halfway between two metrical positions at the sixteenth-note level. Each flash, a white circle against black background with a diameter spanning 80% of the screen height, was presented at the centre of the screen in one of these five temporal positions for a duration of 140ms. Accordingly, flashes in position 0 started just before, and overlapped with, the critical event, and flashes in flash position 3 started just before, and overlapped with, the following downbeat.

A trial consisted of two consecutive presentations of the same stimulus, numbered 1 and 2. During Presentation 1, a flash in one of the five positions was presented, and participants were instructed to memorise its location relative to the underlying rhythm. During Presentation 2 no flash was presented: instead, participants were instructed to indicate with a key press the position in time where they remembered the flash to have occurred during Presentation 1. Presentation 2 followed Presentation 1 without interruption, in order to preserve the metrical pulse, and the two presentations were distinguished by the colour of the screen background (black and grey, respectively). The succession of trials was self-paced, as participants could determine with a key press when to start a new trial.

7.2.4 General procedure

The experiment took place online and consisted of two parts. In the first part, participants engaged in the flash-reproduction task described in Section 7.2.3, implemented with the software PsychoPy (Peirce et al., 2019) and hosted on the platform Pavlovia.org. Trials were divided into 6 blocks of 30 trials each. All trials in a block featured the same stimulus, so that each stimulus was adopted in two blocks. The design by blocks was meant to ensure that listeners would consistently form the same interpretation for every presentation of each stimulus. In particular, we intended to avoid effects of surprise and retrospective reanalysis for stimuli sharing the same beginning (e.g., preparation and syncopation). For the same reason, before each block, participants were exposed to one presentation of the rhythm that would be employed in all trials in that block, so that listeners were not surprised by the resolution of ambiguity halfway through a stimulus. For each participant, the blocks' order was chosen uniformly at random under the constraint that two consecutive blocks were not associated with the same stimulus. In each block, 6 identical trials for each of the 5 possible flash locations were presented in random order. In the second part of the experiment, the Musical Training, Perceptual Abilities, and Emotion subscales of the Goldsmith Music Sophistication Index questionnaire (Müllensiefen et al., 2014) were administered. The whole experiment, including consent, instructions, and three practice trials, lasted between 60 and 70 minutes.

7.2.5 Estimates of processing costs

In each stimulus, we compute the structural integration and the memory storage costs after the definitions given in Sections 7.1.2 and 7.1.3. Structural integration costs, computed after the corresponding equation in Section 7.1.4, are only incurred when a flash overlaps with a rhythmic event that happens to be the head of a constituent, to be integrated in a pre-existing parse. In the Split stimulus (Figure 7.7a), this occurs at flash position 0, which overlaps with the critical event, and flash position 3, which overlaps with the downbeat event in bar 2. The critical event attaches to the immediately preceding event, resulting in no predicted integration costs as no events intervene in the span of the attachment. The downbeat of bar 2 is the head of constituent B and attaches backwards to the downbeat of bar 1, the head of constituent A, with 3 beats intervening in between. Since each one of these 3 beats hosts a sounded event, the structural integration cost at flash position 3 is 3.

In the Syncopation stimulus (Figure 7.7b), the critical event is interpreted as a displaced occurrence of the head of constituent B, hence it integrates backwards to the head of constituent A. The separation is again 3 beats, but only two of them host sounded events, resulting in a structural integration cost of 2 at flash position 0 that overlaps with the critical event.

In the Preparation stimulus (Figure 7.7c), non-zero structural integration costs are incurred at flash position 3, which overlaps with the downbeat of bar 2 attaching backwards to the downbeat of bar 1. No structural integration costs are incurred in flash position 0, overlapping with the critical event, as the latter is here the non-head child of a preparation rule application. Instead, click positions 1, 2, and 3 occurring between the critical event and the following downbeat incur in non-zero memory storage costs: after the critical event is encountered and interpreted as an upbeat, the occurrence of the downbeat is syntactically required to form a grammatical rhythm. As one constituent head is expected, such memory storage cost is unitary.

7.2.6 Analysis

We encoded flash migration as the difference δ between the reported location of the flash as recorded by the participants' key press and the objective onset of the flash as presented. The bi-modal presentation of auditory rhythmic stimuli and visual flashes may introduce

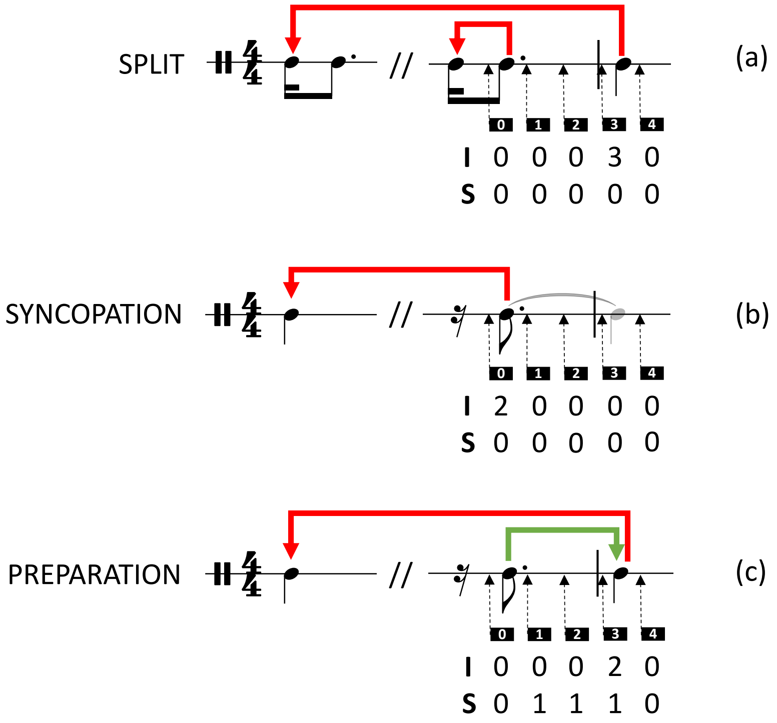


Figure 7.7 – Predicted structural integration (I) and memory storage costs (S) for each flash position (labelled 0 to 4) in the Split (a), Preparation (b), and Syncopation (c) stimuli. Red arrows indicate dependencies between the children of a (left-headed) split rule application, green arrows indicate dependencies between the children of a (right-headed) preparation rule application. In the latter case, all temporal locations comprised between the event that opens the (preparation) dependency and the event that closes it bear a memory-storage processing cost. In the case of our Preparation stimulus (c), this includes flash locations 1–3, that just happen to occur in the span of the preparation dependency (green arrow).

systematic but unknown offsets in the reported flash positions, for example due to slower processing of visual stimuli relative to simultaneous auditory stimuli (Robinson et al., 2018). As a consequence, we shall not interpret or analyse the flash migration in absolute terms (e.g., flashes being reported absolutely earlier, $\delta < 0$, or later, $\delta > 0$, relative to their objective locations). In our analyses, we rather considered relative changes in δ across conditions. All non-categorical variables were standardized to null mean and unit standard deviation σ prior to analysis, while removing 55 datapoints (0.3%) with missing responses and further excluding 766 datapoints (4.3%) with $|\delta| > 1.96\sigma$ as outliers, resulting in 17179 observations being included in the analysis. Data were then analysed with Bayesian mixed-effects models provided with weakly informative priors (t(3,0,1); Gelman et al., 2008) and implemented in the R package *brms* (Bürkner, 2018). For individual coefficient estimates (β) we report the estimated error (EE) as well as evidence ratios (*Odds*) for the regression coefficients or some function of the regression coefficients to be strictly larger or smaller than zero. From a frequentist perspective, evidence ratios can be interpreted as significant (*) at a .05 confidence level when exceeding 38. Data and analyses are available as supplementary materials S2 and

S3, respectively, at https://osf.io/gb2pc/?view_only=404edbe75c35431ba3b8fb1a2d9df5d9.

Differences across flash positions and stimuli

As a preliminary exploration of the data, we tested whether participants' responses differed across conditions, i.e., across stimuli and flash positions. In particular, observing differences across stimuli in each flash position would be consistent with the hypothesis that different interpretations of the auditory rhythmic stimuli interact with the visually-presented flashes. Pairwise differences between conditions were statistically evaluated with the Bayesian mixed-effects model

$$\begin{gathered}\delta \sim Flash + Stimulus + Flash \times Stimulus + \\ (1|\text{TrialWithinBlock}) + (1|\text{Block}) + (1|Participant)\end{gathered}$$

which captures the effect of flash position (a categorical variable with 5 levels) in each stimulus (a categorical variable with levels Split, Preparation and Syncopation) while controlling for possible effects of learning and fatigue (with random intercepts by block index, 1 to 6, and trial number within a block, 1 to 30), and participant index.

Model comparison: syntactic and non-syntactic predictors.

We then tested the relative contribution of syntactic and non-syntactic predictors to the observed flash migration. Bayesian mixed-effects models predicting flash displacement were implemented with different types of predictors. A baseline model M_{Baseline} was defined as

$$\delta \sim (1|Flash) + (1|\text{TrialWithinBlock}) + (1|\text{Block}) + \left(1\middle|\text{Participant}\right), \tag{7.1}$$

where the random intercept by Flash position, encoded as a categorical variable, accounts for any effects that are constant across stimuli in each flash position. This includes a possible serial-order effect across flash positions (for example, the central-tendency bias observed in Section 7.3.1) as well the effect of the metrical grid: in fact, all stimuli shared a 4/4 metrical grid, reinforced by the count-in bar at the beginning of each stimulus, so that each flash position occupied the same location relative to the metrical grid in all three stimuli. We then implemented models capturing syntactic as well as two types of non-syntactic predictors (sequential and metric).

In model $M_{\text{Sequential}}$, additional predictors accounting for the proximity of Flashes to sounded events in the rhythmic stimuli were added to M_{Baseline}. Specifically, if $\{x_{t_n}\}_n$ is the sequence of events in a given stimulus, for a flash presented at time $x_{t_n} < t < x_{t_{n+1}}$ we computed the ratio $\rho = \frac{|t-t_n|}{|t-t_{n+1}|}$ indicating to what extent the flash location was closer to the preceding ($\rho < 1$) or the following ($\rho > 1$) sounded event. We also computed the size of the silent region surrounding the flash as the temporal distance $\gamma = t_{n+1} - t_n$ between the preceding and the following event. Model $M_{\text{Sequential}}$ then accounts for the effect of proximity to sounded events in terms of the term $\rho + \gamma + \rho \times \gamma$, to be added to M_{Baseline}: this reflects how δ is influenced by the relative proximity to the preceding or following event, and how much this effect is modulated by the absolute size of the involved time intervals.

In model M_{Metric} we accounted for the possible effects of metrical entrainment, and for the possibility that flashes are attracted towards metrical beats. We estimated the overall strength of metric entrainment for each stimulus by means of a state-of-the-art pulse-clarity metric evaluated on the audio files of the stimuli (including the initial metronome), after Pironio et al. (2021). We then quantified an attraction coefficient modelling attraction of the flash towards the nearest metrical beat relative to the location of the flash. The magnitude of the attraction was modelled as proportional to the metrical weight of the nearest beat (quantified arbitrarily as 1 for the quarter-note level, 2 for the bar level). The directionality of the attraction was modelled through the sign of the attraction coefficient (negative if the flash followed the nearest beat, positive if the flash preceded the metrical beat). Model M_{Metric} included then the full interaction term $PulseClarity \times AttractionCoefficient$.

In model $M_{\text{Syntactic}}$, the effect of processing operations on flash migration was accounted for by adding structural integration cost I and memory storage cost S at a given flash position, as well as their interaction $I \times S$, as additional monotonic ordinal predictors (Bürkner and Charpentier, 2020) into M_{Baseline} (see Figure 7.7). Note that, in the DLT, structural integration and memory storage costs are assumed to contribute additively to the behavioural effects. Since we are generalising dependency-locality principles to a different domain, we remain agnostic in this respect and rather model the contributions from the two processing operations independently, leaving it as an empirical problem to determine whether the coefficients of the two contributions are the same (consistently with the DLT predictions in language) or not.

Finally, models $M_{Syntactic+Metric}$, $M_{Sequential+Metric}$, $M_{Syntactic+Sequential}$, and $M_{Syntactic+Sequential+Metric}$ were defined as including combinations of the aforementioned fixed effects. The performance of all models was then compared under leave-one-out cross-validation with Pareto-smoothed importance sampling (PSIS-LOO; Vehtari et al., 2017). Differences in the estimated out-of-sample predictive fit (expected log pointwise predictive density, *elpd*) quantify the extent to which adding or removing a predictor results in capturing a greater proportion of the data's variance beyond that which is simply justified by the sheer number of parameters. As a consequence, a model achieves higher *elpd* than another if its predictors are both more parsimonious and effective.

Contribution of structural integration and memory storage.

Once assessed whether syntactic predictors contribute significantly to the observed flash migration, we investigated the individual effects of structural integration and memory storage, as well as of the non-syntactic predictors. This was achieved by inspecting the corresponding coefficient estimates in the best-performing model as identified by the model-comparison approach.

7.3 Results

7.3.1 Differences across flash positions and stimuli

Figure 7.8 shows the average displacement δ for each flash position and each stimulus across all participants. A general trend for decreasing δ from early to late flash positions is observable. This may reflect a tendency for listeners to report clicks towards the middle of the region of interest (Hollingworth, 1910). The evidence ratios estimated by the model presented in Section 7.2.6, reported in Figure 7.8, support with strong evidence that, in each flash position, the displacement of the reported location of the flash relative to its objective location differed across stimuli. This observation rules out that features that are common across stimuli, such as the underlying metrical grid, may account on their own for the listeners' behaviour. Listeners' responses may rather be influenced by stimulus-specific features, such as the specific location of sounded events and their syntactic relatedness. The relevance of syntactic and non-syntactic predictors was then investigated with a model-comparison approach.

7.3.2 Model comparison: syntactic and non-syntactic predictors

Results of the model comparison introduced in Section 7.2.6 are reported in Figure 7.9. The best performing model $M_{Syntactic+Sequential+Metric}$ comprised both syntactic and non-syntactic predictors; individually, syntactic (as formalised in $M_{\text{Syntactic}}$) and non-syntactic (as formalised in $M_{Sequential+Metric}$) predictors both outperformed the baseline model. This suggests that both types of stimulus features contributed to shaping the observed flash migration effects. Syntactic predictors alone accounted for a larger proportion of the variance compared to the baseline model ($\Delta elpd = -50.5$, $SE = 10.1$) and to sequential predictors alone ($\Delta elpd = -28.4$, $SE = 7.7$). Furthermore, the latter had little to no incremental predictive power over the syntactic predictors, as reflected in the non-significant difference between $M_{Syntactic+Sequential}$ and $M_{\text{Syntactic}}$ ($\Delta elpd = -2.0$, $SE = 3.4$). Attraction to meter further accounted for an independent proportion of the variance, the best performing model being the one including all syntactic, sequential, and metrical predictors ($\Delta elpd = 74.9$, $SE = 12.9$ over the baseline model). Overall, models including syntactic predictors outperformed models that only included non-syntactic (sequential or metrical) ones, as indicated by the red compared to the black datapoints in Figure 7.9. This supports the hypothesis that processing operations involved in parsing rhythms into an interpretation carry predictive value towards flash migration.

7.3.3 Contribution of structural integration and memory storage

The effects of syntactic and non-syntactic predictors were quantified through the corresponding coefficients in the top-performing model $M_{Syntactic+Sequential+Metric}$ from Section 7.3.2, as discussed in Section 7.2.6. The distributions of the estimated effects of all predictors are displayed in Figure 7.10. Strong evidence supports that both structural in-

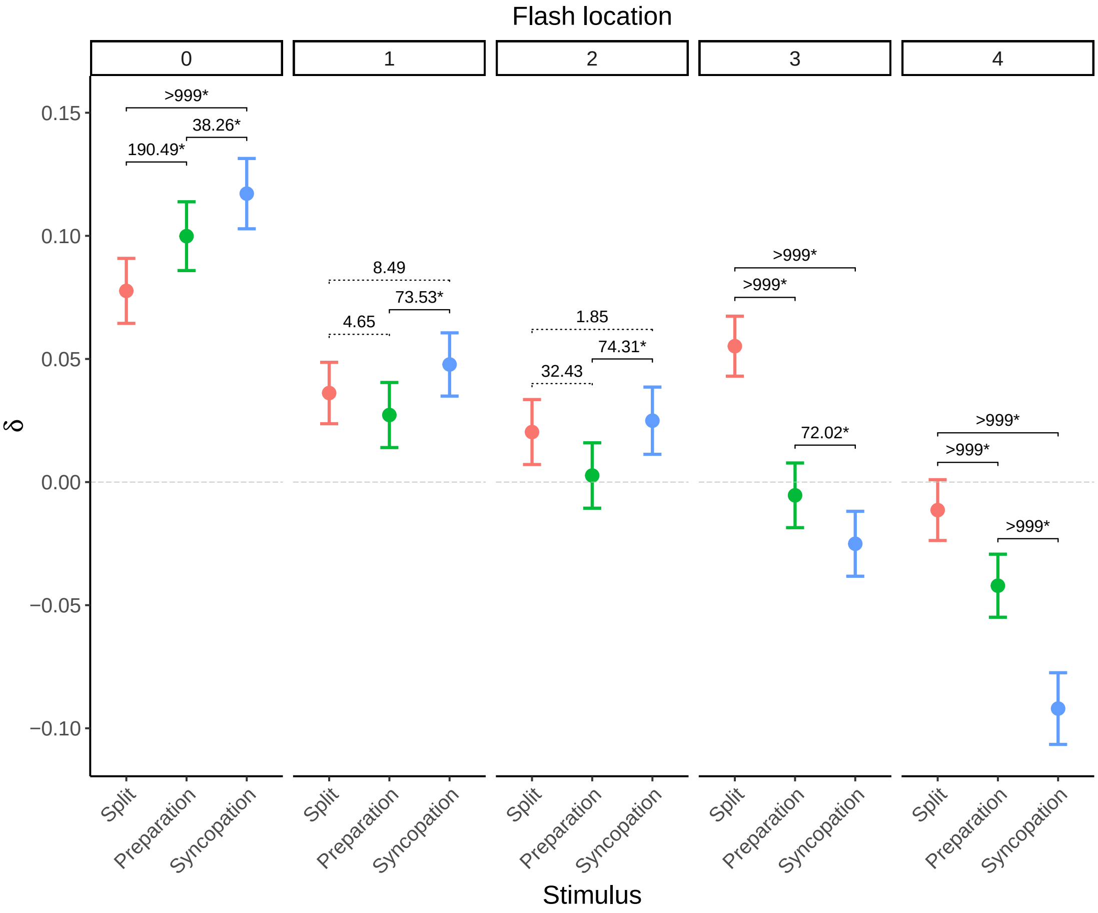


Figure 7.8 – Average flash displacement δ (in seconds) for each Flash position (0 to 4) in Split (red), Preparation (green), and Syncopation (blue) stimuli. Error bars indicate 95% confidence intervals around the mean. For each pair of stimuli, we also report the evidence ratios for the corresponding flash displacements in each position to be different (asterisks mark differences that can be interpreted as significant from a frequentist perspective).

tegration and memory storage costs predicted flash migration with independent contributions, which were further distinguishable by their opposite directionality. Specifically, structural integration costs were found to correspond to displacements in the "late" direction ($\beta = 0.10,\ EE = 0.03,\ Odds\left(\beta > 0\right) > 9999^*$), whereas memory storage costs were associated with displacements in the "early" direction ($\beta = 0.05,\ EE = 0.02,\ Odds\left(\beta > 0\right) = 1749^*$). This suggests that both structural integration and memory storage played a role in shaping listeners' behavioural responses. Only little evidence was found for an interaction effect ($\beta = -0.09,\ EE = 0.14,\ Odds\left(\beta > 0\right) = 23.58$). Strong evidence was found that beat clarity influenced the flash migration ($\beta = 0.13,\ EE = 0.05,\ Odds\left(\beta > 0\right) = 433.44^*$), but no significant evidence was found for effects of other non-syntactic predictors (all Odds$\left(\beta \gtrless 0\right) < 38$).

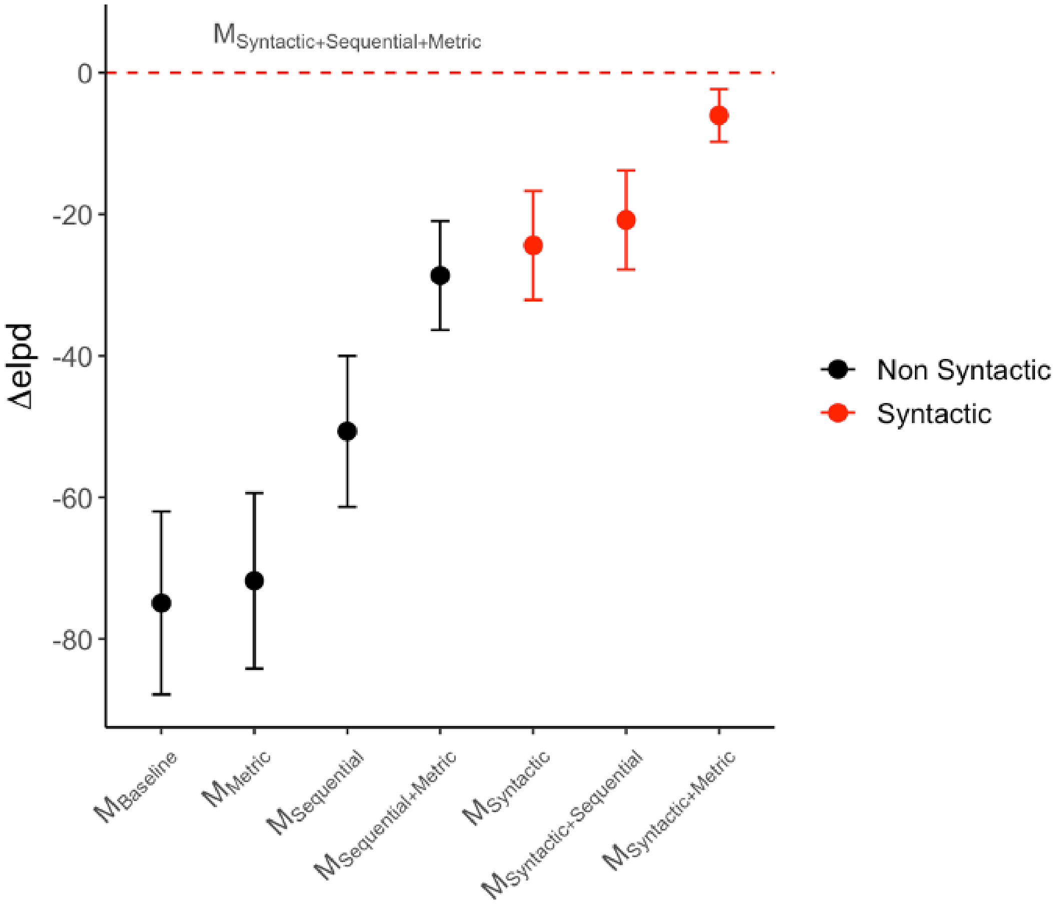


Figure 7.9 – Difference in estimated out-of-sample predictive fit (expected log pointwise predictive density, elpd) of models $M_{\text{Syntactic}}$, $M_{\text{Sequential}}$, M_{Metric} and M_{Baseline}, as well as their combinations, relative to the best performing model $M_{Syntactic+Sequential+Metric}$. Dots and error bars indicate means and their standard error, respectively. Models including syntactic predictors (red) outperform models that do not include syntactic predictors (black).

7.4 Discussion

In this paper, we proposed a framework to investigate whether syntactic-parsing processes contribute to the interpretation of rhythmic structure during listening, and exemplified the approach with a small-scale behavioural experiment. Following the proposal by Rohrmeier (2020a), the underlying syntactic competence was modelled as an Abstract Context-Free Grammar formalising three elementary rhythmic relationships as (families of) generative rules: split, preparation, and syncopation. The process of rhythmic interpretation was then understood as constructing a derivation of a given rhythm incrementally, integrating newly encountered events into pre-existing partial derivations that are stored in memory. As a first empirical test, the displacement between the reported location and the objective location of visual flashes presented while listeners were attending rhythmic stimuli was quantified in a behavioural

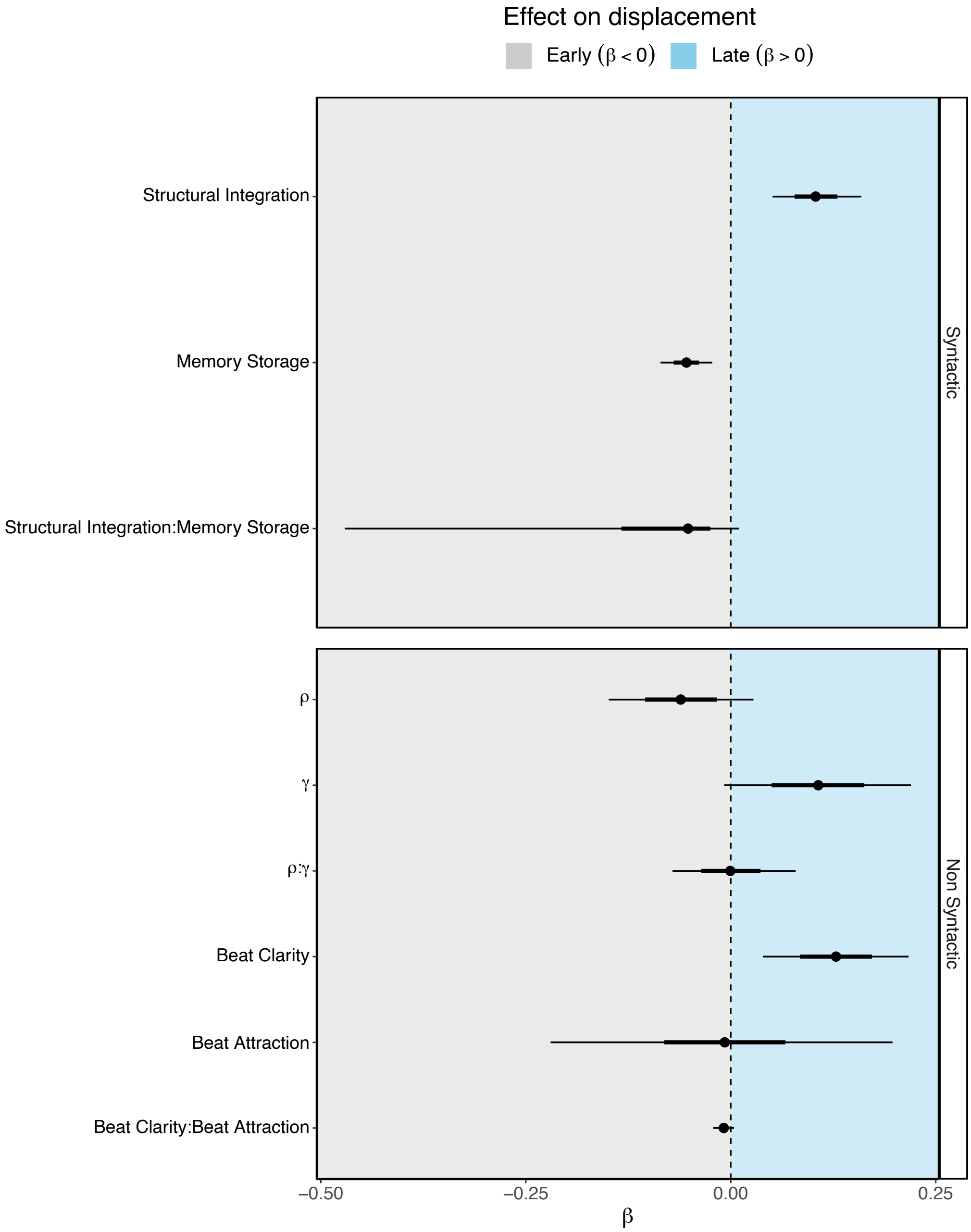


Figure 7.10 – Posterior distributions of the effect β of each predictor on the observed flash displacement, as estimated by model $M_{Syntactic+Sequential+Metric}$. For each distribution, we report the evidence ratio in favor of the hypothesis that the effect is strictly positive or negative. Structural Integration and Memory Storage costs significantly predict displacement in the late (>0) and early (<0) direction, respectively.

task. Strong evidence was found that such displacement was influenced by processing operations as predicted by a model of syntactic processing inspired by psycholinguistic literature (Gibson, 2000). While these results are consistent with the proposed theoretical framework, after controlling for several confounding factors, the small-scale experiment should not be taken on its own as final support of the theory, and rather provides an empirical foundation for future work based on the proposed perspective. In the following, we discuss the significance and limitation of these results, placing the present framework in the context of previous music cognition literature and of ongoing discourses on the generality of cognitive processing across domains such as music and language.

7.4.1 Existence of rhythmic interpretation

The theoretical framework assumes that listeners form a representation of rhythmic structure based on specific syntactic relationships, each being associated with an interpretation: splits as "rebounds", preparations as goal-directed expectancy, and syncopations as temporal displacement. The characteristics of these interpretations are reflected in the properties of the grammar (e.g., the arity and headedness of the rules; Figure 7.2), in the shape and inner-node labelling of the resulting derivation trees (Figure 7.3), and, eventually, in the processing operations that are computationally required to parse a rhythmic surface into such derivation trees (Figure 7.4). We thus interpret the empirical results – within the limits of the small-scale test – as being consistent with the cognitive relevance of rhythmic interpretation, as it is characterised in music theory based on introspection and analysis of the repertoires.

The proposed theoretical framework captures a complementary phenomenology compared to theories that explain aspects of the experience of music in terms of expectation and predictive coding (Rohrmeier and Koelsch, 2012; Vuust et al., 2022) and, in particular, to models of the online processing of grouping and meter. For example, a model like IDyOM (Pearce and Wiggins, 2012) is capable of accounting for the phenomenon of perceived segmentation (among others), characterises it in terms of the predictability of subsequent events (loosely speaking, boundaries occur when a highly predictable event does not afford deterministic predictions about possible continuations), and explains the phenomenon as a result of implicitly acquired statistical regularities (Hansen et al., 2021; Pearce, Ruiz, et al., 2010). While such a model does characterise and explain the percept of group boundaries, hence the "cartography" of grouping, it is not meant to account for the idea that events within a group are linked together by some specific relationship, that different groups may be held together by qualitatively different relations, and that different groups may relate to one another hierarchically. Similarly, the theory of dynamic attending (Large and Jones, 1999) accounts for the phenomenon of metrical strength as perceived, characterises it as periodic peaks of heightened attention, and explains it in terms of neural and behavioural entrainment. While this theory does explain many aspects of what it feels like for an event to *be* in a strong or in a weak metrical position (in terms of, e.g., improved pitch discrimination; Jones et al., 2002), it is not meant to account for the idea of an event being *displaced* relative to its intended metrical location, as in the case of syncopation. The notion that musical events are in specific kinds of relationships with one another (e.g., an event being a preparation of another) and, recursively, with all other events, is instead the focus here. Groupings are then understood as being "held together" by these relations. In this sense, the notion of grouping emerging from a syntactic account differs from other notions of grouping as they are discussed in the literature (cf. Parncutt, 1994), such as periodic grouping (which is based on meter alone) and sequential grouping (based on the temporal proximity of events). The observation that syntactic predictors carry incremental predictive value over those pertaining to meter and sequential proximity supports the meaningfulness of this distinction in perception.

The peculiarity of the present approach is that it explicitly accounts for latent relationships linking events with one another; the observed behavioural responses are then explained as a result of the cognitive availability of a representation of such relationships, or of the process of construing such representation. In particular, syntactic relationships may afford predictions towards future events and underlie part of the phenomenology associated with predictive coding and active inference in music (Patel and Morgan, 2017; Rohrmeier and Koelsch, 2012; Vuust et al., 2022). Our results support this perspective in the context of musical rhythms, enriching prior evidence that integrated representations of melodic and harmonic structure as a whole are formed during listening (Cecchetti et al., 2021; Koelsch et al., 2013; Martínez, 2018; Rohrmeier and Widdess, 2017), and that such representations are even revised retrospectively if necessary to ensure global coherence (Cecchetti et al., 2022). The non-local nature of the hypothesised dependencies further supports that such representations can be hierarchical rather than sequential (Rohrmeier and Pearce, 2018), consistently with prior evidence pertaining to pitch and tonal harmony (Dibben, 1994; Herff, Bonetti, et al., 2023; Herff, Harasim, et al., 2021; Koelsch et al., 2013; Martins et al., 2017; Serafine et al., 1989). Note that this framework does not require, and the present results do not suggest, that the relationships underlying such representations are explicitly available to conscious awareness. It is possible that listeners may experience what it feels like for an event to be, e.g., preparatory towards another without being able to verbalise such experience or its origin. In particular, conscious awareness may well only emerge with training and introspection, and future research may investigate the role of musical expertise in modulating the observed effects.

7.4.2 Structural integration and memory storage implement rhythmic syntactic processing

Results are further consistent with the view that interpreting rhythm involves (at least) two types of processing operations – structural integration and memory storage – each contributing with a distinguishable effect to the observed flash migration. In particular, the occurrence of higher structural integration costs resulted in flashes being reported relatively later. In contrast, flashes occurring in regions with non-zero memory storage cost tended to be reported relatively earlier. By contrast, in the DLT, structural integration and memory storage costs are assumed to contribute additively to overall behavioural effects. The directionality of the observed effects, while not necessitated by the theory, may be explained as reflecting the different nature of the two processing operations, as well as a tendency for structural units to "preserve their structural integrity by resisting interruption" as reported by J. A. Fodor and Bever (1965, p. 415). Structural integration entails to retrieve a preceding event across the attachment region (i.e., the region spanning from the current event to the target of the attachment, lying in the past). Extraneous stimuli (such as a visual flash) that interfere with accessing the earlier event from memory may be perceptually displaced as to be excluded from the attachment region, thus reported later. On the contrary, memory storage entails that listeners keep track of expected future events. A tendency to remove extraneous stimuli from the timespan where such cognitive demands are required would lead to report such interfering stimuli as happening before the expectation is opened, i.e., relatively early compared to the expectancy-inducing event. It has also been suggested (Holmes and Forster, 1972) that the perceptual encoding of a sounded event that initiates a right-headed constituent may be slightly delayed while its head is awaited to form a complete constituent and the currently incomplete constituent is stored in working memory. An extraneous stimulus may then be perceived to have occurred somewhat earlier relative to the perceptually encoded location of the preceding sounded event. Nevertheless, these explanations are speculative and further research is necessary to investigate possible mechanistic accounts for these results.

The moment-by-moment predictions on the complexity of individual processing operations were obtained by adapting a psycholinguistic theory of sentence processing, the Dependency Locality Theory (Gibson, 2000), to the musical domain. In the DLT, the cognitive effort associated with implementing processing operations depends on the reduced accessibility of past events due to the interference of intervening events. The results are consistent with this hypothesis in the musical domain, as well as more generally with the hypothesis that computationally analogous operations to those implemented during sentence comprehension are involved in music processing (Fedorenko et al., 2009; Jackendoff, 1991; Katz and Pesetsky, 2011; Patel, 2010). However, the quantitative estimates of processing complexity adopted in this study should be seen as coarse preliminary heuristics, as they are based on a qualitative analogy with the linguistic model. Future research will need to investigate what factors contribute to the cognitive load associated with processing operations. In particular, the assumption that non-locality could be quantified by the number of intervening beat-level events was motivated by the relevance of metrical beats as a music-theoretically meaningful measurement-unit of temporal distance (Grove Music Online, 2001), as well as markers of salience (Large and Jones, 1999). Nevertheless, this heuristic is certainly too coarse, as it does not account for sub-beat-level events nor hypermeter, and it may need data-driven refinement in the future. Due to the music-theoretical and perceptual salience of metrical beats, though, it is plausible that other measures of non-locality would also correlate to some extent with the one proposed here, at least on the time-scale that was relevant in this study. It is also possible that memory decay due to interference, which underlies the hypothesised processing complexity estimates, may operate differently for words than for musical stimuli, the latter being remarkably robust to such interference (Herff et al., 2019). However, as prior results on the topic only pertain to complete idiomatic melodies (Herff, Olsen, and Dean, 2018; Herff, Olsen, Dean, and Prince, 2018), it is likely that such robustness to interference would not involve individual events or segments smaller than, at least, entire constituents. Furthermore, rhythm-only patterns have been shown not to benefit from such robustness (Herff, Olsen, Prince, et al., 2018), so that this issue may only become relevant when investigating features of music other than rhythm alone.

Future research may also focus on identifying locality thresholds beyond which processing of dependencies fails. Predictors of cognitive load increase monotonically with dependency distance, yet it is likely that listeners are unable to track arbitrarily long dependencies in real-time listening (Cook, 1987). This is not unlikely what happens in language: there, too, comprehension fails or is impeded when processing costs exceed some threshold. However, in language, sentences that challenge such processing threshold are rare, as mutual understanding is typically a speaker's primary concern. On the contrary, in music, the time span of music-theoretically predicted dependencies is potentially longer than any reasonable integration threshold would allow (possibly extending all the way to the duration of entire pieces). This may reflect that the capacity of listeners to form a complete representation of syntactic structure for an entire passage of music during first-pass listening is not a necessary condition for music to fulfill many of its functions (including aesthetic and social ones). On the contrary, it would make little sense for language users to produce sentences that exceed processing threshold, thus being unintelligible in real-time reading or conversation (an exception to this may be certain artistic uses of language in the context of poetry and literature, where ease of production and ease of comprehension during first pass reading/listening is not a priority). For example, listeners may still appreciate music by only perceiving somewhat local relations (i.e., only representing disconnected sub-trees for different portions of the piece). Nevertheless, listeners with a higher degree of expertise may be able to integrate over larger and larger segments of the piece, possibly relying on multiple hearings and explicit domain knowledge. As a consequence, composers may still find it valuable to embed complex structure over large timespans in their pieces (e.g., in terms of hypermeter or form), for listeners to potentially "discover" them. It exceeds the scope of the present article to make specific predictions in this respect, but investigating processing-breakdown thresholds represents an open empirical question that future research may address (also) in light of the modelling approach proposed here.

7.4.3 Syntactic predictors outperform non-syntactic predictors

In addition to the hypothesised processing operations, the reported position of the flashes were modelled in terms of several other factors that could account for effects of proximity to other sounded events, or differences in metrical entrainment across the stimuli. However, such local effects failed to fully explain the observed data, and the predictive power of syntactic predictors was robust to the addition of several control parameters of non-syntactic nature. For instance, the Split and the Preparation stimuli were identical from the critical events onwards, whereas participants' responses in Flash positions 3 and 4 (following the critical event) showed significant differences. Similarly, the Preparation and the Syncopation stimulus were identical up until the critical event, whereas participants' responses showed marked differences between the two stimuli in flash positions 0 and 1 – both in the immediate proximity of the critical event. Overall, the flash's distance from neighbouring events failed to provide a significant predictive advantage towards participants' responses relative to syntactic predictors alone. In principle, these observations do not rule out effects due to the global distribution of sounded events, which was different across the three stimuli. Nevertheless, such effects do not seem to be the result of local perceptual interference due to individual sounded events. In light of the present results, a more likely cause of the observed results was the integration of the entire rhythm into a unitary representation, potentially due to syntactic processing as hypothesised here.

The three stimuli adopted in the study were designed so as to suggest, as unambiguously as possible, three different interpretations while preserving the location of one specific event. This posed significant constraints on the creation of the stimuli, thus limiting the number of items to test and the generalisability of the present results. Nevertheless, the proposed syntactic framework offered parsimonious and effective predictors as well as a principled explanation to the observed behavioural responses from a large sample of participants, whereas local perceptual predictors such as the proximity to sounded events did not significantly contribute to the predictiveness of the tested models even for this limited set of stimuli. Overall, the present results represent an initial proof-of-concept for the proposed approach, calling for further confirmatory investigation to support and refine the theoretical framework proposed here. In particular, further research will need to investigate the phenomenon over a greater variety of rhythmic surfaces, including ecological listening conditions and ecological stimuli from existing repertoires, in order to consolidate these observations in a more representative setting. In light of additional empirical evidence, aspects of the present theory, and in particular the specific heuristics adopted to estimate processing costs, may be refined in the future.

7.4.4 Computational, algorithmic, and implementational analogy of language and music

Different musical dimensions, such as pitch and rhythm, are also relevant in spoken language and, more generally, processing of such auditory parameters may bear analogies with linguistic processing at the computational, algorithmic, and implementational level. In particular, rhythmic aspects of linguistic prosody share many similarities and interactions with musical rhythm (Fiveash et al., 2021; Huron and Ommen, 2006; Patel and Daniele, 2003), although musical and prosodic rhythm are distinguishable phenomena, particularly with respect to the role of metricality (Ding et al., 2017; London, 2012; Patel, 2006). This study is framed in the context of a general computational analogy between language and music, whereby the emergence of interpretation is understood as a solution to the combinatorial problem of structuring sequentially-presented events into a hierarchical network of syntactic relationships (Katz and Pesetsky, 2011; Lerdahl and Jackendoff, 1983a). The present results contribute to support this analogy at Marr's computational level by providing evidence that a grammar (e.g., Rohrmeier, 2020a), if interpreted as a model of the listeners' competence for rhythmic structure, carries predictive value towards listeners' behavioural responses (cf. Herff, Harasim, et al., 2021 for converging evidence in the context of harmony). The present results also provide preliminary empirical grounding to theoretical proposals that the experience of musical structure may be understood as the result of cognitive operations algorithmically analogous to those of a parser, as detailed by Jackendoff (1991). Concurrently with recent evidence showing the existence of retrospective revision in music, analogously to linguistic garden-path effects (Cecchetti et al., 2022), the present results support a processing architecture entailing mechanisms of online incremental parsing as well as post-hoc strategies to resolve ambiguity (cf. Steedman, 2000).

The present study for the first time formulates and tests moment-by-moment predictions on the execution of cognitive processes implementing specific operations such as integration and storage in the context of music. In particular, building on previous studies predicting processing difficulty or breakdown in music (Berent and Perfetti, 1993; Koelsch, Gunter, et al., 2000; Patel et al., 1998; see also Koelsch, 2013; Patel, 2010), the present paradigm hypothesises explicit structural relationships (encoded as grammar rules) and processing operations rather than generic syntactic violations. Furthermore, this allows the present study to investigate processing complexity in idiomatic stimuli, rather than syntactically implausible ones. Building up on this approach, future research may aim at specifying a fully detailed algorithmic implementation of the hypothesised parser, for example in the form of a left-corner parser consistent with the framework presented above (Gibson, 1991). Based on such a model, it may be possible to further investigate the boundaries of the algorithmic analogy between language and music with respect to the resolution of ambiguity (Gibson, 1998), parallel vs. serial processing (Gibson and Pearlmutter, 2000; Lewis, 2000), and the relationship of syntactic processing with emotional responses associated with surprise and tension (Jackendoff, 1991; Lehne et al., 2013; Rohrmeier, 2013).

Finally, behavioural (Fedorenko et al., 2009; Fiveash and Pammer, 2012; Slevc et al., 2009; Van de Cavey and Hartsuiker, 2016) and neuroscientific evidence (Calma-Roddin and Drury, 2020; Koelsch, 2006; Koelsch, Gunter, et al., 2000; Maess et al., 2001; Patel, 1998) suggests that linguistic and music syntactic processing share neural resources at Marr's implementational level (however, see Chen et al., 2021 for a contrasting view). The theoretical framework presented here is largely independent of whether the hypothesised computational and algorithmic features shared by linguistic and musical processing are implemented by the same neural resources. In particular, although rhythmic regularity has been shown to interfere with linguistic syntactic processing (Fiveash et al., 2023; Jung et al., 2015), it is also plausible that a domain-general rhythm-processing system would be shared between music and linguistic prosody, rather than or alongside linguistic syntax. However, the perspective proposed in this study may also contribute to the study of implementational analogies between language and music. In the absence of a precise formulation of a parsing model for music, previous research on this topic has compared possibly computationally different phenomena, such as attachment complexity and recoverable linguistic garden-path effects, on one hand, and ungrammatical musical violations (e.g., out-of-key chords, scrambled melodies), on the other. Building up on the proposed framework, future research may rather compare the neural resources recruited by specific processing operations that are algorithmically analogous across the two domains, such as structural integration or memory storage during online incremental parsing.

Overall, this study represents a proof-of-concept that predictions from a theory of musical syntax can be turned into a model of online parsing specifying the nature, the time-course and the complexity of the relevant computations, and that such predictions correlate with observable behavioural effects. From this perspective, these results contribute to frame the introspective and analytical insights drawn from the music-theoretical discourse in the context of a computational theory of cognition (Cecchetti et al., 2020; Harasim, 2020; Rohrmeier, 2013), based on the notions of syntactic competence (Chomsky, 1965) and syntactic processing as parsing (Jackendoff, 1991, 2002a).

7.5 Conclusion

Music theory predicts a rich phenomenology associated with interpreting music, i.e., attributing functional relationships to musical events. Interpretation is hypothesised to emerge as a result of cognitive processing, and to be computationally equivalent to syntactic parsing under a generative grammar. In this paper, we have presented a framework that allows to turn such music-theoretically motivated hypotheses into an empirical paradigm, useful to investigate the underlying cognitive processes by exploiting their interaction with competing cognitive tasks. Our approach extends previous research in two directions: (1) by quantifying hypotheses about specific parsing operations, as opposed to more general notions of processing complexity and syntactic violation, and (2) by investigating syntactic dependency relations in the rhythm rather than the pitch dimension of music. This approach also represents a step towards characterising structural parsing in music beyond the computational level of description, as future research may further formalise the parsing operations into a full-fledged algorithmic account. Information about the time-course and the complexity of processing computations, as characterised in the present framework, may contribute in this direction. We showed preliminary evidence that interpretations are formed during listening to musical rhythms, and that the moment-by-moment execution of cognitive processes leading to such interpretations may be modelled as the implementation of a syntactic parser relying on structural integration and memory storage. We propose this as a first step towards a theoretically and empirically grounded understanding of the emergence of musical interpretation as a cognitive phenomenon, on similar grounds as the emergence of linguistic syntactic interpretation is investigated in psycholinguistics.

8 Musical garden paths

Abstract

While theoretical and empirical insights suggest that the capacity to represent and process complex syntax is crucial in language as well as other domains, it is still unclear whether specific parsing mechanisms are also shared across domains. Focusing on the musical domain, we developed a novel behavioral paradigm to investigate whether a phenomenon of syntactic revision occurs in the processing of tonal melodies under analogous conditions as in language. We present the first proof-of-existence for syntactic revision in a set of tonally ambiguous melodies, supporting the relevance of syntactic representations and parsing with language-like characteristics in a non-linguistic domain. Furthermore, we find no evidence for a modulatory effect of musical training, suggesting that a general cognitive capacity, rather than explicit knowledge and strategies, may underlie the observed phenomenon in music.

8.1 Introduction

Syntactic parsing accounts for the computational operation of inferring representations of syntactic structure from sequential inputs (Jurafsky and Martin, 2009; Sipser, 2012). For syntactic parsing to be understood as a model of cognition beyond the pure *computational* level of description (Marr, 1982) it is necessary to account for how processing is implemented at the *algorithmic* level through cognitive parsing strategies that cope with ambiguity, limited memory resources, and with the temporal unfolding of parsing itself (Narayanan and Jurafsky, 1998; Vogelzang et al., 2017). In particular, the revision mechanisms that deal with ambiguity and temporarily misled syntactic interpretations are thoroughly investigated in psycholinguistics (J. Fodor and Ferreira, 1998; Kaan and Swaab, 2003).

This Chapter is published as: Cecchetti, G., Herff, S. A., & Rohrmeier, M. A. (2022). Musical Garden Paths: Evidence for Syntactic Revision Beyond the Linguistic Domain. *Cognitive Science*, **46**(7). GC conceptualised the study, implemented the experiment, analysed the data, and wrote the original manuscript with supervision by SAH and MAR.

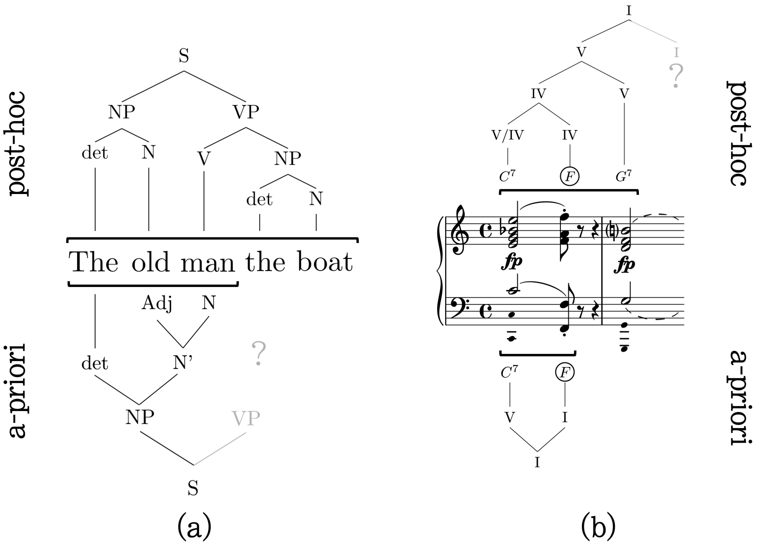


Figure 8.1 – Garden-path effect and retrospective revision in language and music. **(a)** A-priori (bottom) and post-hoc (top) interpretations of a garden-path sentence. **(b)** A similar phenomenon is predicted to occur in music (Rohrmeier, 2013), as exemplified here with the changing syntactic interpretation hypothesized to occur in the opening of Beethoven's First symphony before (a-priori, bottom) and after (post-hoc, top) the presentation of the third chord. Musical syntactic interpretations are represented as syntactic trees according to Rohrmeier and Neuwirth (2015), and question marks indicate open dependency relations, entailing expectations of future events.

Syntactic organization has also been argued to govern non-linguistic stimuli such as music (Fitch et al., 2005; Fitch and Martins, 2014; Jackendoff, 2007; Lerdahl and Jackendoff, 1983a; Patel, 2010), but it has not been empirically investigated whether effects analogous to garden-path effects occur in music and whether the parsing strategies involved in the processing of such structures would resemble those observed in language. In addressing this issue, this paper presents explicit perceptual evidence for a revision effect to occur in the processing of tonal melodies.

8.1.1 Syntactic revision in language

Structural representations emerge incrementally as a sentence is gradually presented and parsed (Frazier, 1987; Marslen-Wilson, 1973). When processing ambiguous sequences, the representation of structure as perceived may be updated retrospectively upon encountering new information, as prototypically exemplified by the recovery from so-called *garden-path* effects (Figure 8.1a; Ferreira and Henderson, 1991; Frazier, 1978). While reading the first part of the garden-path-sentence in Figure 8.1a ("The old man..."), the most likely interpretation is to understand "man" as a noun and to expect a Verb Phrase to follow (a-priori interpretation, bottom). After exposure to the second part of the sentence ("... the boat"), the previously most likely interpretation is replaced by a different one where "man" serves as a verb (post-hoc interpretation, top). Note how parsing "[...] the boat" serves here as a critical event that requires the most likely syntactic role of the word “man” to change retrospectively from noun to verb, although by this time “man” lies in the past. This change is retrospective because the interpretation that is most likely after the critical event may differ from the interpretation that is most likely before the critical event not only in terms of how it accounts for the critical event and those that follow, but also in terms of how it accounts for the events that precede the critical event. The occurrence of such a retrospective change of interpretation is associated with cognitively demanding recovery processes that manifest themselves, e.g., in slower reading times (Frazier and Rayner, 1982; Meseguer et al., 2002) and characteristic patterns of brain activity (Meltzer and Braun, 2011) following the critical event itself.

Theoretical and empirical literature in linguistics presents diverging accounts of which processing mechanisms underlie sentence processing and, specifically, the garden-path effect and recovery from it (Sprouse and Lau, 2013). In particular, in cases of syntactic ambiguity, it is debated whether only one syntactic representation is parsed at any given time (serial parsing) or rather several alternatives among the plausible ones are parsed simultaneously (parallel parsing). From a serial-processing perspective, behavioural and ERP evidence is interpreted as suggesting that separate early and late processes are involved in sentence comprehension (Friederici, 1995; Friederici and Mecklinger, 1996): the former are argued to implement a first parsing attempt that rapidly assigns a structural interpretation to the incoming information, while the latter implement any adjustments to the outcome of the early processing (e.g., through reanalysis) if incompatible information (e.g., the critical event in a garden-path sentence) is presented (Meltzer and Braun, 2011). However, alternative interpretations of ERPs (Hagoort, 2003) alongside conflicting behavioural evidence (Hickok, 1993; Nicol and Pickering, 1993; Trueswell et al., 1994) rather supports a ranked parallel perspective whereby multiple coexisting interpretations are continuously ranked and eventually pruned based, e.g., on lexical (MacDonald et al., 1994; McRae et al., 1998; Trueswell et al., 1994) and complexity (Gibson, 1991) constraints. Serial and parallel models are not easy to disambiguate, as in many cases they make broadly compatible predictions (Gibson and Pearlmutter, 2000; Lewis, 2000). In particular, while garden-path effects have a natural explanation within a serial-processing perspective (Frazier and Rayner, 1982; Friederici, 1995), parallel-processing models can also offer alternative explanations for the same effects (Gibson and Pearlmutter, 2000): as a consequence, observing a retrospective change of the most likely syntactic interpretation as perceived before and after the critical event does not rule out either family of accounts, although the proposed underlying mechanism would be different. Specifically, in serial-processing accounts such retrospective change results from the need to generate a new representation of the stimulus once encountering the critical event *re-analysis*; (Frazier and Rayner, 1982). Differently, in a parallel perspective, it is the likelihood and ranking of alternative coexisting parses that is updated (*re-ranking*; Gibson and Pearlmutter, 2000). In either case, it is possible to define a “preferred” interpretation as the only (in a serial account) or top-ranked (in a parallel account) representation that is generated by the processor at any given time. The phenomenological effect exemplified in Figure 8.1a, the switch from one preferred interpretation to a different one, is then characterized by the following qualitative conditions:

(1) One of the plausible interpretations of an ambiguous stimulus, the a-priori interpretation, is initially preferred;

(2) A critical event occurs that is unlikely under the (currently preferred) a-priori interpretation;

(3) If the critical event is consistent with an alternative structural interpretation of the entire stimulus, including the critical event and those preceding it, this new post-hoc interpretation becomes the preferred one.

Although the term *revision* is often employed equivalently to *reanalysis*, for conciseness we use here the term (*syntactic) revision* to refer to a phenomenon occurring under conditions (1)-(3), which is a plausible phenomenological manifestation of either a putative *reanalysis* process as well as of a putative *reranking* process in terms of a change of preferred interpretation (depending on the assumed underlying serial or parallel model). In this study, we seek to demonstrate the existence of this phenomenological effect of syntactic revision in a non-linguistic domain such as music.

8.1.2 Syntax and syntactic revision in music

Complex hierarchical structure has been theorized in the musical domain (Lerdahl and Jackendoff, 1983a; Rohrmeier, 2011, 2020b; Schenker, 1935; Steedman, 1984), where syntax captures idiom-specific recursive dependency relationships linking musical events which in turn motivate corresponding patterns of creation and resolution of expectancy (Cecchetti et al., 2020; Rohrmeier, 2013). Some degree of formal analogy between such syntactic structures in music and those in language has been highlighted repeatedly in the literature (e.g., Baroni et al., 1983; Bernstein, 1976; Jackendoff, 2009; Katz and Pesetsky, 2011). It has also been proposed that the listeners' experience of abstract musical structure is the result of a parsing process based on generative rules (Berent and Perfetti, 1993; Jackendoff, 1991; Rohrmeier and Pearce, 2018), and that common neural and cognitive resources are involved in linguistic and musical syntactic processing (Koelsch, 2013; Patel, 2010). The properties of such a putative musical syntactic processor have been discussed on theoretical grounds. Jackendoff (1991), arguing in favour of a parallel processing architecture, predicted a musical "retrospective reanalysis" effect based on a "selection function" singling out one preferred parse in the presence of ambiguity:

> The processor is computing multiple analyses in parallel, and [(1)] evidence has accumulated for one of these to be chosen as most plausible by the selection function. However, [(2)] subsequent events in the musical surface lead to a relative reweighting of the analyses being computed by the processor. The selection

> function thereby [(3)] “changes horses in midstream” jumping to a different analysis. The phenomenological effect of such an occurrence will be a *“retrospective reanalysis”* of the passage as it is heard. (p. 223)

As highlighted by the numbering added in brackets, this prediction is fully compatible with conditions (1)-(3). Nevertheless, experimental evidence for the very existence of such a garden-path effect in music has yet to be found.

Musical garden paths are frequently presented as a compositional device in analytical accounts of Western tonal music (Caplin, 1998; Lewin, 1986; N. J. Martin and Vande Moortele, 2014; Rohrmeier, 2013; Schmalfeldt, 2017; Temperley, 2001a). A common example is displayed in Figure 8.1b. While listening to the opening bars of Beethoven’s Symphony op. 21, the most likely a-priori interpretation for the first two chords (bottom) is replaced by a new post-hoc one when the third chord is encountered (top). Note how the F chord (circled), initially likely heard as a tonic (I in the key of F major), may be reinterpreted as a subdominant (IV in the key of C major) when the third chord intervenes.

However, unlike the effect of updating expectations over future events (Pearce and Wiggins, 2012; Sears et al., 2020), the perceptual and cognitive nature of revision of musical structure has not received much empirical attention. Additionally, it is even unclear whether revision should exist at all in music: while the success of the parsing process in language is subject to the evolutionary pressure of effectively formulating (Friederici et al., 2017) and communicating (Pinker and Jackendoff, 2005) propositional content, the nature of musical communicative interactions may not require arbitrary specificity and deterministic agreement among interactants (Cross, 2009; Fitch, 2006; Jackendoff, 2009). If reaching an unambiguous and definitive parse would not be crucial in music as it is in language, especially in the absence of formal musical training, it is not granted that processing musical structure would rely on spontaneous strategies to repair failed parsing attempts, requiring the formation and maintenance of structural representations that may be subject to a retrospective update. As a consequence, even if music theory predicts the existence of musical syntactic revision, such a phenomenon may occur spontaneously with lesser frequency or even not occur at all during music listening.

Empirical approaches to syntactic processing in music have identified musical counterparts of neural markers (Patel et al., 1998) that are known to be associated with ambiguity (Frisch et al., 2002) and second-pass (re)analysis (Friederici and Mecklinger, 1996; Kaan and Swaab, 2003; Osterhout et al., 1994) in the linguistic domain. Behavioural interference between the linguistic garden-path effect and generic musical syntactic violations (Slevc et al., 2009) has also been demonstrated, but the unambiguous nature of the musical stimuli (as opposed to the linguistic ones) does not afford the inference that the competing cognitive processes were analogous at the computational and algorithmic level. Overall, such evidence is consistent with analogous or concurrent processing between musical stimuli and linguistic garden-path sentences, possibly relying on cognitive-control resources shared across domains (Ogg et al., 2019; Slevc and Okada, 2015). However, cross-domain processing interference alone does not prove the substitution of a previously active representation with a different one: showing the existence of a phenomenon with this feature would be necessary to identify revision. In other words, evidence from cross-domain resource sharing shows that some aspect of the implementation of processing is shared, not necessarily that the same revision processes (as characterized in (1)-(3) above) are performed. Furthermore, despite the abundance of theoretical examples of musical ambiguity (Jackendoff, 1991; Rohrmeier, 2013; Slevc and Okada, 2015), no empirical studies have directly addressed this phenomenon by adopting revisable musical stimuli in a controlled experimental setting, while previous attempts to specifically contrast linguistic and musical syntactic revision with harmonically ambiguous stimuli comparable to garden-path sentences have led to inconclusive results (Ross, 2014).

8.1.3 Aims and hypotheses

Compared to the linguistic case, establishing a phenomenology of musical processing is hampered by the methodological difficulties of capturing perceptual correlates of syntactic representations in music. In language, the availability of specific syntactic representations may be tested through explicit verbalization or semantic matching (e.g., matching sentences with visual representations of their meaning; Meltzer and Braun, 2011), which is not straightforward to achieve in music. In particular, while most speakers can explicitly report their interpretation of a sentence, it is not to be expected that music listeners, especially untrained listeners, would be able to do the same. To address this issue, the present paradigm was designed to prompt behavioral responses that can be read as proxies of syntactic interpretations, even in the absence of semantic references. By accessing listener's syntactic interpretation of ambiguous tonal melodies, we aim at testing whether such interpretations were revised from an a-priori to a different post-hoc one as a consequence of a disambiguating critical event perceived as unlikely (hence surprising) under the a-priori interpretation. Overall, in analogy to linguistic syntactic revision, we hypothesize that a phenomenon unfolding as outlined in (1)-(3) above and exemplified in Figure 8.1 occurs in the processing of ambiguous tonal melodies upon presentation of a disambiguating critical event. We further assess whether such an effect is based on a general cognitive capacity or rather explicit domain-knowledge by considering the impact of formal musical training.

8.2 Methods

8.2.1 Participants

Sixty-two participants (median age 25.5, range 18-74) took part in an online study (ethics approval number HREC 037-2020). The sample represents a wide range of musical expertise, as reflected by a median score 30 (range 7-45) in the Musical Training subscale of the GoldMSI (Müllensiefen et al., 2014). For comparison, Müllensiefen et al. report a mean score of 26.52 ($SD = 11$) for a large validation sample of Western listeners. All participants reported

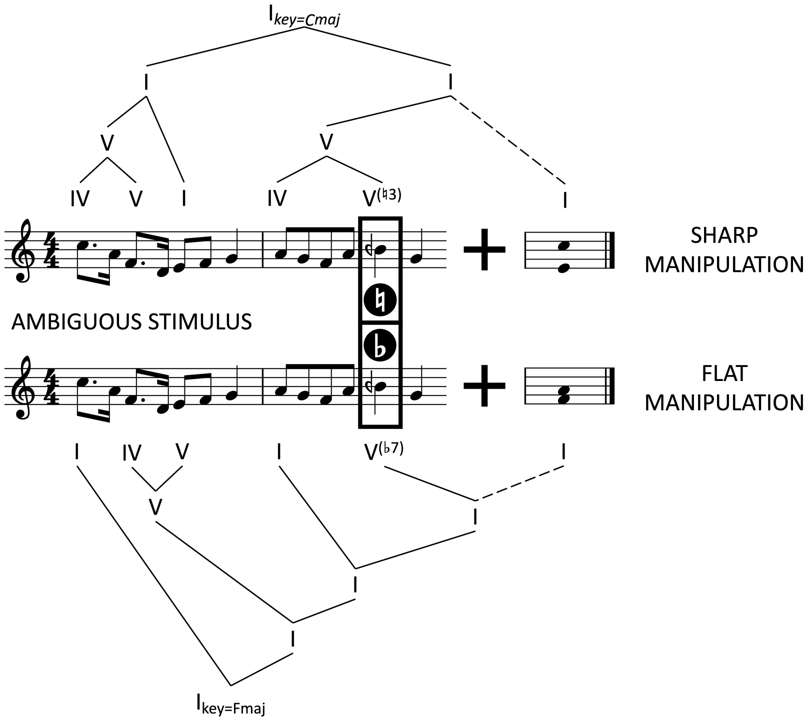


Figure 8.2 – A tonally ambiguous stimulus, with a key-defining note mistuned by a quarter tone (in the box). The presentation of a sharp or flat manipulation, in the form of a two-voiced chord (root and third of C major or F major, respectively), at the end of the melody may bias the interpretation of the stimulus towards the corresponding key. We also highlight two different tree analyses after Rohrmeier and Neuwirth (2015), exemplifying the two interpretations. The B half-flat may be heard either (top) as the leading-tone in a dominant (V) chord in C, to be tuned upwards as a B natural, or rather (bottom) as the seventh of a dominant chord in F, to be tuned downwards as a B flat. Since the note is tuned halfway between B and B flat, both interpretations are plausible until the chord is presented.

close familiarity with at least one genre within Western musical practices (e.g., classical, Jazz, Rock/Pop).

8.2.2 Stimuli

Fifteen distinct original melodies, collectively ranging from C4 to G5 with 440Hz tuning and each spanning 2 bars in 4/4 meter at 120bpm, were synthesized in MuseScore 3.5.0 in the default piano timbre. In their original transposition, melodies were composed with the goal of being interpretable in the key of C-major in the absence of any accidentals, while by flattening pitch B to B flat they can be interpreted in the key of F-major. Specifically, each melody affords to be harmonized with idiomatic chord progressions in either one of the two keys, given the appropriate accidentals. Ambiguous stimuli were then obtained from each melody by mistuning all occurrences of pitch B by a quarter-tone, halfway between B and B flat, as highlighted by the box in Figure 8.2. These 15 stimuli, each comprising some mistuned notes, are used in the main experimental task. All stimuli are available as Supplementary Material S1.

Importantly, modes and keys are not only sets of notes (e.g., scales), but come with specific typical melodic and harmonic motions that determine (functional) relationships between notes (Bostwick et al., 2018; Large et al., 2016; Lerdahl, 2001). Hearing a melody in a key results in attributing interpretations to each note, specifying their relationships to all other notes (cf. Schenker, 1935). These key-specific relationships may be expected to be updated when a melodic excerpt is suddenly perceived in a different key. As an example, Figure 8.2 reports two alternative tree analyses for one of the melodies based on the established generative grammar for tonal harmony proposed by Rohrmeier and Neuwirth (2015). Specifically, the formalism models hierarchical harmonic structure in terms of an Abstract Context Free Grammar (Harasim et al., 2018) based on two rule types: preparation, $X \rightarrow YX$, and prolongation, $X \rightarrow XX$, where X and Y stand for chord symbols (expressed, e.g., as Roman numerals). The two tree analyses capture the syntactic constituency structure that a listener may perceive when interpreting the melody in C major or F major, respectively. The tuning of note B as a B natural or B flat disambiguates between the two alternatives: a B natural may be interpreted as the third of a V ("leading tone") in C major, which then prepares a C major chord; a B flat may be interpreted in F major as the seventh of a V, which then prepares an F major chord, or as the third of a ii, which then prepares a C major chord as the V of F.

In this framework, revision occurs when a listener's preferred parse for the melody, captured by one of the two tree structures, is made implausible by the occurrence of, e.g., a key-defining chord at the end of the melody, and eventually the listener's preferred parse for the entire stimulus including the chord is best represented by the other tree structure, consistently with conditions (1)-(3). Crucially, since the C major scale contains no B flat and the F major scale contains no B natural, a listener hearing the quarter-tone note B half flat as a $\widehat{7}$ in the key of C major might find more appropriate to replace the quarter-tone note with an equal-tempered B natural, while a listener hearing the quarter-tone note as a $\widehat{4}$ in the key of F major may find more appropriate to replace it with a B flat. In other words, the preferred equal-tempered approximate tuning of the quarter-tone note can be used as a proxy to infer the listener's syntactic interpretation.

Note that, in principle, listeners may have perceptual biases that deviate from equal-tempered tuning, thus making quarter-tones only an approximation of the perceptual midpoint between scale tones. In order to mitigate this potential source of variability, we make sure listeners are primed to equal temperament by presenting equal-tempered melodies throughout the experiment, and we further account for systematic individual biases across participants by allowing for the corresponding random effect in our analyses.

8.2.3 Experimental task

In the presentation phase of each trial, a stimulus was played randomly in one of the 12 chromatic transpositions. Over the 30 trials of the experimental task, each stimulus was presented twice in random order with different manipulations. No proximity constraints were imposed by design, but the effect of proximity between presentations of the same stimulus was explicitly investigated in the analysis. In each trial, the manipulation consisted in the presentation of a two-voiced chord that removed the key ambiguity: in the original transposition, a C-major chord constituted the *sharp* manipulation (cf. the top staff in Figure 8.2, where the manipulation is displayed as the chord in the last bar) and a F-major chord constituted the *flat* manipulation (cf. the bottom staff in Figure 8.2). Following the manipulation, two behavioral variables were measured:

Surprise Rating. Participants were asked to rate how surprising the final chord sounded to them. Ratings were provided on a quasi-continuous Visual Analog Scale (Hayes and Patterson, 1921) ranging from *Expected* to *Surprising.*

Tuning Response. Participants were then presented with the stimulus again, additionally transposed by an ascending or descending tritone and preceded by 3s of white noise to minimize proactive interference from the manipulation at the end of the first presentation towards the second presentation. This time, the stimulus was interrupted right before the first occurrence of the de-tuned pitch. Participants were then instructed to select a note to continue the melody in the way that most closely resembled how they remembered the melody itself from the presentation phase of that trial. Two options were given, corresponding to the *sharp* (in the original transposition, B) and *flat* (B flat) tuning of the mistuned note respectively. Each option was associated randomly with a key on the participants' keyboard (Q or P), and participants could play either option arbitrarily until they were ready to confirm their response with another key press.

In order to address the hypothesis that a phenomenon occurring under conditions (1)-(3) (cf. Introduction) plays a role in the processing of the ambiguous stimuli, we need to be able to access and compare the a-priori and the post-hoc interpretations of each melody. This is achieved through the two behavioral measures. In each trial, the Surprise Rating carries information concerning the a-priori interpretation of the melody. A low Surprise Rating suggests that a strong a-priori interpretation of the stimulus was formed which happened to be the same as the one implied by the manipulation (e.g., C-major for the *sharp* manipulation in the original transposition), whereas a high Surprise Rating suggests that the a-priori interpretation was different to the one implied by the manipulation. Specifically, a high Surprise Rating indicates that the manipulation was perceived as having a low likelihood conditional to the a-priori interpretation up to that point in the melody (Pearce and Wiggins, 2012), thus potentially serving as a critical event consistently with condition (2). In turn, the selection of one tuning over the other in the Tuning Response is a proxy of the post-hoc interpretation, as it captures a representation of the melody that participants accessed after the manipulation had been presented.

Since each melody was presented with both manipulations, we can compare the corresponding Surprise Ratings. A small difference between the two Surprise Ratings suggests either that no strong a-priori interpretation was formed in either trial, so that both manipulations resulted in only average surprise, or that the a-priori interpretation had changed across the two trials, resulting in the two different manipulations being perceived as similarly surprising. None of these scenarios matches both conditions (1) and (2) in our working definition of revision: condition (1) requires an a-priori interpretation to be preferred for the melody prior to, and independently of, the manipulation, so that only one of the two manipulations is perceived as an unlikely critical event inducing the need for revision to occur as per condition (2). On the contrary, a large difference between the two Surprise Ratings indicates that an interpretation is strongly preferred for the melody in both trials, and that in the more surprising trial this interpretation is at odds with the manipulation, so that in such a trial both conditions (1) and (2) are satisfied. If revision occurs in this trial, we further expect the post-hoc interpretation, as captured by the Tuning Response, to be updated in accordance with the manipulation, so that condition (3) is also satisfied. In other words, if a melody undergoes revision, we expect its two Tuning Responses to be different from each other and in music-theoretical agreement with the manipulation in the respective trials. While it is possible that this latter scenario may also occur alongside a small difference in Surprise Ratings, we conservatively only interpret a systematic co-occurrence of this circumstance with a large difference in Surprise Ratings as evidence for the occurrence of revision as defined at the outset in terms of conditions (1)-(3).

8.2.4 General procedure

After providing informed consent, participants performed a memory task based on a continuous-recognition paradigm (Herff, Olsen, and Dean, 2018; Herff, Olsen, Dean, and Prince, 2018; Shepard and Teghtsoonian, 1961), where the 15 ambiguous stimuli were presented twice in random order and transposition. Results from the memory task are reported in (Cecchetti et al., 2021). During the memory task, participants had the chance to familiarize themselves with the stimuli, potentially (but not necessarily) settling on some preferred parsing that could eventually be prone to be revised. Whether this happened or not does not impact the interpretation of the results reported in this study. Following the memory task, participants took part in the main revision experiment described above and finally answered the Goldsmith Music Sophistication Index (MSI) questionnaire (Müllensiefen et al., 2014). The entire experimental session lasted 45-60 minutes in total.

8.2.5 Analysis

For each melody and participant, a new variable is defined, *Congruency*, with the three categories *Congruent*, *Incongruent* and *Stable*. A melody falls in the *Congruent* category if the Tuning Responses to both presentations of the melody are consistent with the manipulation adopted in the corresponding trial (e.g., *sharp* Tuning Response in a trial where the *sharp* manipulation was presented, and *vice versa*). A melody falls in the *Incongruent* category if the opposite happens in both trials, while it falls in the *Stable* category if the Tuning Response is the same irrespective of the manipulation. By definition, the occurrence of revision for a given melody as music-theoretically predicted would place that melody into the *Congruent* category. In addressing our main hypothesis that syntactic revision occurs in music, we show then that *Congruent* Tuning Responses are more likely as the difference in Surprise Rating (*DiffSurprise*) between the two presentations of the same melody increases.

Data were analyzed with Bayesian mixed-effects models provided with weakly informative priors ($t(3{,}0{,}1)$; Gelman et al., 2008) and implemented with the R package *brms* (Bürkner, 2018). All non-categorical variables were standardized to null mean and unit standard deviation, and we then fitted the model:

$$\begin{aligned}\text{Congruency} \sim &\text{DiffSurprise} + \text{MusicalTraining}+\\ &\text{DiffSurprise} \times \text{MusicalTraining} + \text{InterveningTrials}+\\ &(1|\text{Participant}) + (1|\text{Stimulus}),\end{aligned}$$

where $DiffSurprise = MaxSurprise - MinSurprise$ and *MusicalTraining* is quantified by the corresponding subscale of the MSI. We also tested for an effect of the number of intervening trials (*InterveningTrials*) separating the two presentations of the same melody, to account for potential interference effects (Herff and Czernochowski, 2019; Herff, Olsen, and Dean, 2018). Specifically, it is possible that the two trials of the same stimulus interfere more strongly with one another, the closer they are together. The model also allows for random effects accounting for the individual variability across participants and for the differences across the 15 stimuli. We report coefficient estimates (β), their estimated error (*EE*), as well as evidence ratios (*Odds*) for the individual hypotheses (a given coefficient being larger or smaller than zero), labelled as 'significant' (*) at a .05 confidence level when exceeding 19. Data and code are available as Supplementary Materials S2 and S3.

8.3 Results

Stable responses (56.67% of the total 930 observations) were most likely in general ($p < .001$ in a one-sided binomial test). In other words, participants tended to form a definite *a-priori* interpretation of each stimulus that was usually retained irrespectively of the manipulation, even if the latter was found in conflict with the interpretation itself. However, in the cases in which changes in the post-hoc interpretation occurred, participants responses were not distributed randomly between *Congruent* and *Incongruent* responses. Instead, we found strong evidence that *DiffSurprise* ($\beta = .16$, $EE = .07$, $\text{Odds}\left(\beta > 0\right) = 188.47^*$) carries predictive power towards the *Congruency* of Tuning Responses. Specifically, as shown in Figure 8.3 by the upward trend in the blue line, the predicted probability of observing *Congruent* responses for a stimulus increases with the difference between the two surprise ratings, and significantly exceeds the likelihood of observing *Incongruent* responses for $DiffSurprise > 0.91$ (Odds$\left(0.91 \cdot DiffSurprise > (Intercept_1 + Intercept_2)/2\right) = 19.01^*$; see S4 for details). As discussed in Section 8.2.3, this suggests that a process satisfying conditions (1)-(3) outlined in

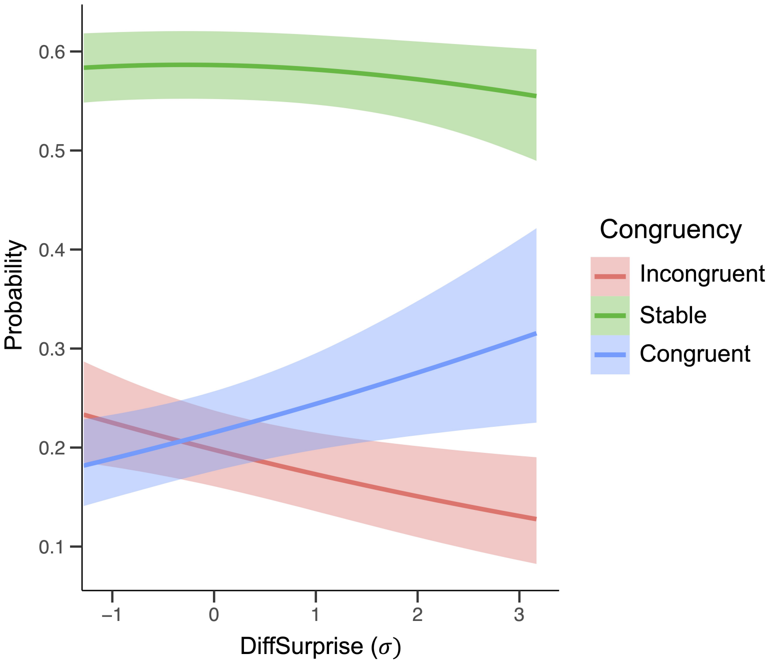


Figure 8.3 – Predicted probability for participants' post-hoc interpretations of a melody to be in accordance with the manipulation in both (Congruent category, blue), neither (Incongruent category, red) or only one (Stable category, green) of the two presentations of the same melody, expressed as a function of the difference in Surprise Rating (DiffSurprise, expressed in standard-deviation units from the mean) between the two presentations of the melody. The prevalence of Congruent over Incongruent responses for increasing values of DiffSurprise supports the hypothesis that syntactic revision occurs in the processing of the musical stimuli. Stable responses are most likely in general.

the Introduction is observed, and thus supports the occurrence of syntactic revision in music.

We only observed weak to no evidence that this effect was further shaped by *MusicalTraining* (β = .11, EE = .08, $Odds(\beta > 0) = 12.28$), its interaction with *DiffSurprise* (β = -.04, EE = .06, $Odds(\beta > 0) = 0.36$), or *InterveningTrials* (β = -.04, EE = .06, $Odds(\beta < 0) = 2.40$).

8.4 Discussion

In this study, a novel behavioral paradigm was developed to test for the existence of a perceptual correlate of syntactic revision in the domain of music. Data support that the retrograde integration of new information, as provided by the manipulation presented at the end of a stimulus, results in updated syntactic representations, as captured by Surprise and Tuning Responses. Specifically, a bias in favor of revising the preferred tonal interpretation of the ambiguous stimuli in accordance with the manipulation, rather than randomly, was observed selectively in conditions that match psycholinguistic accounts of garden-path-sentence processing (e.g., Frazier, 1978).

We characterized syntactic revision as a phenomenon whereby (1) listeners form a preferred a-priori interpretation of an ambiguous stimulus, (2) a critical event occurs that is unlikely under the a-priori interpretation, and (3) a change occurs from the a-priori interpretation, which is preferred prior to the critical event, to a different post-hoc interpretation consistent with the critical event. As a consequence, we hypothesized that, if a phenomenon of musical revision exists, the occurrence of conditions (1) and (2) would increase the likelihood of condition (3).

Our stimuli comprised ambiguous melodies allowing participants to form two plausible interpretations. At the end of each melody, a disambiguating final chord was presented and a Surprise Rating was measured. Two different chords were used, corresponding to the two different interpretations. Given a melody, the observation that one chord was perceived as expected and the other as surprising indicates that listeners had formed the same a-priori interpretation of the melody in both cases, and that under this a-priori interpretation the surprising chord was perceived as unlikely, i.e., as a critical event. A large difference in surprise for a given melody indicates then that conditions (1) and (2) were fulfilled in the more surprising trial.

Following the disambiguating chord, we then asked participants to report their memory of the melody as a proxy of the post-hoc interpretation. We observed that, as the difference in surprise increased, participants were not only more likely to report a post-hoc interpretation consistent with the chord in trials where the chord was perceived as highly expected, but crucially also in trials where the chord was perceived as highly surprising. This supports that conditions (1)-(2), captured by the difference in surprise, were predictive towards (3), captured by the reported post-hoc interpretation – which is consistent with the occurrence of musical revision.

We also observed that participants tended to form a strong and stable *a-priori* interpretation of the melodies that was maintained regardless of the number of intervening trials separating the two presentations, so that a surprising manipulation did not deterministically result in participants reporting a revised post-hoc interpretation. Note that, in the linguistic domain, comprehenders also often fail to recover from garden-path effects, rather sticking to some form of good-enough parsing (Ferreira et al., 2001). In the musical domain, this behaviour may be even more typical, consistently with the understanding that musical syntax serves a different purpose than linguistic syntax (Lerdahl and Jackendoff, 1983a; Rohrmeier, 2020b). In particular, if ungrammaticality is penalized to a minor extent by communicative pressure (cf. Temperley, 2004) in the musical domain, parsing strategies that deal with recovery from temporary failure may not be systematically adopted. Listeners may then ignore conflicting information, or fail to integrate it into a unique coherent parsing together with previous events.

Since a surprising manipulation did only probabilistically, rather than deterministically, lead to revision, one may wonder how often revision occurs when processing stimuli that may allow for it, as those adopted in this study. In this respect, our results should not be taken as a characterization of the frequency of occurrence of revision: as we conservatively focus on a sufficient condition for conditions (1)-(3) to be met, such frequency may be underrepresented by our analysis. However, this exceeds the scope of this proof-of-concept study, which has the goal to establish whether revision occurs at all in musical stimuli.

Characterizing what other conditions influence the occurrence of revision in music is another relevant issue open for further research. Musical training is a natural candidate in this respect. However, while all participants were generally familiar with Western music, no significant effect of explicit musical training on the *Congruency* of *Tuning Responses* was found, nor did musical training modulate the strength of the observed effect of *DiffSurprise*. Based on these results, the occurrence of musical syntactic revision does not seem to be explained as a byproduct of explicit domain-specific knowledge, as is acquired through formal training, but likely as the manifestation of a fundamental cognitive operation relying on an implicit syntactic competence, in broad formal analogy to the linguistic one.

It should also be noted that, in each trial of our paradigm, the Tuning Response was determined by parsing a second presentation of the melody. This second parse, in principle, may have been independent of memory of the first (Jackendoff, 1991). The observed systematic dependency of Surprise Ratings and Tuning Responses, however, supports that the two parses were not entirely independent of one another. In other words, the paradigm could be conceived of as inducing a priming effect of the first presentation on the parsing of the second one. Note that the final chord of the first presentation (i.e., the manipulation) is unlikely to have primed the Tuning Response on its own, since the second presentation was transposed. As a consequence, it is a representation integrating both the ambiguous melody and the manipulation that may have primed the Tuning Response in the second presentation, and we showed evidence that this representation may have been revised retrospectively in some trials at least. In principle, the observed bias on the Tuning Response in highly surprising trials may have originated at any point before the participant's Tuning Response, not necessarily while still actively parsing the first presentation of the stimulus. As a consequence, how closely the time course of the phenomenon identified in this study matches the time course of syntactic reanalysis processes in language (Friederici and Mecklinger, 1996; Meltzer and Braun, 2011) remains to be investigated.

More generally, empirical research is still far from characterizing the processing of musical structure in comparable detail as linguistic processing. In language, the existence of a garden-path effect and related phenomena in readers/listeners constrains the characteristics of any plausible model of linguistic processing (cf., e.g., Lewis, 2000), and based on these phenomenological observations it is then possible to debate, e.g., whether a model of parsing should separate syntactic from semantic processing and which one has functional priority (Frazier, 1978; Hagoort, 2003; Sprouse and Lau, 2013), whether processing is serial or parallel (Lewis, 2000), or what factors determine the preferred choices of the human parser (Gibson and Pearlmutter, 1998). A precondition for addressing this type of questions in music is to observe a solid base of phenomena that constrain the properties of a putative musical syntactic processor. In particular, the very existence of an effect of musical syntactic revision has been a crucial yet unsupported assumption of theoretical accounts of musical processing

(Jackendoff, 1991; Rohrmeier, 2013; Slevc and Okada, 2015), which now finds preliminary empirical grounding.

In principle, the phenomenon reported here is strictly musical in nature, and it exceeds the scope of the present study to specify the analogy with its linguistic counterpart beyond the broad characterization in terms of conditions (1)-(3) our approach assumed. In particular, our results cannot disambiguate between serial or parallel accounts of syntactic processing, as our assumptions are broad enough to be consistent with both. Further research will also need to clarify whether existing evidence for shared neural and cognitive mechanisms (Fedorenko et al., 2009; Koelsch, 2013; Patel, 2003) accounts for the phenomenological analogy concerning musical and linguistic revision established in this study – in particular, whether such shared substrate supports the whole of the processing pipeline (from parsing proper over error detection to revision), and to what extent domain-specific and domain-generic processing modules are involved (Peretz and Coltheart, 2003; Peretz and Zatorre, 2005). In this respect, our study offers a novel empirical paradigm that succeeds in singling out the occurrence of musical revision and that may complement future studies that wish to selectively target this phenomenon.

Overall, by providing the first explicit evidence for syntactic revision of tonally-structured music by Western-enculturated listeners, our results show that syntactic revision as a cognitive mechanism is spontaneously deployed in the processing of non-linguistic stimuli. The present findings suggest that syntactic representations of music develop incrementally during listening, and representations of syntactic relationships linking events in the past appear to persist in memory. Specifically, such representations are prone to retrograde interference due to an ensuing syntactically-related critical event and yet robust to merely sensory interference from intervening unrelated trials. Although we found a prevalence of Stable responses, suggesting that revision may not occur frequently in general, the very existence of this phenomenon even in a smaller number of cases challenges the sufficiency of models that only afford the update of expectations towards future events, such as simple Markov chains or n-gram models predicting surface events, as cognitive models of musical structure (cf. Rohrmeier, 2013; Rohrmeier and Pearce, 2018). In particular, modelling a revision effect requires the existence of latent structure (in terms of, e.g., hidden states or non-terminals) encoding alternative, abstract interpretations of a given surface, such as the harmonic function of a given pitch collection and the syntactic relatedness of different harmonic functions. From this perspective, a given musical surface is ambiguous insofar as it can be generated by a multiplicity of latent-structure encodings. Our findings further suggest the necessity of a processing architecture that allows for the retrospective change of the latent structure that is interpreted as generating a portion of the musical surface that belongs in the past: updating transitional probabilities towards states generating events in the future (as in Markov or n-gram models) is not sufficient to account for the observed effect. In summary, these results support the hypothesis that syntactic models that account for abstract musical dependencies based on latent structure are required in describing the cognitive underpinnings of the musical experience, and that qualitatively analogous parsing strategies as those observed in language are likely deployed in music perception.

Coda Part IV

9 General discussion

9.1 Summary

In this thesis, we have presented empirical results from several behavioural studies addressing the cognitive reality of structural representations in music, as well as of some processing mechanisms underlying their emergence. By "structural representations", we mean here encodings of the structural relations linking events in the musical surface with one another, as discussed in Chapter 1. In particular, results in Chapter 4 suggest that, when we retrieve a melody from memory, we do not only have access to its sensory make-up but also to the tonal relationships between tones: in other words, listeners form representations of structure that are stored in memory alongside representations of sensory information. Building up on these results, the study presented in Chapter 5 adopts a structural priming paradigm to show that a prime stimulus can influence the perception of a later target stimulus depending on the congruency of their structural interpretations. Consistently with how this phenomenon is interpreted in language (Branigan and Pickering, 2017), these results further support that key-independent structural representations are encoded in memory in abstraction from sensory information, so that the activation of one such representation due to the prime stimulus can influence the activation of the same or a different representation for the target stimulus. Importantly, the priming effect occurs even if representations are task-irrelevant, thus supporting their emergence as a result of automatic, implicit processing. Furthermore, the observed priming effect reflects information about structure at different hierarchical levels, supporting the hierarchicity of the underlying structures.

Having established the cognitive relevance of a notion of structural representation, Chapter 6 investigated one component of structural representations for a specific musical idiom, the harmony of extended tonality. Results indicate that a music-theoretically motivated categorisation of harmonic functions is predictive of the similarity of listener's expectations towards future events. In particular, a clear distinction is observed between expectancy-inducing harmonies, falling in the subdominant and the dominant functional categories, and non-expectancy-inducing harmonies, falling in the tonic functional category. This observation partially supports the external explanatory value of music-theoretical accounts of extended tonality in terms of functional substitutability. However, these results also likely reflect the contemporary listener's implicit familiarity with Pop repertoires, where authentic and plagal motions tend to have comparable status.

We then addressed some aspects of an algorithmic characterisation of processing. First, in Chapter 7, we proposed a framework for testing hypotheses about the time-course of incremental parsing. By applying dependency-locality principles (Gibson, 1998) to a grammar for rhythmic structure (Rohrmeier, 2020a), we formulated predictions about the cognitive load required by the moment-by-moment execution of parsing computations. These model-driven estimates of processing complexity predicted the inaccuracy in reporting the temporal location of visual flashes presented concurrently with listening to musical rhythms. The observed effect was robust to competition with control variables reflecting surface features of the music, thus supporting the role of syntactic parsing as a cognitive computation implementing the inference of structural representations. Note that a notion of interpretation in terms of the inference of structural relations among events is common in the domains of harmony and melody, but its explicit formalisation in the domain of rhythm is a recent addition in empirical research (Rohrmeier, 2020a; Sioros et al., 2018), despite being almost a common-place in the musical discourse. As a consequence, besides representing a first attempt to model the time-course of syntactic parsing in the musical domain, these results also provide preliminary support for the applicability of a notion of structural interpretation to the domain of rhythm.

As discussed in Chapter 3, a model of incremental parsing should also account for mechanisms of ambiguity resolution and recovery from misled parses. Chapter 8 reports the first evidence for a phenomenon of garden-path resolution in the musical domain, including the retrospective revision of the structural interpretation of events belonging in the past. The existence of garden-path effects is well known in language (J. Fodor and Ferreira, 1998; Frazier, 1987) and has been extensively hypothesised on theoretical grounds in the musical domain (Jackendoff, 1991; Rohrmeier, 2013; Temperley, 2001a), yet it had never been observed empirically before.

9.2 Towards a model of parsing across Marr's levels

Taken together, results from Chapters 6–8 identify some cornerstones that may inform a theory of structural parsing in music. In particular, results are compatible with the operation of a parser that forms representations of structural relations incrementally over time. The automatic and implicit operation of the parser is supported by the observation of priming effects when the emerging representations are task-irrelevant (Chapter 5), and is consistent with the automatic nature of syntax-related ERPs elicited by task-irrelevant manipulations (Koelsch et al., 2007; Koelsch et al., 2002). Such a parser would deploy computations pertaining to (at least)

Memory storage: the projection of hypothesised future nodes, corresponding to events that are implied by the current structural representation but have not been encountered yet.

Structural integration: the attachment of newly-encountered events into a pre-existing (partial) representation, whereby the newly-encountered event is attached to an earlier event (the target of the attachment) thus establishing a structural relation between them.

Results from Chapter 6 are consistent with the view that, when encountered, events can be perceived as members of syntactic categories distinguished by the potential to form structural relations with other events. In this understanding, syntactic categories correspond to different projections towards future events: for example, in the context of extended tonal harmony, results show that dominant- and subdominant-functioning chords may be understood as projecting definite expectations towards future events, whereas tonic-functioning chords do not imply hypothesised nodes and thus fail to engender consistent projections.

Memory storage and structural integration entail the recruitment of cognitive computational resources. In particular, memory storage requires the maintenance of a representation of the hypothesised nodes comprised in the projections of previously-encountered events: thus, the parser incurs in memory-storage costs over the entire time-span between the occurrence of an event that projects a hypothesised node, until the occurrence of an event that fills that hypothesised node. Processing costs pertaining to structural integration are incurred as the parser attempts to access a representation of the target of the attachment: intervening events may interfere with memory retrieval, resulting in increased processing complexity. Overall, in this view, processing complexity reflects constraints of dependency locality: structural relations that span longer period of time result in higher processing costs.

Results in Chapter 8 further constrain such a putative parser to operate, minimally, a ranking of competing interpretations in the presence of ambiguity. This is compatible with a parallel parsing architecture complemented by a *selection function* that singles out the most likely interpretation, as hypothesised by Jackendoff (1991), as well as with a serial parsing architecture that is forced to re-analyse the entire stimulus from scratch if the initial parse is misled. The findings presented here cannot disambiguate between such contrasting parsing models, and future research may build up on the present results to test more fine-grained hypotheses about the algorithmic nature of parsing.

More generally, we see the approach presented here as paving the way for bridging across the different Marr (1982) levels in characterising the emergence of representations of musical structure. Future research may test competing algorithmic theories of parsing that differ in terms of what kind of operations are performed, when they are performed, and how they contribute to processing complexity. In particular, certain crucial features of a cognitively plausible parser have not been tested in the musical domain. For example, in language, the observation that left- and right-branching structures are less demanding, in terms of processing, than center-embedded ones rules out purely bottom-up and purely top-down parsing strategies as algorithmic accounts of incremental parsing (Resnik, 1992). A cognitively plausible parser for language would rather need to incorporate both top-down and bottom-up features, such as in a left-corner parser for context-free grammars (Gibson, 1991; Resnik,

1992) or a parser for combinatory categorial grammars (Steedman, 2000). In music, center-embedding does result in increased processing complexity (Ma et al., 2022), but a direct comparison between the processing complexity of center-embedded, right-branching, and left-branching structures has not been tested. Future research may focus on investigating this issue to determine what incremental-parsing strategies are more likely to reflect the operation of the human "processor" for musical structure.

Furthermore, the parsing process outlined in Chapter 7 does not take into account many forms of structural cues that musical surfaces typically provide in terms of texture, articulation, etc. For example, in the classical style, a cadence is not only a harmonic-contrapuntal schema but is also associated with characteristic textural features (e.g., change of harmonic rhythm or motivic liquidation; Caplin, 1998) that make it easily identifiable irrespectively of harmonic syntax. Furthermore, cadences mark important formal anchor points that imply long-distance structural relations: for instance, the tonic that closes the cadence marking the Essential Structural Closure (ESC; Hepokoski and Darcy, 2011) syntactically attaches back to the tonic that typically opens a sonata-form piece. If familiar with this coupling of harmonic syntax and form, listeners may infer the non-local structural relation between the initial tonic and the ESC-tonic of a sonata movement upon identifying the ESC cadence without the need of maintaining a representation of the open dependency (a "hypothesised node" in Gibson's (1991) terminology) throughout the movement. As a consequence, instead of predicting sustained memory-storage costs over an extended period of time plus an instantaneous structural-integration cost, such an inference strategy would rather only predict a structural-integration cost at the cadence, originating from the memory-retrieval effort. It is unclear whether cue-based or schema-based inference (cf. Gjerdingen and Bourne, 2015 on a construction-grammar approach to musical structure) may fully and efficiently reflect the combinatorial complexity of music to the same extent as incremental parsing, but two such forms of inference may rather coexist to account for a flexible capacity to process novelty as well as parsimonious heuristics for dealing with prototypical structures.

Crucially, any such theories will need to be compatible with a computational-level understanding of the mapping from surface to structural interpretation: in other words, in this view, theories of parsing are constrained by the "shape" of possible structural interpretations as internally-explanatory derivations of the musical surface under some generative model. In this way, music-theoretical insights can be faithfully translated into computational-level theories (Lerdahl and Jackendoff, 1983a; Rohrmeier, 2020b), for which algorithmic strategies with concrete cognitive implications may be hypothesised and tested.

Available behavioural evidence should also be complemented with brain-imaging studies addressing the interface of the algorithmic and the implementational level of description, by identifying the brain structures and processes that support the execution of the hypothesised computations. In particular, algorithmic-level theories of parsing may help attributing functional interpretations to brain imaging findings. Future research may, for example, investigate whether estimates of processing complexity associated with specific parsing computations –

such as the ones outlined in Chapter 7 – modulate the amplitude of ERP signatures such as the N5 on an event-by-event basis. This would support and enrich the interpretation of the N5 as reflecting the execution of structural-integration operations (Koelsch, 2011). Similarly, known markers of syntactic reanalysis in language, such as the P600 ERP (Friederici, 2002), have been proposed to reflect similar algorithmic operations in the domain of music (Koelsch et al., 2005), which my be supported by investigating the occurrence of such EEG signatures in conjunction with the phenomenon of musical garden-paths identified in Chapter 8.

9.3 Syntax, language, and Bayesian cognition

In discussing these phenomena, we have characterised the inference of structure in music in close analogy with linguistic parsing. This is particularly useful insofar as research in (psycho)linguistics offers to research in music decades of successful examples of how a computational-level understanding of the type of structures encountered in a domain can be transformed into cognitively-plausible algorithmic-level theories of processing. Appealing to such an analogy resonates with a large body of literature discussing the putative sharedness of cognitive and neural resources between the two domains (Koelsch, 2013; Patel, 2010). Nevertheless, the conceptual framework and the findings presented here are agnostic to the sharedness hypothesis, in the sense that any analogy at the computational or algorithmic level does not imply a common implementation or the recruitment of domain general resources.

On the other hand, more detailed accounts of musical processing at the algorithmic level may help to assess whether the observed sharedness of implementational resources can be interpreted as reflecting the execution of common algorithmic operations across the two domains. For example, a standard paradigm such as Slevc et al.'s (2009) cross-domain self-paced-reading task may be employed to probe the effects of word-by-word and chord-by-chord processing costs resulting from algorithmically congruent parsing computations (e.g., structural integration, memory storage, reanalysis, etc. across domains) in fully grammatical (as opposed to syntactically broken) stimuli. This would allow us to investigate whether the previously observed interference between language and music is the result of analogous algorithmic operations that implement the "ordinary" parsing process in the respective domains and that compete over shared cognitive resources, as opposed to mechanisms that deal with violations specifically.

Being adaptations of context-free grammars (Harasim et al., 2018; Rohrmeier, 2020a, 2020b), the grammar models underlying our studies share many formal similarities with analogous formal models that have been historically proposed for linguistic structure (Chomsky, 1957),[1]. These models are particularly suitable to capture nested, non-crossing structural dependencies in sequential strings of symbols forming a single stream, which covers some aspects of musical organisation in a satisfactory yet simplified way (Rohrmeier, 2020b). In the future,

[1] In general, linguistic syntax exceeds context-free complexity (Savitch et al., 1987; Steedman, 2000); the formal complexity of musical syntax is debated (Rohrmeier, 2020b; Rohrmeier and Pearce, 2018).

the expressive capacity of current models may be expanded in order to account for musical features that are problematic under the constraints of context-freeness, including (1) repetition structure (Finkensiep et al., 2023), (2) crossing dependencies (Wagner, 1995), (3) the coexistence of dependencies linking events and dependencies linking processes (Ren et al., 2023), and (4) the intricacies of multiple streams in (implicitly or explicitly) polyphonic music (Finkensiep, 2023). In all these respects, the generative models adopted in music may grow significantly apart from those adopted in modelling linguistic structure, which, for example, do not need to account for polyphony (Finkensiep, 2023).

The analogy between linguistic and musical processing in this context is based on the understanding that both language speakers and music listeners face the challenge of inferring latent structural relations when exposed to sensory stimuli in the respective domains, after learning from supervised and unsupervised exposure. Grammar-based incremental parsing is proposed as a viable algorithmic strategy to implement such inference, in accordance with probabilistic accounts of language acquisition and processing (Chater and Manning, 2006; Clark, 2017; Manning and Schütze, 1999; Sprouse and Lau, 2013). In this sense, the approach we presented here falls within the scope of the more general framework of Bayesian cognition (Chater et al., 2010; Clark, 2013b; Griffiths et al., 2008). Consistently with the Bayesian-cognition perspective, the present approach addresses the emergence of structural hearing in terms of (1) characterising the musical surface as the result of a generative process, (2) learning such a candidate generative model through exposure to a repertoire, and (3) inferring (representations of) the latent causes of observed stimuli by inverting the hypothesised generative process. Crucially, this approach does not commit to the ontological reality of such latent causes: entities and relations may well be fully "illusory", as long as they represent useful mediators of cognitive capacities such as prediction. Accordingly, the modelling perspective we adopted here is complementary, not alternative, to existing theories of behaviour and brain function such as predictive coding, which is increasingly protagonist of the music-cognition discourse (Rohrmeier and Koelsch, 2012; Vuust et al., 2022). Specifically, predictive-coding accounts are often agnostic with respect to the specific nature of the predictive model that generates predictions, against which prediction error is computed. Grammar-based syntactic modelling fulfils the role of specifying (and testing) the nature of the representations that a predictive model may exploit to generate such predictions. In turn, prediction-optimisation may constitute one of the constraints that determine the acquisition of a given grammar.

Models of probabilistic parsing will make it possible to directly compare, and possibly integrate, the phenomenology of structural hearing, as characterised here in terms of grammar-based representations, in the context of the advancing literature on predictive coding in music. Research on the computational modelling of music already advanced formalisations of musical structure in the form of probabilistic grammars (Abdallah et al., 2016; Finkensiep and Rohrmeier, 2021; Harasim et al., 2018). Implementations of probabilistic parsing have been shown to perform above chance after learning from annotated corpora (Finkensiep, 2023; Foscarin et al., 2023; Harasim, 2020), but in their current form they face difficulties that limit their direct applicability in empirical work. Such difficulties include (1) scaling up to pieces of arbitrary length, due to the combinatorial explosion of possible parses, and (2) achieving qualitatively satisfactory performance in replicating human-expert analytical insights. Both problems will need to be addressed in the future with innovative computational approaches, possibly including the combination of symbolic and deep-learning models (Foscarin et al., 2023). In particular, solving the latter problem also faces two additional hurdles. First, the limited availability of annotated corpora, which represents a major obstacle for training (cf. Ericson et al., 2023; Foscarin et al., 2023; Harasim et al., 2020). Second, the fact that analytical decisions made by humans are influenced by a variety of cues coming from all aspects of the musical surface: this means that a model of harmony that has no access to a representation of rhythm, or a model of rhythm that has no access to a representation of harmony, are unlikely to be able to infer the same interpretation as a human who has access to both. Addressing this issue may benefit from models that integrate grammars governing different domains of musical organisation (such as harmony and rhythm, cf. Harasim et al., 2019).

Overall, the refinement of parsing models, even non-cognitively-plausible ones, will contribute significantly to research on the cognition of musical structure, in particular with respect to the learnability of musical idioms (Harasim, 2020) and to modelling the ambiguity that is intrinsic to music. Furthermore, from a methodological perspective, it is important to mention that the stimuli to be employed in psychological and neuroscientific experiment currently have to be handcrafted by human experts on a case-by-case basis: this constrains both the scalability of experiments, as well as formal control on the features of the stimuli. The availability of probabilistic grammars and parsing models will greatly facilitate the construction of controlled stimuli to be employed in large-scale experiments, by automatising the attribution of structural interpretations to surfaces. As a consequence, we hope that our approach to the cognition of musical structure will motivate, and benefit from, advances in fields that are seemingly less related to the cognitive sciences such as music information retrieval, for the development and computational validation of new parsing models (Finkensiep, 2023; Foscarin et al., 2023; Harasim et al., 2018), and corpus studies, for the curation of annotated corpora useful for training such models (Ericson et al., 2023; Hentschel, Moss, et al., 2021) as well as for identifying structural features emerging in well-specified repertoires (Laneve et al., 2023; Moss et al., 2019; White, 2022).

9.4 Listeners and listenings

In the experiments and the examples comprised in this thesis, we have focused on models of structure that are conceived to reflect the specific rhythmic-harmonic idiom of Western extended tonality, from the early common practice to some instances of Jazz, Rock, and Pop music. Since the object of modelling is the listener's subjective experience, and the (expert) listener's introspection is an important tool of investigation, it is methodologically relevant to commit to a specific idiom, and that this idiom falls within the scope of the authors' own cultural familiarity and training. On the other hand, such a commitment necessarily undermines the generalisability of the findings to other musical practices: are, e.g., structural priming

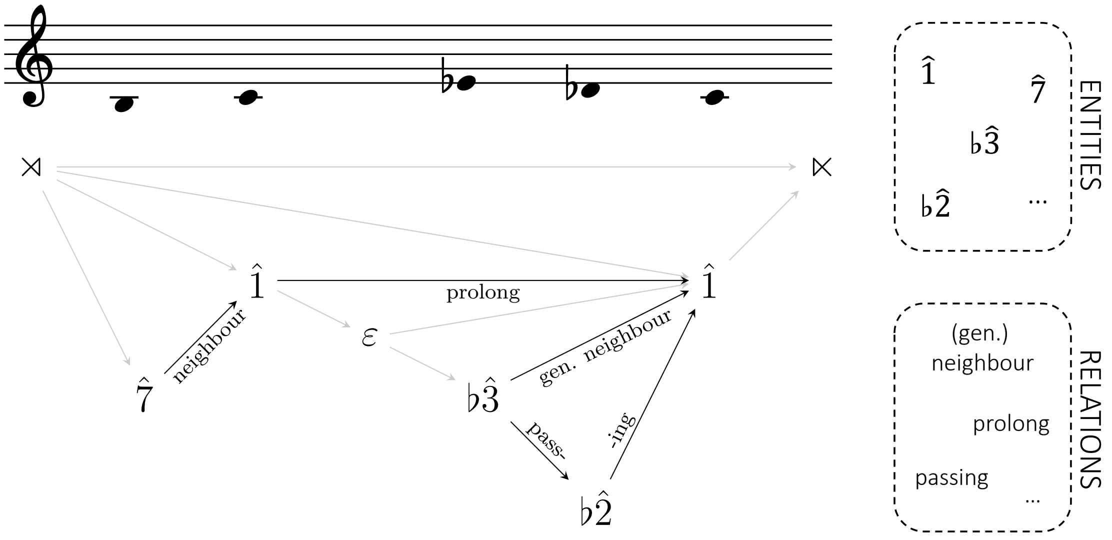


Figure 9.1 – Structural interpretation of a melody in the idiom of North-Indian classical music, after Finkensiep et al. (2019). The entities comprise scale degrees in the *Multani* rāga, which can be related in terms of neihbouring motions and passing steps. The interpretation is expressed as a labelled graph, similarly to Figure 1.5.

and garden-path effects phenomena that only apply to the Western-enculturated listener, and to (extended) tonal repertoires? The studies presented here were not designed to answer this question. However, our results do constitute a proof of existence of certain structural-processing phenomena in the musical domain *tout-court*: we cannot tell whether they exist for all listeners, but we have novel evidence that music as a universal human cognitive domain – distinct from, e.g., language – can *in principle* engender these phenomena. Recruiting the specific cognitive architecture that underlies the processing and representational phenomena we have identified is then one of the options for music makers (across musical cultures) to manipulate the listening experience in their community of listeners.

Future research may investigate the kind of structural representations that listeners my infer in idioms other than (extended) tonality: for example, Figure 9.1 exemplifies a graph-grammar modelling the interpretation of North-Indian classical music. Such grammars would reflect idiom-specific structuring principles, bearing their own implications towards plausible parsing mechanisms. In particular, different idioms may appeal to different entities and relations as their structural primitives. It is still possible that some forms of musical organisation may be widespread or even universal across cultures (Mehr et al., 2019; Savage et al., 2021). By comparing Figure 9.1 to Figure 1.5, for example, we see that – while the latter also comprises harmonic relations that are specific to Western tonality – both comprise somewhat analogous kinds of relations between individual tones (passing, neighbour). Underlying this analogy is the shared use of modes as a discrete and hierarchically-organised space of pitch-classes, and criteria of melodic motion favouring small intervals (passing and neighbour tones) and skips between relatively stable tones (generalised neighbours). However, analogies of this kind have to be considered carefully on a case-by-case basis, relying on the introspective intuition of listeners that are culturally familiar with each given idiom.

An assumption of our modelling framework – to be further investigated empirically in other cultural settings – is that parsing surfaces into structural representations is an automatic manifestation of the ordinary effort to "make sense" of the incoming sensory stimuli in terms of (possibly illusory) latent causes. However, it is to be expected that the resulting representations would only converge to some common "grammar" within communities of similarly enculturated listeners (i.e., listeners that are familiar with the same musical repertoire). Even among listeners with a shared cultural background, formal musical expertise may represent an additional factor of individual variability. In our studies, we found only marginal effects of musical expertise. In particular musical training was not a significant mediator of musical garden-path effects. However, investigating the role of musical expertise was not the main focus of our studies, and remains open for future research to assess. In particular, musical training may impact both the convergence of the listener's inferred representations with those that can be predicted by music-theoretical means, as well as aspects of processing, including the capacity to deploy cognitive functions such as attention, memory, prediction, and auditory scene analysis to identify structural cues which in turn may influence the inference process.

More generally, both cultural context and musical expertise within a cultural context may impact the listener's experience of structure also by influencing the *mode* of listening. Musical structure is, almost by definition, latent, and as such, it is something that may be *discovered*. However, engaging with the process of discovery is not necessarily something that listeners ought to do, depending on their personal attitude (possibly reflecting their training) as well as cultural biases. After all, musical cultures differ widely in terms of how music is meant to be engaged with: relevant distinctions include presentational vs. participatory (Nettl, 1999, cf. Cross, 2014), primarily notational vs. primarily oral (e.g., Shelemay, 2008), or composed vs. improvised (e.g., Schuiling, 2022). Each of these dimensions may change the role of the listener with respect to their agency in shaping their own musical experience, as well as the social and environmental setting where the act of listening takes place. Many genres in Western musical tradition are somewhat peculiar in this respect, as they assume a listener that is "just" listening, possibly in darkness and silence, and possibly multiple times, to an "immutable" work (Levinson, 1980). An attitude of *discovery* towards what is *latent* in the music may play a larger role in musical idioms that share these characteristics. Of course, there is nothing structural "in the music" (or rather, in the sounds) that is not attributed to it in the mind of the listener, but the listener's attitude may certainly impact what cognitive processes are recruited and how they are prioritised during listening. This has two implications towards empirical research. First, culture-specific modes of engagement with the music should be taken into account in cross-cultural work concerned with the inference of musical structure, noting that no listening modality is intrinsically primary and all should be explored, while bearing in mind that some are more representative of a specific musical practice than others.

As a corollary, an active endeavour of discovery based on attentive, skilled, creative, and repeated listening should become part of the repertoire of experimental settings, complementing the standard empirical approaches where the result of first-pass listening only is probed. Structural hearing, as it is described in the music-theoretical discourse, is not (necessarily) an all-or-nothing phenomenon: different degrees of active engagement may result in the emergence of richer or coarser, more complete or more partial representations. Tracking the spectrum of possible manifestations of a given music theory, and of structural representations for individual surfaces, as a function of the listener's agency and expertise is an aspect that was largely neglected in the present studies – which mainly aimed to established proofs of existence for certain phenomena. Nevertheless, this represents a desirable and challenging avenue for future empirical research – one that may bring this music-theoretically inspired psychology closer to the actual practices of musicians.

9.5 Conclusions

In conclusion, we have proposed an empirical approach for investigating the emergence of structural hearing during listening to music. The listener's experience of structure is understood as the result of inferring latent structural relations that constitute useful explanations of the intentional arrangement of the musical surface. Grammar-based incremental parsing is proposed as an algorithmic strategy for implementing such inference.

Results from our behavioural experiments support the cognitive relevance of persistent memory encodings of structural representations that are abstracted from sensory information and emerge implicitly and automatically during listening. In the context of (extended) tonal harmony, such representations are consistent with an underlying grammar allowing for hierarchical structural relations among abstract functional categories of harmonic entities. The processes that result in the emergence of such representations reflect many features that have been theorised and observed for cognitively-plausible linguistic parsers, such as structural priming, dependency-locality effects, and garden-paths.

Overall, by providing a first proof-of-existence for a number of concrete phenomena pertaining to the emergence of structural "illusions", this work demonstrates a viable approach for integrating elusive music-theoretical introspection into the purview of empirical methodologies in the cognitive sciences. Our results pave the way for an investigation of structural processing in music to a level of detail that is comparable with that achieved in psycholinguistics for language. Building up on these results, future research may aim at further bridging the gap between Marr's levels by (1) refining our understanding of candidate computational-level grammar models for individual musical idioms, (2) testing more fine-grained parsing models at the algorithmic level, and (3) exploring their implications towards brain function at the implementational level.

Bibliography

Abbott, J., Hamrick, J., & Griffiths, T. (2013). Approximating Bayesian inference with a sparse distributed memory system. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35).

Abdallah, S., Gold, N., & Marsden, A. (2016). Analysing symbolic music with probabilistic grammars. In D. Meredith (Ed.), *Computational Music Analysis* (pp. 157–189). Springer.

Abrams, K., & Bever, T. G. (1969). Syntactic structure modifies attention during speech perception and recognition. *Quarterly Journal of Experimental Psychology*, *21*(3), 280–290. https://doi.org/10.1080/14640746908400223

Agmon, E. (1995). Functional Harmony Revisited: A Prototype-Theoretic Approach. *Music Theory Spectrum*, *17*(2), 196–214. https://doi.org/10.2307/745871

Allanbrook, W. J. (2008). Mozart's K331, First Movement: Once More, with Feeling. In D. Mirka & K. Agawu (Eds.), *Communication in Eighteenth-Century Music* (pp. 254–282). Cambridge University Press. https://doi.org/10.1017/CBO9780511481376.010

Anzuoni, E., Dutto, F., Mcleod, A., Moss, F. C., & Rohrmeier, M. (2021). A Historical Analysis of Harmonic Progressions Using Chord Embeddings. *Proceedings of the 18th Sound and Music Computing Conference*. https://doi.org/10.5281/zenodo.5038909

Arnell, K. M. (2006). Visual, auditory, and cross-modality dual-task costs: electrophysiological evidence for an amodal bottleneck on working memory consolidation. *Perception & Psychophysics*, *68*(3), 447–457. https://doi.org/10.3758/bf03193689

Arnell, K. M., & Jolicœur, P. (1999). The attentional blink across stimulus modalities: Evidence for central processing limitations. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(3), 630–648. https://doi.org/10.1037/0096-1523.25.3.630

Arthur, C. (2018). A Perceptual Study of Scale-degree Qualia in Context. *Music Perception*, *35*(3), 295–314. https://doi.org/10.1525/mp.2018.35.3.295

Asano, R. (2021). The evolution of hierarchical structure building capacity for language and music: a bottom-up perspective. *Primates*. https://doi.org/10.1007/s10329-021-00905-x

Asano, R., & Boeckx, C. (2015). Syntax in language and music: what is the right level of comparison? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00942

Asano, R., Boeckx, C., & Seifert, U. (2021). Hierarchical control as a shared neurocognitive mechanism for language and music. *Cognition*, *216*. https://doi.org/10.1016/j.cognition.2021.104847

Bibliography

Atalay, N. B., & Tekman, H. G. (2006). Integration of non-diatonic chords into diatonic sequences: Scrambling sequences with secondary dominant chords. *9th International Conference on Music Perception and Cognition*.

Augustine. (2012). *Augustine Confessions: Augustine Confessions: Volume 1: Introduction and Text* (J. J. O'Donnell, Ed.). Oxford University Press.

Baird, A., & Samson, S. (2014). Music evoked autobiographical memory after severe acquired brain injury: Preliminary findings from a case series. *Neuropsychological Rehabilitation*, *24*(1), 125–143. https://doi.org/10.1080/09602011.2013.858642

Baltieri, M., & Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *The Behavioral and Brain Sciences*, *42*, e218. https://doi.org/10.1017/S0140525X19001353

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*(3), 241–253. https://doi.org/10.1080/net.12.3.241.253

Baroni, M. (1999). Musical Grammar and the Study of Cognitive Processes of Composition. *Musicae Scientiae*, *3*(1), 3–21. https://doi.org/10.1177/102986499900300101

Baroni, M., Maguire, S., & Drabkin, W. (1983). The Concept of Musical Grammar. *Music Analysis*, *2*(2), 175–208. https://doi.org/10.2307/854248

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. https://doi.org/10.1016/j.jml.2012.11.001

Beach, D. (1990). The Cadential Six-Four as Support for Scale-Degree Three of the Fundamental Line. *Journal of Music Theory*, *34*(1), 81–99. https://doi.org/10.2307/843863

Beach, D. (1995). Phrase Expansion: Three Analytical Studies. *Music Analysis*, *14*(1), 27–47. https://doi.org/10.2307/853961

Benjamin, L., Fló, A., Al Roumi, F., & Dehaene-Lambertz, G. (2023). Humans parsimoniously represent auditory sequences by pruning and completing the underlying network structure. *eLife*, *12*, e86430. https://doi.org/10.7554/eLife.86430

Bent, I. (1990). *Analisi musicale* (W. Drabkin, Ed.). EDT.

Berent, I., & Perfetti, C. A. (1993). An on-line method in studying music parsing. *Cognition*, *46*(3), 203–222. https://doi.org/10.1016/0010-0277(93)90010-S

Bergé, P., & Neuwirth, M. (Eds.). (2015). *What is a cadence? theoretical and analytical perspectives on cadences in the classical repertoire*. Leuven University Press.

Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.675558

Berman, S., Friedman, D., & Cramer, M. (1991). ERPs during continuous recognition memory for words and pictures. *Bulletin of the Psychonomic Society*, *29*(2), 113–116. https://doi.org/10.3758/BF03335209

Bernstein, L. (1976). *The Unanswered Question: Six Talks at Harvard*. Harvard University Press.

Berwick, R. C. C., & Weinberg, A. S. (1984). *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition*. MIT Press.

Bever, T. G., Lackner, J. R., & Kirk, R. (1969). The underlying structures of sentences are the primary units of immediate speech processing. *Perception & Psychophysics*, *5*(4), 225–234. https://doi.org/10.3758/BF03210545

Bharucha, J. J. (1987). Music Cognition and Perceptual Facilitation: A Connectionist Framework. *Music Perception*, *5*(1), 1–30. https://doi.org/10.2307/40285384

Bharucha, J. J., & Stoeckig, K. (1986). Reaction time and musical expectancy: Priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 403–410. https://doi.org/10.1037/0096-1523.12.4.403

Bharucha, J. J., & Stoeckig, K. (1987). Priming of chords: Spreading activation or overlapping frequency spectra? *Perception & Psychophysics*, *41*(6), 519–524. https://doi.org/10.3758/BF03210486

Bharucha, J. J., & Todd, P. M. (1989). Modeling the Perception of Tonal Structure with Neural Nets. *Computer Music Journal*, *13*(4), 44–53. https://doi.org/10.2307/3679552

Biamonte, N. (2008). Augmented-Sixth Chords vs. Tritone Substitutes. *Music Theory Online*, *14*(2).

Bianco, R., Novembre, G., Keller, P. E., Kim, S.-G., Scharf, F., Friederici, A. D., Villringer, A., & Sammler, D. (2016). Neural networks for harmonic structure in music perception and action. *NeuroImage*, *142*, 454–464. https://doi.org/10.1016/j.neuroimage.2016.08.025

Bianco, R., Novembre, G., Keller, P. E., Scharf, F., Friederici, A. D., Villringer, A., & Sammler, D. (2015). Syntax in Action Has Priority over Movement Selection in Piano Playing: An ERP Study. *Journal of Cognitive Neuroscience*, *28*(1), 41–54. https://doi.org/10.1162/jocn_a_00873

Bigand, E., & Pineau, M. (1997). Global context effects on musical expectancy. *Perception & Psychophysics*, *59*(7), 1098–1107. https://doi.org/10.3758/BF03205524

Bigand, E., Tillmann, B., Poulin-Charronnat, B., & Manderlier, D. (2005). Repetition priming: Is music special? *The Quarterly Journal of Experimental Psychology Section A*, *58*(8), 1347–1375. https://doi.org/10.1080/02724980443000601

Bigand, E. (1990). Abstraction of Two Forms of Underlying Structure in a Tonal Melody. *Psychology of Music*, *18*(1), 45–59. https://doi.org/10.1177/0305735690181004

Bigand, E., Delbé, C., Poulin-Charronnat, B., Leman, M., & Tillmann, B. (2014). Empirical evidence for musical syntax processing? Computer simulations reveal the contribution of auditory short-term memory. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00094

Bigand, E., & Parncutt, R. (1999). Perceiving musical tension in long chord sequences. *Psychological Research*, *62*(4), 237–254. https://doi.org/10.1007/s004260050053

Bigand, E., Parncutt, R., & Lerdahl, F. (1996). Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, *58*(1), 125–141. https://doi.org/10.3758/BF03205482

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., & D'Adamo, D. A. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology:*

Human Perception and Performance, *29*(1), 159–171. https://doi.org/10.1037/0096-1523.29.1.159

Bisesi, E. (2017). Measuring and Modelling Perceived Distance Among Collections in Post-Tonal Music: Music theory Meets Music Psychology. *9th European Music Analysis Conference*.

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*(2), 163–186. https://doi.org/10.1016/0010-0277(89)90022-X

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*(1), 1–39. https://doi.org/10.1016/0010-0277(90)90035-I

Boltz, M., & Jones, M. R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, *18*(4), 389–431. https://doi.org/10.1016/0010-0285(86)90005-8

Boretz, B. (1977). What Lingers on (, When the Song Is Ended). *Perspectives of New Music*, *16*(1), 102–109. https://doi.org/10.2307/832851

Born, G. (2011). Music and the materialization of identities. *Journal of Material Culture*, *16*(4), 376–388. https://doi.org/10.1177/1359183511424196

Bornkessel, I. D., Fiebach, C. J., & Friederici, A. D. (2004). On the cost of syntactic ambiguity in human language comprehension: an individual differences approach. *Cognitive Brain Research*, *21*(1), 11–21. https://doi.org/10.1016/j.cogbrainres.2004.05.007

Bostwick, J., Seror, G. A., & Neill, W. T. (2018). Tonality Without Structure. Using Drones to Induce Modes and Convey Moods. *Music Perception*, *36*(2), 243–249. https://doi.org/10.1525/mp.2018.36.2.243

Branigan, H. P., & Pickering, M. J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*, e282. https://doi.org/10.1017/S0140525X16002028

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., & Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, *24*(6), 489–506. https://doi.org/10.1007/BF02143163

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 468–481. https://doi.org/10.1037/0278-7393.31.3.468

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Bradford.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brown, J., Tan, D., & Baker, D. J. (2021). The Perceptual Attraction of Pre-Dominant Chords. *Music Perception*, *39*(1), 21–40. https://doi.org/10.1525/mp.2021.39.1.21

Brown, M., & Dempster, D. J. (1989). The Scientific Image of Music Theory. *Journal of Music Theory*, *33*(1), 65. https://doi.org/10.2307/843666

Bruford, F., Lartillot, O., McDonald, S., & Sandler, M. (2020). Multidimensional similarity modelling of complex drum loops using the GrooveToolbox. *21st International Society for Music Information Retrieval Conference*.

Bruner, C. L. (1984). The Perception of Contemporary Pitch Structures. *Music Perception*, *2*(1), 25–39. https://doi.org/10.2307/40285280

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Bürkner, P.-C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420–451. https://doi.org/10.1111/bmsp.12195

Cadwallader, A. (1992). More on Scale Degree Three and the Cadential Six-Four. *Journal of Music Theory*, *36*(1), 187–198. https://doi.org/10.2307/843914

Cadwallader, A., & Gagne, D. (2010). *Analysis of Tonal Music: A Schenkerian Approach*. Oxford University Press.

Calma-Roddin, N., & Drury, J. (2020). Music, Language, and The N400: ERP Interference Patterns Across Cognitive Domains. *Scientific Reports*, *10*, 1–14. https://doi.org/10.1038/s41598-020-66732-0

Cambouropoulos, E. (2010). The Musical Surface: Challenging Basic Assumptions [Publisher: SAGE Publications Ltd]. *Musicae Scientiae*, *14*(2), 131–147. https://doi.org/10.1177/10298649100140S209

Campos, A., & Fuentes, L. (2016). Musical Studies and the Vividness and Clarity of Auditory Imagery. *Imagination, Cognition and Personality*, *36*(1), 75–84. https://doi.org/10.1177/0276236616635985

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(01). https://doi.org/10.1017/S0140525X99001788

Caplin, W. E. (2002). Theories of musical rhythm in the eighteenth and nineteenth centuries. In T. Christensen (Ed.), *The Cambridge History of Western Music Theory* (1st ed., pp. 657–694). Cambridge University Press. https://doi.org/10.1017/CHOL9780521623711.023

Caplin, W. E. (1998). *Classical form: a theory of formal functions for the instrumental music of Haydn, Mozart, and Beethoven*. Oxford University Press.

Capuzzo, G. C. (2004). Neo-Riemannian Theory and the Analysis of Pop-Rock Music. *Music Theory Spectrum*, *26*(2), 177–200. https://doi.org/10.1525/mts.2004.26.2.177

Cariani, P. A., & Delgutte, B. (1996). Neural Correlates of the Pitch of Complex Tones. I. Pitch and Pitch Salience. *Journal of Neurophysiology*, *76*(3), 1698–1716.

Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of north India. *Journal of Experimental Psychology. General*, *113*(3), 394–412.

Cecchetti, G., Herff, S. A., Finkensiep, C., Harasim, D., & Rohrmeier, M. (2023). Hearing functional harmony in jazz: A perceptual study of music-theoretical accounts of extended tonality. *Musicae Scientiae*, *27*(3), 672–697. https://doi.org/10.1177/10298649221122245

Cecchetti, G., Herff, S. A., Finkensiep, C., & Rohrmeier, M. (2020). The experience of musical structure as computation: what can we learn? *Rivista di Analisi e Teoria Musicale*, *26*(2), 91–127. https://doi.org/10.53152/1032

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

Cecchetti, G., Herff, S. A., & Rohrmeier, M. (2022). Musical Garden Paths: Evidence for Syntactic Revision Beyond the Linguistic Domain. *Cognitive Science*, *46*(7), e13165. https://doi.org/10.1111/cogs.13165

Cecchetti, G., Herff, S. A., & Rohrmeier, M. (in review). Priming of abstract harmonic structure in music.

Cecchetti, G., Tomasini, C. A., Herff, S. A., & Rohrmeier, M. (2023). Interpreting Rhythm as Parsing: Syntactic-Processing Operations Predict the Migration of Visual Flashes as Perceived During Listening to Musical Rhythms. *Cognitive Science*, *47*(12), e13389. https://doi.org/10.1111/cogs.13389

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344. https://doi.org/10.1016/j.tics.2006.05.006

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *WIREs Cognitive Science*, *1*(6), 811–823. https://doi.org/10.1002/wcs.79

Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., Malik-Moraleda, S., Kean, H., Varley, R., & Fedorenko, E. (2021). The human language system does not support music processing. https://doi.org/10.1101/2021.06.01.446439

Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., Malik-Moraleda, S., Kean, H., Varley, R., & Fedorenko, E. (2023). The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cerebral Cortex*, *33*(12), 7904–7929. https://doi.org/10.1093/cercor/bhad087

Cheung, V. K. M., Harrison, P. M. C., Meyer, L., Pearce, M. T., Haynes, J.-D., & Koelsch, S. (2019). Uncertainty and Surprise Jointly Predict Musical Pleasure and Amygdala, Hippocampus, and Auditory Cortex Activity. *Current Biology*, *29*(23), 4084–4092. https://doi.org/10.1016/j.cub.2019.09.067

Cheung, V. K. M., Meyer, L., Friederici, A. D., & Koelsch, S. (2018). The right inferior frontal gyrus processes nested non-local dependencies in music. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-22144-9

Chomsky, N. (1957). *Syntactic structures*. de Gruyter.

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

Chomsky, N. (2002). Perspectives on language and mind. In *On Nature and Language* (pp. 45–60). Cambridge University Press. https://doi.org/10.1017/CBO9780511613876.003

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509. https://doi.org/10.1017/S0140525X08004998

Clark, A. (2013a). Learning Trees from Strings: A Strong Learning Algorithm for some Context-Free Grammars. *Journal of Machine Learning Research*, *14*(111), 3537–3559.

Clark, A. (2017). Computational Learning of Syntax. *Annual Review of Linguistics*, *3*(1), 107–123. https://doi.org/10.1146/annurev-linguistics-011516-034008

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. https://doi.org/10.1017/S0140525X12000477

Clayton, M. (2012). What is Entrainment? Definition and applications in musical research. *Empirical Musicology Review*, *7*(1-2), 49–56. https://doi.org/10.18061/1811/52979

Cohen, A. J. (2002). Music cognition and the cognitive psychology of film structure. *Canadian Psychology/Psychologie canadienne*, *43*(4), 215–232. https://doi.org/10.1037/h0086918

Cohen, M. A., Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychonomic Bulletin & Review*, *18*(3), 586–591. https://doi.org/10.3758/s13423-011-0074-0

Cohn, N. (2020). Your Brain on Comics: A Cognitive Model of Visual Narrative Comprehension. *Topics in Cognitive Science*, *12*(1), 352–386. https://doi.org/10.1111/tops.12421

Cohn, R. (1999). As wonderful as star clusters: instruments for gazing at tonality in Schubert. *Nineteenth-Century Music.*, *22*(3), 213–232.

Cohn, R. (2007). Hexatonic Poles and the Uncanny in Parsifal. *The Opera Quarterly*, *22*(2), 230–248. https://doi.org/10.1093/oq/kbl008

Cohn, R. (2012). *Audacious Euphony: Chromatic Harmony and the Triad's Second Nature*. Oxford University Press, USA.

Cohn, R. (2016). On Hexatonic Poles. *Music Analysis*, *35*(1), 134–138. https://doi.org/10.1111/musa.12063

Cohn, R. (2020). Meter. In A. Rehding & S. Rings (Eds.), *The Oxford Handbook of Critical Concepts in Music Theory* (pp. 207–233). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190454746.013.9

Cook, N. (1987). The Perception of Large-Scale Tonal Closure. *Music Perception: An Interdisciplinary Journal*, *5*(2), 197–205. https://doi.org/10.2307/40285392

Cook, N. (1999). Analysing Performance and Performing Analysis. In N. Cook & M. Everist (Eds.), *Rethinking Music* (pp. 239–261). Oxford University Press.

Cross, I. (1998). Music Analysis and Music Perception. *Music Analysis*, *17*(1), 3. https://doi.org/10.2307/854368

Cross, I. (2009). The evolutionary nature of musical meaning. *Musicae Scientiae*, *13*(2), 179–200. https://doi.org/10.1177/1029864909013002091

Cross, I. (2014). Music and communication in music psychology. *Psychology of Music*, *42*(6), 809–819. https://doi.org/10.1177/0305735614543968

Cuddy, L. L., Cohen, A. J., & Mewhort, D. J. K. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 869–883. https://doi.org/10.1037/0096-1523.7.4.869

Cuddy, L. L., Cohen, A. J., & Miller, J. (1979). Melody recognition: The experimental application of musical rules. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *33*(3), 148–157. https://doi.org/10.1037/h0081713

Cuddy, L. L., Sikka, R., & Vanstone, A. (2015). Preservation of musical memory and engagement in healthy aging and Alzheimer's disease. *Annals of the New York Academy of Sciences*, *1337*(1), 223–231. https://doi.org/10.1111/nyas.12617

Cusimano, M., Hewitt, L., Tenenbaum, J. B., & McDermott, J. H. (2018). Auditory scene analysis as Bayesian inference in sound source models. *MIT web domain.*

Danielsen, A., Johansson, M., & Stover, C. (2023). Bins, Spans, and Tolerance: Three Theories of Microtiming Behavior. *Music Theory Spectrum, 45.* https://doi.org/10.1093/mts/mtad005

Davis, S. (2006). Implied Polyphony in the Solo String Works of J. S. Bach: A Case for the Perceptual Relevance of Structural Expression. *Music Perception: An Interdisciplinary Journal, 23*(5), 423–446. https://doi.org/10.1525/mp.2006.23.5.423

Davis, S. (2011). Stream Segregation and Perceived Syncopation: Analyzing the Rhythmic Effects of Implied Polyphony in Bach's Unaccompanied String Works. *Music Theory Online, 17*(1), 15.

Dean, R. T. (Ed.). (2011). *The Oxford Handbook of Computer Music.* Oxford University Press.

de Clercq, T., & Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music, 30*(1), 47–70. https://doi.org/10.1017/S026114301000067X

Deliege, I. (1987). Grouping Conditions in Listening to Music: An Approach to Lerdahl & Jackendoff's Grouping Preference Rules. *Music Perception, 4*(4), 325–359. https://doi.org/10.2307/40285378

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics, 28*(5), 381–389. https://doi.org/10.3758/BF03204881

Deutsch, D. (1999). Grouping Mechanisms in Music. In D. Deutsch (Ed.), *The Psychology of Music (Second Edition)* (pp. 299–348). Academic Press. https://doi.org/10.1016/B978-012213564-4/50010-X

DeVoto, M. (2004). *Debussy and the Veil of Tonality: Essays on His Music.* Pendragon Press.

Di Stefano, N., Vuust, P., & Brattico, E. (2022). Consonance and dissonance perception. A critical review of the historical sources, multidisciplinary findings, and main hypotheses. *Physics of Life Reviews, 43,* 273–304. https://doi.org/10.1016/j.plrev.2022.10.004

Dibben, N. (1994). The Cognitive Reality of Hierarchic Structure in Tonal and Atonal Music. *Music Perception: An Interdisciplinary Journal, 12*(1), 1–25. https://doi.org/10.2307/40285753

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

Doğantan-Dack, M. (2001). Upbeat. *Grove Music Online.*

Doll, C. (2017). *Hearing Harmony* (University of Michigan Press).

Dowling, W. J. (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception & Psychophysics, 14*(1), 37–40. https://doi.org/10.3758/BF03198614

Dowling, W. J., Tillman, B., & Ayers, D. F. (2001). Memory and the experience of hearing music. *Music Perception, 19*(2), 249–276. https://doi.org/10.1525/mp.2001.19.2.249

Drabkin, W., Pasticci, S., & Pozzi, E. (1995). *Analisi schenkeriana. Per un'interpretazione organica della struttura musicale.* LIM.

Dubiel, J. (1990). "When You are a Beethoven": Kinds of Rules in Schenker's "Counterpoint". *Journal of Music Theory, 34*(2), 291–340. https://doi.org/10.2307/843840

Dubiel, J. (2000). Analysis, Description, and What Really Happens. *Music Theory Online*, *6*(3).

Dubiel, J. (2017). Music Analysis and Kinds of Hearing-As. *Music Theory and Analysis (MTA)*, *4*(2), 233–242. https://doi.org/10.11116/MTA.4.2.4

Egermann, H., Pearce, M. T., Wiggins, G. A., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, affective & behavioral neuroscience*, *13*(3), 533–553. https://doi.org/10.3758/s13415-013-0161-y

Ellis, R. J., & Jones, M. R. (2009). The role of accent salience and joint accent structure in meter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 264–280. https://doi.org/10.1037/a0013482

Ericson, P., Rammos, Y., & Rohrmeier, M. (2023). Musereduce : a generic framework for hierarchical music analysis. *Music Encoding Conference 2022, Halifax, Canada, May 19-22, 2022*, 40–51.

Euler, L. (1773). De Harmoniae Veris Principiis per Speculum Musicum Represaentatis. In *Novi commentarii Academiae Scientiarum Imperialis Petropolitanae* (pp. 330–353). Typis Academiae Scientarum.

Everett, W. (2004). Making Sense of Rock's Tonal Systems. *Music Theory Online*, *10*(4).

Eysenck, M. W., & Keane, M. T. (2015). *Cognitive psychology: A student's handbook*. Psychology Press. https://doi.org/10.4324/9781315778006

Farbood, M. M. (2016). Memory of a Tonal Center After Modulation. *Music Perception: An Interdisciplinary Journal*, *34*(1), 71–93. https://doi.org/10.1525/mp.2016.34.1.71

Farbood, M. M. (2012). A Parametric, Temporal Model of Musical Tension. *Music Perception: An Interdisciplinary Journal*, *29*(4), 387–428. https://doi.org/10.1525/mp.2012.29.4.387

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*.

Farbood, M. M., & Mavromatis, P. (2018). The mutability of pitch memory in a tonal context. *Psychomusicology: Music, Mind, and Brain*, *28*(1), 1–16. https://doi.org/10.1037/pmu0000205

Farbood, M. M., & Price, K. C. (2017). The contribution of timbre attributes to musical tension. *The Journal of the Acoustical Society of America*, *141*(1), 419–427. Retrieved February 5, 2024, from https://pubs.aip.org/asa/jasa/article/141/1/419/1058232

Featherstone, C. R., Morrison, C. M., Waterman, M. G., & MacGregor, L. J. (2013). Semantics, Syntax or Neither? A Case for Resolution in the Interpretation of N500 and P600 Responses to Harmonic Incongruities. *PLOS ONE*, *8*(11), e76600. https://doi.org/10.1371/journal.pone.0076600

Featherstone, C. R., Waterman, M. G., & Morrison, C. M. (2012). Norming the odd: Creation, norming, and validation of a stimulus set for the study of incongruities across music and language. *Behavior Research Methods*, *44*(1), 81–94. https://doi.org/10.3758/s13428-011-0137-1

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition*, *37*(1), 1–9. https://doi.org/10.3758/MC.37.1.1

Fedorenko, E., & Shain, C. (2021). Similarity of Computations Across Domains Does Not Imply Shared Implementation: The Case of Language Comprehension. *Current Directions in Psychological Science*, *30*(6), 526–534. https://doi.org/10.1177/09637214211046955

Feld, S. (1984). Sound Structure as Social Structure. *Ethnomusicology*, *28*(3), 383–409. https://doi.org/10.2307/851232

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of Garden-Path Sentences: Implications for Models of Sentence Processing and Reanalysis. *Journal of Psycholinguistic Research*, *30*(1), 3–20. https://doi.org/10.1023/A:1005290706460

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*(6), 725–745. https://doi.org/10.1016/0749-596X(91)90034-H

Ferreira, F., & Henderson, J. M. (1998). Syntactic Reanalysis, Thematic Processing, and Sentence Comprehension. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in Sentence Processing* (pp. 73–100). Springer Netherlands. https://doi.org/10.1007/978-94-015-9070-9_3

Finkensiep, C. (2023). *The Structure of Free Polyphony* (Doctoral dissertation). École Polytechnique Fédérale de Lausanne. Lausanne.

Finkensiep, C., Haeberle, M., Eisenbrand, F., Neuwirth, M., & Rohrmeier, M. (2023). Repetition-structure inference with formal prototypes. *Proceedings of the 24th International Society for Music Information Retrieval Conference*.

Finkensiep, C., & Rohrmeier, M. (2021). Modeling and inferring proto-voice structure in free polyphony. *Proceedings of the 22nd International Society for Music Information Retrieval Conference*.

Finkensiep, C., Widdess, R., & Rohrmeier, M. (2019). Modelling the syntax of north indian melodies with a generalized graph grammar. *Proceedings of the 20th International Society for Music Information Retrieval Conference*.

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1933–1955. https://doi.org/10.1098/rstb.2012.0103

Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition*, *97*(2), 179–210. https://doi.org/10.1016/j.cognition.2005.02.005

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

Fitzroy, A. B., & Sanders, L. D. (2015). Musical Meter Modulates the Allocation of Attention across Time. *Journal of Cognitive Neuroscience*, *27*(12), 2339–2351. https://doi.org/10.1162/jocn_a_00862

Fiveash, A. (2018). *The Nature of Syntactic Processing in Music and Language* (Doctoral dissertation). Macquarie University.

Fiveash, A., Bedoin, N., Gordon, R. L., & Tillmann, B. (2021). Processing rhythm in speech and music: Shared mechanisms and implications for developmental speech and language disorders. *Neuropsychology*, *35*, 771–791. https://doi.org/10.1037/neu0000766

Fiveash, A., Ladányi, E., Camici, J., Chidiac, K., Bush, C. T., Canette, L.-H., Bedoin, N., Gordon, R. L., & Tillmann, B. (2023). Regular rhythmic primes improve sentence repetition in children with developmental language disorder. *Science of Learning*, *8*(1), 1–8. https://doi.org/10.1038/s41539-023-00170-1

Fiveash, A., McArthur, G., & Thompson, W. F. (2018). Syntactic and non-syntactic sources of interference by music on language processing. *Scientific Reports*, *8*(1), 17918. https://doi.org/10.1038/s41598-018-36076-x

Fiveash, A., & Pammer, K. (2012). Music and language: Do they draw on similar syntactic working memory resources? *Psychology of Music*, *42*(2), 190–209. https://doi.org/10.1177/0305735612463949

Fodor, J., & Ferreira, F. (1998). *Reanalysis in Sentence Processing*. Springer Science & Business Media.

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, *4*(5), 414–420. https://doi.org/10.1016/S0022-5371(65)80081-0

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Forte, A. (1973). *The structure of atonal music*. Yale University Press.

Foscarin, F., Harasim, D., & Widmer, G. (2023). Predicting Music Hierarchies with a Graph-Based Neural Decoder. *Proceedings of the 24th International Society for Music Information Retrieval Conference*.

Foscarin, F., Jacquemard, F., & Rigaux, P. (2019). Modeling and Learning Rhythm Structure. *Sound and Music Computing Conference (SMC)*.

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies* (Doctoral dissertation). University of Massachusetts.

Frazier, L. (1987). Sentence processing: A tutorial review. In *Attention and performance 12: The psychology of reading* (pp. 559–586). Lawrence Erlbaum Associates, Inc.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325. https://doi.org/10.1016/0010-0277(78)90002-1

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Friederici, A. D. (1995). The Time Course of Syntactic Activation During Language Processing: A Model Based on Neuropsychological and Neurophysiological Data. *Brain and Language*, *50*(3), 259–281. https://doi.org/10.1006/brln.1995.1048

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84. https://doi.org/10.1016/S1364-6613(00)01839-8

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour*, *1*(10), 713–722. https://doi.org/10.1038/s41562-017-0184-4

Friederici, A. D., & Mecklinger, A. (1996). Syntactic parsing as revealed by brain responses: First-pass and second-pass parsing processes. *Journal of Psycholinguistic Research*, *25*(1), 157–176. https://doi.org/10.1007/BF01708424

Friedman, D. (1990). Cognitive Event-Related Potential Components During Continuous Recognition Memory for Pictures. *Psychophysiology*, *27*(2), 136–148. https://doi.org/10.1111/j.1469-8986.1990.tb00365.x

Frisch, S., Schlesewsky, M., Saddy, D., & Alpermann, A. (2002). The P600 as an indicator of syntactic ambiguity. *Cognition*, *85*(3), B83–B92. https://doi.org/10.1016/S0010-0277(02)00126-9

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, *3*.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., ODoherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Gallant, J., & Libben, G. (2019). No lab, no problem: Designing lexical comprehension and production experiments using PsychoPy3. *The Mental Lexicon*, *14*(1), 152–168. https://doi.org/10.1075/ml.00002.gal

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown* (Doctoral dissertation). Carnegie Mellon University.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 94–126). The MIT Press.

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, *2*(7), 262–268. https://doi.org/10.1016/S1364-6613(98)01187-5

Gibson, E., & Pearlmutter, N. J. (2000). Distinguishing Serial and Parallel Parsing. *Journal of Psycholinguistic Research*, *29*(2), 231–240. https://doi.org/10.1023/A:1005153330168

Gjerdingen, R., & Bourne, J. (2015). Schema Theory as a Construction Grammar. *Music Theory Online*, *21*(2).

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582. https://doi.org/10.1007/s11229-015-0762-9

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & Philosophy*, *32*(3), 337–355. https://doi.org/10.1007/s10539-017-9562-6

Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., & Zatorre, R. J. (2019). Predictability and uncertainty in the pleasure of music: a reward for learning? *The Journal of Neuroscience*, 0428–19. https://doi.org/10.1523/JNEUROSCI.0428-19.2019

Goldman, A., Harrison, P. M. C., Jackson, T., & Pearce, M. T. (2021). Reassessing Syntax-Related ERP Components Using Popular Music Chord SequencesA Model-Based Approach. *Music Perception*, *39*(2), 118–144. https://doi.org/10.1525/mp.2021.39.2.118

Gordon, E. E. (1985). Research Studies in Audiation: I. *Bulletin of the Council for Research in Music Education*, (84), 34–50.

Gran, J. (2017). Ornamental and Motivic Integration in Chopin's Op. 9 Nocturnes. *Indiana Theory Review*, *34*(1-2), 23–49. https://doi.org/10.2979/inditheorevi.34.1-2.02

Granroth-Wilding, M., & Steedman, M. (2014). A Robust Parser-Interpreter for Jazz Chord Sequences. *Journal of New Music Research*, *43*(4), 355–374. https://doi.org/10.1080/09298215.2014.910532

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Griffiths, T. L., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. https://doi.org/10.1184/R1/6613682.v1

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic Complexity in Ambiguity Resolution. *Journal of Memory and Language*, *46*(2), 267–295. https://doi.org/10.1006/jmla.2001.2808

Guichaoua, C., Besada, J., Bisesi, E., & Andreatta, M. (2021). The Tonnetz Environment: A Web Platform for Computer-aided "Mathemusical" Learning and Research. *Proceedings of the 13th International Conference on Computer Supported Education*, *1*, 680–689. https://doi.org/10.5220/0010532606800689

Gwilliams, L. (2020). *Towards a Mechanistic Account of Speech Comprehension in the Human Brain* (Doctoral dissertation). New York University.

Haas, B. (2004). *Die neue Tonalität von Schubert bis Webern. Hören und Analysieren nach Albert Simon*. Noetzel.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *NeuroImage*, *20*, S18–S29. https://doi.org/10.1016/j.neuroimage.2003.09.013

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

Halpern, A. R., & Bower, G. H. (1982). Musical Expertise and Melodic Structure in Memory for Musical Notation. *The American Journal of Psychology*, *95*(1), 31–50. https://doi.org/10.2307/1422658

Bibliography

Hamanaka, M., Hirata, K., & Tojo, S. (2006). Implementing “A Generative Theory of Tonal Music". *Journal of New Music Research*, *35*(4), 249–277. https://doi.org/10.1080/09298210701563238

Hamaoui, K., & Deutsch, D. (2004). Perceptual grouping of musical sequences: Pitch and timing as competing cues. *The Journal of the Acoustical Society of America*, *116*(4_Supplement), 2580. https://doi.org/10.1121/1.4785300

Hansen, N. C., Kragness, H. E., Vuust, P., Trainor, L., & Pearce, M. T. (2021). Predictive Uncertainty Underlies Auditory Boundary Perception. *Psychological Science*, *32*(9), 1416–1425. https://doi.org/10.1177/0956797621997349

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

Harasim, D., Finkensiep, C., Ericson, P., O’Donnell, T. J., & Rohrmeier, M. (2020). The Jazz harmony treebank. *International Society for Music Information Retrieval*. https://doi.org/DOI:10.5281/ZENODO.3546040

Harasim, D., Moss, F. C., Ramirez, M., & Rohrmeier, M. (2021). Exploring the foundations of tonality: statistical cognitive modeling of modes in the history of Western classical music. *Humanities and Social Sciences Communications*, *8*(1), 1–11. https://doi.org/10.1057/s41599-020-00678-6

Harasim, D., O’Donnell, T. J., & Rohrmeier, M. (2019). Harmonic Syntax in Time: Rhythm Improves Grammatical Models of Harmony. *Proceedings of the 20th International Society for Music Information Retrieval Conference*.

Harasim, D., Rohrmeier, M., & O’Donnell, T. J. (2018). A generalized parsing framework for generative models of harmonic syntax. *International Society for Music Information Retrieval*.

Harrison, D. (1994). *Harmonic Function in Chromatic Music: A Renewed Dualist Theory and an Account of Its Precedents*. University of Chicago Press.

Hauptmann, M. (1853). *Die Natur der Harmonik und der Metrik*. Breitkopf und Härtel.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic scale. *Psychology Bullettin*, *18*, 98–99.

Heine, E. (2018). Chromatic Mediants and Narrative Context in Film. *Music Analysis*, *37*(1), 103–132. https://doi.org/10.1111/musa.12106

Hentschel, J., Moss, F. C., Neuwirth, M., & Rohrmeier, M. (2021). A semi-automated workflow paradigm for the distributed creation and curation of expert annotations. *Proceedings of the 22nd International Society for Music Information Retrieval Conference*.

Hentschel, J., Neuwirth, M., & Rohrmeier, M. (2021). The Annotated Mozart Sonatas: Score, Harmony, and Cadence. *Transactions of the International Society for Music Information Retrieval*, *4*(1), 67–80. https://doi.org/10.5334/tismir.63

Hepokoski, J., & Darcy, W. (2011). *Elements of Sonata Theory: Norms, Types, and Deformations in the Late-Eighteenth-Century Sonata*. Oxford University Press.

Herff, S. A., Bonetti, L., Cecchetti, G., Vuust, P., Kringelbach, M. L., & Rohrmeier, M. (2023). Hierarchical syntax models of music predict theta power during music listening. https://doi.org/10.1101/2023.05.15.540878

Herff, S. A., Cecchetti, G., Ericson, P., & Cano, E. (2023). Solitary Silence and Social Sounds: Music influences mental imagery, inducing thoughts of social interactions. https://doi.org/10.1101/2023.06.22.546175

Herff, S. A., Cecchetti, G., Taruffi, L., & Déguernel, K. (2021). Music influences vividness and content of imagined journeys in a directed visual imagery task. *Scientific Reports*, *11*(1), 15990. https://doi.org/10.1038/s41598-021-95260-8

Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Primary_atype: Research Subject_term: Human behaviour;Psychology Subject_term_id: human-behaviour;psychology

Herff, S. A., & Czernochowski, D. (2019). The role of divided attention and expertise in melody recognition. *Musicae Scientiae*, *23*(1), 69–86. https://doi.org/10.1177/1029864917731126

Herff, S. A., Dean, R. T., & Olsen, K. N. (2017). Interrater agreement in memory for melody as a measure of listeners' similarity in music perception. *Psychomusicology: Music, Mind, and Brain*, *27*(4), 297–311. https://doi.org/10.1037/pmu0000197

Herff, S. A., Harasim, D., Cecchetti, G., Finkensiep, C., & Rohrmeier, M. (2021). Hierarchical syntactic structure predicts listeners' sequence completion in music. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*.

Herff, S. A., Herff, C., Milne, A., Johnson, G., Shih, J., & Krusienski, D. (2020). Prefrontal High Gamma in ECoG Tags Periodicity of Musical Rhythms in Perception and Imagination. *eNeuro*, *7*(4), ENEURO.0413–19.2020. https://doi.org/10.1523/ENEURO.0413-19.2020

Herff, S. A., Olsen, K. N., Anic, A., & Schaal, N. K. (2019). Investigating cumulative disruptive interference in memory for melodies, words, and pictures. *New Ideas in Psychology*, *55*, 68–77. https://doi.org/10.1016/j.newideapsych.2019.04.004

Herff, S. A., Olsen, K. N., & Dean, R. T. (2018). Resilient memory for melodies: The number of intervening melodies does not influence novel melody recognition. *Quarterly Journal of Experimental Psychology*, *71*(5), 1150–1171. https://doi.org/10.1080/17470218.2017.1318932

Herff, S. A., Olsen, K. N., Dean, R. T., & Prince, J. (2018). Memory for melodies in unfamiliar tuning systems: Investigating effects of recency and number of intervening items. *Quarterly Journal of Experimental Psychology*, *71*(6), 1367–1381. https://doi.org/10.1177/1747021817734978

Herff, S. A., Olsen, K. N., Prince, J., & Dean, R. T. (2018). Interference in memory for pitch-only and rhythm-only sequences: *Musicae Scientiae*. https://doi.org/10.1177/1029864917695654

Hickok, G. (1993). Parallel parsing: Evidence from reactivation in garden-path sentences. *Journal of Psycholinguistic Research*, *22*(2), 239–250. https://doi.org/10.1007/BF01067832

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434. https://doi.org/10.1016/j.tics.2007.09.004

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Hoch, L., Poulin-Charronnat, B., & Tillmann, B. (2011). The Influence of Task-Irrelevant Music on Language Processing: Syntactic and Semantic Structures. *Frontiers in Psychology, 2*, 112. https://doi.org/10.3389/fpsyg.2011.00112

Hoerl, C. (2013). Husserl, the Absolute Flow, and Temporal Experience. *Philosophy and Phenomenological Research*, *86*(2), 376–411.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.

Hollingworth, H. L. (1910). The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469. https://doi.org/10.2307/2012819

Holmes, V. M., & Forster, K. I. (1970). Detection of extraneous signals during sentence recognition. *Perception & Psychophysics*, *7*(5), 297–301. https://doi.org/10.3758/BF03210171

Holmes, V. M., & Forster, K. I. (1972). Click location and syntactic structure. *Perception & Psychophysics*, *12*(1), 9–15. https://doi.org/10.3758/BF03212836

Honing, H. (2008). Structure and interpretation of rhythm in music. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford Handbook of Music Psychology* (p. 36). Oxford University Press.

Huron, D. (2006). *Sweet anticipation: music and the psychology of expectation*. MIT Press.

Huron, D., & Ommen, A. (2006). An Empirical Study of Syncopation in American Popular Music, 1890–1939. *Music Theory Spectrum*, *28*(2), 211–231. https://doi.org/10.1525/mts.2006.28.2.211

Huron, D., & Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology: A Journal of Research in Music Cognition*, *12*(2), 154–171. https://doi.org/10.1037/h0094110

Husserl, E. (1964). *Phenomenology of Internal Time Consciousness*. Indiana University Press.

Ishida, K., & Nittono, H. (2022). Relationship between early neural responses to syntactic and acoustic irregularities in music. *European Journal of Neuroscience*, *56*(12), 6201–6214. https://doi.org/10.1111/ejn.15856

Jackendoff, R. (1987). *Consciousness and the Computational Mind* (1st edition). MIT Press.

Jackendoff, R. (1991). Musical Parsing and Musical Affect. *Music Perception: An Interdisciplinary Journal*, *9*(2), 199–229. https://doi.org/10.2307/40285529

Jackendoff, R. (2002a). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jackendoff, R. (2002b). How the competence model can constrain theories of processing. In *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jackendoff, R. (2007). *Language, consciousness, culture: essays on mental structure*. MIT Press OCLC: ocm71352682.

Jackendoff, R. (2009). Parallels and Nonparallels between Language and Music. *Music Perception*, *26*(3), 10.

Jackendoff, R., & Lerdahl, F. (2006). The capacity for music: What is it, and what's special about it?, 40.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *WIREs Cognitive Science*, *2*(1), 8–21. https://doi.org/10.1002/wcs.80

Jacoby, N., Tishby, N., & Tymoczko, D. (2015). An Information Theoretic Approach to Chord Categorization and Functional Harmony. *Journal of New Music Research*, *44*(3), 219–244. https://doi.org/10.1080/09298215.2015.1036888

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339. https://doi.org/10.1016/j.jml.2017.01.004

Jakubowski, K., Finkel, S., Stewart, L., & Müllensiefen, D. (2017). Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 122–135. https://doi.org/10.1037/aca0000090

Janata, P., Birk, J. L., Horn, J. D. V., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The Cortical Topography of Tonal Structures Underlying Western Music. *Science*, *298*(5601), 2167–2170. https://doi.org/10.1126/science.1076262

Jonaitis, E. M., & Saffran, J. R. (2009). Learning Harmony: The Role of Serial Statistics. *Cognitive Science*, *33*(5), 951–968. https://doi.org/10.1111/j.1551-6709.2009.01036.x

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, *13*(4), 313–319. https://doi.org/10.1111/1467-9280.00458

Jung, H., Sontag, S., Park, Y. S., & Loui, P. (2015). Rhythmic Effects of Syntax Processing in Music and Language. *Frontiers in Psychology*, *6*, 1762. https://doi.org/10.3389/fpsyg.2015.01762

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing* (2nd edition). Prentice Hall.

Kaan, E., & Swaab, T. Y. (2003). Repair, Revision, and Complexity in Syntactic Analysis: An Electrophysiological Differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Kaschak, M. P. (2006). What this construction needs is generalized. *Memory & Cognition*, *34*(2), 368–379. https://doi.org/10.3758/BF03193414

Katz, J., Chemla, E., & Pallier, C. (2015). An Attentional Effect of Musical Metrical Structure. *PloS One*, *10*(11). https://doi.org/10.1371/journal.pone.0140895

Katz, J., & Pesetsky, D. (2011). The identity thesis for language and music. http://ling.auf.net/lingbuzz/000959.

Keiler, A. (1978). Bernstein's "The Unanswered Question" and the Problem of Musical Competence. *The Musical Quarterly*, *64*(2), 195–222.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kirlin, P. B. (2014). A Data Set for Computational Studies of Schenkerian Analysis. *Proceedings of the 15th International Society for Music Information Retrieval Conference.*

Kirlin, P. B. (2008). A Framework for Automated Schenkerian Analysis. *Proceedings of the 9th International Society for Music Information Retrieval Conference.*

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian Inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511984037

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443–15448. https://doi.org/10.1073/pnas.1300272110

Koelsch, S. (2006). Significance of Broca's Area and Ventral Premotor Cortex for Music-Syntactic Processing. *Cortex*, *42*(4), 518–520. https://doi.org/10.1016/S0010-9452(08)70390-3

Koelsch, S. (2011). Toward a Neural Basis of Music Perception – A Review and Updated Model. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00110

Koelsch, S. (2013). *Brain and Music*. John Wiley & Sons.

Koelsch, S., Busch, T., Jentschke, S., & Rohrmeier, M. (2016). Under the hood of statistical learning: A statistical MMN reflects the magnitude of transitional probabilities in auditory sequences. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep19741

Koelsch, S., Gunter, T., Schröger, E., & Friederici, A. D. (2003). Processing Tonal Modulations: An ERP Study. *Journal of Cognitive Neuroscience*, *15*(8), 1149–1159. https://doi.org/10.1162/089892903322598111

Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between Syntax Processing in Language and in Music: An ERP Study. *Journal of Cognitive Neuroscience*, *17*(10), 1565–1577. https://doi.org/10.1162/089892905774597290

Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain Indices of Music Processing: "Nonmusicians" are Musical. *Journal of Cognitive Neuroscience*, *12*(3), 520–541. https://doi.org/10.1162/089892900562183

Koelsch, S., Jentschke, S., Sammler, D., & Mietchen, D. (2007). Untangling syntactic and sensory processing: An ERP study of music perception. *Psychophysiology*, *44*(3), 476–490. https://doi.org/10.1111/j.1469-8986.2007.00517.x

Koelsch, S., Maess, B., & Friederici, A. D. (2000). Musical syntax is processed in the area of Broca: an MEG study. *NeuroImage*, *11*(5), S56. https://doi.org/10.1016/S1053-8119(00)90990-X

Koelsch, S., Schroger, E., & Gunter, T. C. (2002). Music matters: Preattentive musicality of the human brain. *Psychophysiology*, *39*(1), 38–48. https://doi.org/10.1111/1469-8986.3910038

Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive Processes and the Peculiar Case of Music. *Trends in Cognitive Sciences*, *23*(1), 63–77. https://doi.org/10.1016/j.tics.2018.10.006

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368.

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, *11*(3), 346–374. https://doi.org/10.1016/0010-0285(79)90016-1

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford University Press.

Krumhansl, C. L. (1996). A Perceptual Analysis of Mozart's Piano Sonata K. 282: Segmentation, Tension, and Musical Ideas. *Music Perception: An Interdisciplinary Journal*, *13*(3), 401–432. https://doi.org/10.2307/40286177

Krumhansl, C. L. (1998). Perceived Triad Distance: Evidence Supporting the Psychological Reality of Neo-Riemannian Transformations. *Journal of Music Theory*, *42*(2), 265–281. https://doi.org/10.2307/843878

Krumhansl, C. L. (2002). Music: A Link Between Cognition and Emotion. *Current Directions in Psychological Science*, *11*(2), 45–50. https://doi.org/10.1111/1467-8721.00165

Kunert, R., & Slevc, L. R. (2015). A Commentary on: "Neural overlap in processing music and speech". *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00330

Ladefoged, P., & Broadbent, D. E. (1960). Perception of Sequence in Auditory Events. *Quarterly Journal of Experimental Psychology*, *12*(3), 162–170. https://doi.org/10.1080/17470216008416720

Ladinig, O., Honing, H., Háden, G., & Winkler, I. (2009). Probing Attentive and Preattentive Emergent Meter in Adult Listeners without Extensive Music Training. *Music Perception*, *26*(4), 377–386. https://doi.org/10.1525/mp.2009.26.4.377

Laneve, S., Schaerf, L., Cecchetti, G., Hentschel, J., & Rohrmeier, M. (2023). The diachronic development of Debussy's musical style: a corpus study with Discrete Fourier Transform. *Humanities and Social Sciences Communications*, *10*(1), 1–13. https://doi.org/10.1057/s41599-023-01796-7

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*(1), 119–159. https://doi.org/10.1037/0033-295X.106.1.119

Large, E. W., Kim, J. C., Flaig, N. K., Bharucha, J. J., & Krumhansl, C. L. (2016). A Neurodynamic Account of Musical Tonality. *Music Perception: An Interdisciplinary Journal*, *33*(3), 319–331. https://doi.org/10.1525/mp.2016.33.3.319

Large, E. W., & Snyder, J. S. (2009). Pulse and Meter as Neural Resonance. *Annals of the New York Academy of Sciences*, *1169*(1), 46–57. https://doi.org/10.1111/j.1749-6632.2009.04550.x

Lartillot, O. (2011). Cultural impact in listeners' structural understanding of a Tunisian traditional modal improvisation, studied with the help of computational models. *JOurnal of interdisciplinary music studies*, (5-1). https://doi.org/10.4407/jims.2011.07.005

Lartillot, O., & Ayari, M. (2009). Segmentation of Tunisian Modal Improvisation: Comparing Listeners' Responses with Computational Predictions. *Journal of New Music Research*, *38*(2), 117–127. https://doi.org/10.1080/09298210903194071

Lehne, M., & Koelsch, S. (2015). Tension–resolution patterns as a key element of aesthetic experience: Psychological principles and underlying brain mechanisms. In J. P. Huston, M. Nadal, F. Mora, L. F. Agnati, & C. J. C. Conde (Eds.), *Art, Aesthetics, and the Brain* (pp. 285–302). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199670000.003.0014

Lehne, M., Rohrmeier, M., Gollmann, D., & Koelsch, S. (2013). The Influence of Different Structural Features on Felt Musical Tension in Two Piano Pieces by Mozart and

Mendelssohn. *Music Perception: An Interdisciplinary Journal*, *31*(2), 171–185. https://doi.org/10.1525/mp.2013.31.2.171

Leino, S., Brattico, E., Tervaniemi, M., & Vuust, P. (2007). Representation of harmony rules in the human brain: Further evidence from event-related potentials. *Brain Research*, *1142*, 169–177. https://doi.org/10.1016/j.brainres.2007.01.049

Lendvai, E. (1971). *Bela Bartok: An Analysis of His Music*. Kahn & Averill Publishers.

Lerdahl, F. (2001). *Tonal Pitch Space*. Oxford University Press.

Lerdahl, F. (2014). Tension and Expectation in a Schubert Song. In L. F. Bernstein & A. Rozin (Eds.), *Musical Implications: Essays in Honor of Eugene Narmour* (pp. 255–274). Pendragon Press.

Lerdahl, F. (2015). Concepts and Representations of Musical Hierarchies. *Music Perception: An Interdisciplinary Journal*, *33*(1), 83–95. https://doi.org/10.1525/mp.2015.33.1.83

Lerdahl, F. (2019). *Composition and Cognition: Reflections on Contemporary Music and the Musical Mind*. University of California Press.

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

Lerdahl, F., & Jackendoff, R. (1983b). An Overview of Hierarchical Structure in Music. *Music Perception: An Interdisciplinary Journal*, *1*(2), 229–252.

Lerdahl, F., & Krumhansl, C. L. (2007). Modeling Tonal Tension. *Music Perception: An Interdisciplinary Journal*, *24*(4). https://doi.org/10.1525/mp.2007.24.4.329

Lester, J. (1979). Articulation of Tonal Structures as a Criterion for Analytic Choices. *Music Theory Spectrum*, *1*, 67–79. https://doi.org/10.2307/745779

Lester, J. (1982). *Harmony in Tonal Music* (1st edition). McGraw-Hill College.

Lester, J. (1995). Performance and analysis: interaction and interpretation. In J. Rink (Ed.), *The Practice of Performance* (pp. 197–216). Cambridge University Press. https://doi.org/10.1017/CBO9780511552366.010

Levine, M. (1989). *The Jazz Piano Book*. Sher Music.

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

Levinson, J. (1980). What a Musical Work Is. *The Journal of Philosophy*, *77*(1), 5. https://doi.org/10.2307/2025596

Levinson, J. (1997). *Music in the Moment*. Cornell University Press.

Levitin, D. J., Grahn, J. A., & London, J. (2018). The Psychology of Music: Rhythm and Movement. *Annual Review of Psychology*, *69*(1), 51–75. https://doi.org/10.1146/annurev-psych-122216-011740

Levitin, D. J., & Menon, V. (2003). Musical structure is processed in “language” areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. *NeuroImage*, *20*(4), 2142–2152. https://doi.org/10.1016/j.neuroimage.2003.08.016

Lewin, D. (1986). Music Theory, Phenomenology, and Modes of Perception. *Music Perception: An Interdisciplinary Journal*, *3*(4), 327–392. https://doi.org/10.2307/40285344

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*(1), 93–115. https://doi.org/10.1007/BF01708421

Lewis, R. L. (2000). Falsifying Serial and Parallel Parsing Models: Empirical Conundrums and An Overlooked Paradigm. *Journal of Psycholinguistic Research*, *29*(2), 241–248. https://doi.org/10.1023/A:1005105414238

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000_25

London, J. (2012). Three Things Linguists Need to Know About Rhythm and Time in Music. *Empirical Musicology Review*, *7*(1-2), 5–11. https://doi.org/10.18061/1811/52973

London, J., Himberg, T., & Cross, I. (2009). The Effect of Structural and Performance Factors in the Perception of Anacruses. *Music Perception*, *27*(2), 103–120. https://doi.org/10.1525/mp.2009.27.2.103

Longuet-Higgins, H. C. (1978). The Perception of Music. *Interdisciplinary Science Reviews*, *3*(2), 148–156. https://doi.org/10.1179/030801878791926065

Longuet-Higgins, H. C., & Lee, C. S. (1982). The Perception of Musical Rhythms. *Perception*, *11*(2), 115–128. https://doi.org/10.1068/p110115

Loui, P. (2012). Learning and Liking of Melody and Harmony: Further Studies in Artificial Grammar Learning. *Topics in Cognitive Science*, *4*(4), 554–567. https://doi.org/10.1111/j.1756-8765.2012.01208.x

Loui, P., & Wessel, D. (2008). Learning and Liking an Artificial Musical System: Effects of Set Size and Repeated Exposure. *Musicae Scientiae*, *12*(2), 207–230. https://doi.org/10.1177/102986490801200202

Loui, P., Wessel, D. L., & Kam, C. L. H. (2010). Humans Rapidly Learn Grammatical Structure in a New Musical Scale. *Music Perception*, *27*(5), 377–388. https://doi.org/10.1525/mp.2010.27.5.377

Ma, X., Ding, N., Tao, Y., & Yang, Y. F. (2018a). Differences in Neurocognitive Mechanisms Underlying the Processing of Center-Embedded and Non–embedded Musical Structures. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00425

Ma, X., Ding, N., Tao, Y., & Yang, Y. F. (2018b). Syntactic complexity and musical proficiency modulate neural processing of non-native music. *Neuropsychologia*, *121*, 164–174. https://doi.org/10.1016/j.neuropsychologia.2018.10.005

Ma, X., Tao, Y., & Yang, Y. (2022). Factors inducing complexities in musical embedded structure processing. *Neuropsychologia*, *169*, 108153. https://doi.org/10.1016/j.neuropsychologia.2022.108153

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037//0033-295X.101.4.676

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

MacRitchie, J., Herff, S. A., Procopio, A., & Keller, P. E. (2018). Negotiating between individual and joint goals in ensemble musical performance. *Quarterly Journal of Experimental Psychology*, *71*(7), 1535–1551. https://doi.org/10.1080/17470218.2017.1339098

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: an MEG study. *Nature Neuroscience*, *4*(5), 540–545. https://doi.org/10.1038/87502

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Margulis, E. H. (2005). A Model of Melodic Expectation. *Music Perception*, *22*(4), 663–714. https://doi.org/10.1525/mp.2005.22.4.663

Margulis, E. H. (2007). Surprise and Listening Ahead: Analytic Engagements with Musical Tendencies. *Music Theory Spectrum*, *29*(2), 197–217. https://doi.org/10.1525/mts.2007.29.2.197

Marmel, F., Tillmann, B., & Dowling, W. J. (2008). Tonal expectations influence pitch perception. *Perception & Psychophysics*, *70*(5), 841–852. https://doi.org/10.3758/PP.70.5.841

Marmel, F., Perrin, F., & Tillmann, B. (2011). Tonal Expectations Influence Early Pitch Processing. *Journal of Cognitive Neuroscience*, *23*(10), 3095–3104. https://doi.org/10.1162/jocn.2011.21632

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Marslen-Wilson, W. (1973). Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature*, *244*(5417), 522–523. https://doi.org/10.1038/244522a0

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71. https://doi.org/10.1016/0010-0277(80)90015-3

Martin, H. (2023). On the Tonic Added-Sixth Chord in Jazz. *Music Theory Online*, *29*(2).

Martin, N. J., & Vande Moortele, S. (2014). Formal Functions and Retrospective Reinterpretation in the First Movement of Schubert's String Quintet: Formal Functions and Retrospective Reinterpretation. *Music Analysis*, *33*(2), 130–155. https://doi.org/10.1111/musa.12025

Martínez, I. C. (2018). Music attending to linear constituent structures in tonal music. *Music & Science*, *1*, 2059204318787763. https://doi.org/10.1177/2059204318787763

Martins, M., Bianco, R., Sammler, D., & Villringer, A. (2017). Cognitive and Neural Mechanisms Underlying the Generation of Motor Hierarchies, 2.

Mathias, B., Zamm, A., Gianferrara, P. G., Ross, B., & Palmer, C. (2020). Rhythm Complexity Modulates Behavioral and Neural Dynamics During Auditory–Motor Synchronization. *Journal of Cognitive Neuroscience*, *32*(10), 1864–1880. https://doi.org/10.1162/jocn_a_01601

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356. https://doi.org/10.1016/j.tics.2010.06.002

McClelland, J. L., & Cleeremans, A. (2009). Connectionist Models. In B. Tim, C. Axel, & W. Patrick (Eds.), *The Oxford Companion to Consciousness.* Oxford University Press.

McClelland, R. (2006). Extended Upbeats in the Classical Minuet: Interactions with Hypermeter and Phrase Structure. *Music Theory Spectrum*, *28*(1), 23–56. https://doi.org/10.1525/mts.2006.28.1.23

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, *535*(7613), 547–550. https://doi.org/10.1038/nature18635

McFee, B., Nieto, O., Farbood, M. M., & Bello, J. P. (2017). Evaluating Hierarchical Structure in Music Annotations. *Frontiers in Psychology*, *8*. Retrieved February 5, 2024, from https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.01337

McGowan, J. (2010). Riemann's Functional Framework for Extended Jazz Harmony. *Intégral*, *24*, 115–133.

McGowan, J. (2011). Psychoacoustic Foundations of Contextual Harmonic Stability in Jazz Piano Voicings. *Journal of Jazz Studies*, *7*(2), 156. https://doi.org/10.14713/jjs.v7i2.13

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, *38*(3), 283–312. https://doi.org/10.1006/jmla.1997.2543

Meeus, N. (2000). Toward a Post-Schoenbergian Grammar of Tonal and Pre-tonal Harmonic Progressions. *Music Theory Online*, *6*(1).

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Meltzer, J. A., & Braun, A. R. (2011). An EEG–MEG Dissociation between Online Syntactic Comprehension and Post Hoc Reanalysis. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00010

Merkey, P., & Posner, E. (1984). Optimum cyclic redundancy codes for noisy channels. *IEEE Transactions on Information Theory*, *30*(6), 865–867. https://doi.org/10.1109/TIT.1984.1056971

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, *30*(4), 551–561. https://doi.org/10.3758/BF03194956

Meyer, L. B. (1956). *Emotion and Meaning in Music.* The University of Chicago Press.

Meyer, L. B. (1957). Meaning in Music and Information Theory. *The Journal of Aesthetics and Art Criticism*, *15*(4), 412. https://doi.org/10.2307/427154

Miller, I. (1984). *Husserl, perception, and temporal awareness.* Cambridge, Mass. : MIT Press.

Milne, A. J., & Herff, S. A. (2020). The perceptual relevance of balance, evenness, and entropy in musical rhythms. *Cognition*, *203*, 104233. https://doi.org/10.1016/j.cognition.2020.104233

Milne, A. J., & Holland, S. (2016). Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, *10*(1), 59–85. https://doi.org/10.1080/17459737.2016.1152517

Milne, A. J., Laney, R., & Sharp, D. B. (2015). A Spectral Pitch Class Model of the Probe Tone Data and Scalic Tonality. *Music Perception: An Interdisciplinary Journal*, *32*(4), 364–393.

Milne, A. J., Laney, R., & Sharp, D. B. (2016). Testing a spectral model of tonal affinity with microtonal melodies and inharmonic spectra. *Musicae Scientiae*, *20*(4), 465–494. https://doi.org/10.1177/1029864915622682

Milne, A. J., Smit, E. A., Sarvasy, H. S., & Dean, R. T. (2023). Evidence for a universal association of auditory roughness with musical stability. *PloS One*, *18*(9), e0291642. https://doi.org/10.1371/journal.pone.0291642

Mirka, D. (2009). *Metric Manipulations in Haydn and Mozart: Chamber Music for Strings, 1787-1791*. Oxford University Press.

Moore, A. (1995). The so-called 'flattened seventh' in rock. *Popular Music*, *14*(2), 185–201. https://doi.org/10.1017/S0261143000007431

Moore, B. C. J., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1591), 919–931. https://doi.org/10.1098/rstb.2011.0355

Morgan, R. P. (1978). The Theory and Analysis of Tonal Rhythm. *The Musical Quarterly*, *64*(4), 435–473.

Moshaver, M. A. (2012). Telos and Temporality: Phenomenology and the Experience of Time in Lewin's Study of Perception. *Journal of the American Musicological Society*, *65*(1), 179–214. https://doi.org/10.1525/jams.2012.65.1.179

Moss, F. C., Neuwirth, M., Harasim, D., & Rohrmeier, M. (2019). Statistical characteristics of tonal harmony: A corpus study of Beethoven's string quartets (C. M. Pinto, Ed.). *PLOS ONE*, *14*(6), e0217242. https://doi.org/10.1371/journal.pone.0217242

Moss, F. C., Neuwirth, M., & Rohrmeier, M. (2022). The line of fifths and the co-evolution of tonal pitch-classes. *Journal of Mathematics and Music*, 1–25. https://doi.org/10.1080/17459737.2022.2044927

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population (J. Snyder, Ed.). *PLoS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Narayanan, S., & Jurafsky, D. (1998). A Bayesian Model of Human Sentence Processing. *Proceedings of the Twelfth Annual Meeting of the Cognitive Science Society*.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. University of Chicago Press.

Nettl, B. (1999). An Ethnomusicologist Contemplates Universals in Musical Sound and Musical Culture. https://doi.org/10.7551/mitpress/5190.003.0032

Neuwirth, M. (2015). Fuggir la Cadenza, or The Art of Avoiding Cadential Closure. In M. Neuwirth & P. Bergé (Eds.), *What Is a Cadence* (pp. 117–156). Leuven University Press.

Nicol, J. L., & Pickering, M. J. (1993). Processing syntactically ambiguous sentences: Evidence from semantic priming. *Journal of Psycholinguistic Research*, *22*(2), 207–237. https://doi.org/10.1007/BF01067831

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin*, *89*(1), 133–162. https://doi.org/10.1037/0033-2909.89.1.133

Novembre, G., & Keller, P. E. (2011). A grammar of action generates predictions in skilled musicians. *Consciousness and Cognition*, *20*(4), 1232–1243. https://doi.org/10.1016/j.concog.2011.03.009

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective Neuronal Entrainment to the Beat and Meter Embedded in a Musical Rhythm. *Journal of Neuroscience*, *32*(49), 17572–17581. https://doi.org/10.1523/JNEUROSCI.3203-12.2012

O'Brien, G., & Opie, J. (2004). Chapter 1 - Notes Toward a Structuralist Theory of Mental Representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in Mind* (pp. 1–20). Elsevier. https://doi.org/10.1016/B978-008044394-2/50004-X

Ogg, M., Okada, B. M., Novick, J. M., & Slevc, L. R. (2019). Updating Musical Tonal Structure in Working Memory Engages Cognitive Control. *Auditory Perception & Cognition*, *2*(1-2), 21–46. https://doi.org/10.1080/25742442.2019.1626686

Online, G. M. (2001). Beat (i).

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain Potentials Elicited by Garden-Path Sentences: Evidence of the Application of Verb Information During Parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786–803.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241–263. https://doi.org/10.1037/h0027272

Parncutt, R. (1988). Revision of Terhardt's Psychoacoustical Model of the Root(s) of a Musical Chord. *Music Perception: An Interdisciplinary Journal*, *6*(1), 65–93. https://doi.org/10.2307/40285416

Parncutt, R. (1994). A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms. *Music Perception*, *11*(4), 409–464. https://doi.org/10.2307/40285633

Parncutt, R. (2011). The Tonic as Triad: Key Profiles as Pitch Salience Profiles of Tonic Triads. *Music Perception: An Interdisciplinary Journal*, *28*(4), 333–366. https://doi.org/10.1525/mp.2011.28.4.333

Parncutt, R., & Hair, G. (2012). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, *5*(2). https://doi.org/10.4407/jims.2011.11.002

Patel, A. D. (1998). Syntactic Processing in Language and Music: Different Cognitive Operations, Similar Neural Resources? *Music Perception*, *16*(1), 27–42. https://doi.org/10.2307/40285775

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681. https://doi.org/10.1038/nn1082

Patel, A. D. (2006). Musical Rhythm, Linguistic Rhythm, and Human Evolution. *Music Perception*, *24*(1), 99–104. https://doi.org/10.1525/mp.2006.24.1.99

Patel, A. D. (2007). Language, music, and the brain: a resource-sharing framework. In P. Rebuschat, M. Rohrmeier, J. A. Hawkins, & I. Cross (Eds.), *Language and Music as Cognitive Systems* (pp. 204–223). Oxford University Press.

Patel, A. D. (2010). *Music, Language, and the Brain* (1st edition). Oxford University Press.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, *87*(1), B35–B45. https://doi.org/10.1016/S0010-0277(02)00187-7

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing Syntactic Relations in Language and Music: An Event-Related Potential Study. *Journal of Cognitive Neuroscience*, *10*(6), 717–733. https://doi.org/10.1162/089892998563121

Patel, A. D., & Morgan, E. (2017). Exploring Cognitive Relations Between Prediction in Language and Music. *Cognitive Science*, *41*(S2), 303–320. https://doi.org/10.1111/cogs.12411

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition* (Doctoral Thesis). City University. London.

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395. https://doi.org/10.1111/nyas.13654

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The Role of Expectation and Probabilistic Learning in Auditory Boundary Perception: A Model Comparison. *Perception*, *39*(10), 1367–1391. https://doi.org/10.1068/p6507

Pearce, M. T., & Rohrmeier, M. (2012). Music Cognition and the Cognitive Sciences. *Topics in Cognitive Science*, *4*(4), 468–484. https://doi.org/10.1111/j.1756-8765.2012.01226.x

Pearce, M. T., & Rohrmeier, M. (2018). Musical Syntax II: Empirical Perspectives. In R. Bader (Ed.), *Springer Handbook of Systematic Musicology* (pp. 487–505). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-55004-5_26

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302–313.

Pearce, M. T., & Wiggins, G. A. (2012). Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science*, *4*(4), 625–652. https://doi.org/10.1111/j.1756-8765.2012.01214.x

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Penel, A., & Jones, M. R. (2005). Speeded Detection of a Tone Embedded in a Quasi-isochronous Sequence: Effects of a Task-Irrelevant Temporal Irregularity. *Music Perception*, *22*(3), 371–388. https://doi.org/10.1525/mp.2005.22.3.371

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, *6*(7), 688–691. https://doi.org/10.1038/nn1083

Peretz, I., Vuvan, D., Lagrois, M.-É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140090. https://doi.org/10.1098/rstb.2014.0090

Peretz, I., & Zatorre, R. J. (2005). Brain Organization for Music Processing. *Annual Review of Psychology*, *56*(1), 89–114. https://doi.org/10.1146/annurev.psych.56.091103.070225

Pezzulo, G., Parr, T., & Friston, K. (2021). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200531. https://doi.org/10.1098/rstb.2020.0531

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz*. https://doi.org/lingbuzz/007180

Pickering, M. J., & Branigan, H. P. (1998). The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language*, *39*(4), 633–651. https://doi.org/10.1006/jmla.1998.2592

Pickering, M. J., & van Gompel, R. P. G. (2006). Syntactic Parsing. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics (Second Edition)* (pp. 455–503). Academic Press. https://doi.org/10.1016/B978-012369374-7/50013-4

Pinker, S., & Jackendoff, R. (2005). *The faculty of language: what's special about it?*

Pironio, N., Slezak, D. F., & Miguel, M. A. (2021). Pulse clarity metrics developed from a deep learning beat tracking model. *Proceedings of the 22nd International Society for Music Information Retrieval Conference*.

Piston, W. (1948). *Harmony* (M. DeVoto, Ed.; Fifth edition). W. W. Norton & Company.

Politimou, N., Douglass-Kirk, P., Pearce, M. T., Stewart, L., & Franco, F. (2021). Melodic expectations in 5- and 6-year-old children. *Journal of Experimental Child Psychology*, *203*, 105020. https://doi.org/10.1016/j.jecp.2020.105020

Polth, M. (2006). Tonalität der Tonfelder: Anmerkungen zu Bernhard Haas (2004), Die neue Tonalität von Schubert bis Webern. Hören und Analysieren nach Albert Simon, Wilhelmshaven: Noetzel. *Zeitschrift der Gesellschaft für Musiktheorie [Journal of the German-speaking Society of Music Theory]*, *3*(1), 167–178. https://doi.org/10.31751/210

Polth, M. (2011). Zur Artikulation von Tonfeldern bei Brahms, Debussy und Stockhausen. *Zeitschrift der Gesellschaft für Musiktheorie [Journal of the German-speaking Society of Music Theory]*, *8*(2), 225–265. https://doi.org/10.31751/645

Polth, M. (2018). The Individual Tone and Musical Context in Albert Simon's *Tonfeldtheorie*. *Music Theory Online*, *24*(4).

Popescu, T., Farrugia, N., Ruge, H., Boneh, O., Bravo, F., Tian, X., & Rohrmeier, M. (2022). *Neural representations of harmonic function in musical imagery* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/ry79k

Popescu, T., Widdess, R., & Rohrmeier, M. (2021). Western listeners detect boundary hierarchy in Indian music: a segmentation study. *Scientific Reports*, *11*(1), 3112. https://doi.org/10.1038/s41598-021-82629-y

Povel, D.-J., & Okkerman, H. (1981). Accents in equitone sequences. *Perception & Psychophysics*, *30*(6), 565–572. https://doi.org/10.3758/BF03202011

Pressnitzer, D., McAdams, S., Winsberg, S., & Fineberg, J. (2000). Perception of musical tension for nontonal orchestral timbres and its relation to psychoacoustic roughness. *Perception & Psychophysics*, *62*(1), 66–80. https://doi.org/10.3758/BF03212061

Prince, J. B. (2011). The integration of stimulus dimensions in the perception of music. *Quarterly Journal of Experimental Psychology*, *64*(11), 2125–2152.

Prince, J. B., Stevens, C. J., Jones, M. R., & Tillmann, B. (2018). Learning of pitch and time structures in an artificial grammar setting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1201–1214. https://doi.org/10.1037/xlm0000502

Prince, J. B., Thompson, W. F., & Schmuckler, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1598.

Prince, J. B., Vuvan, D. T., Schmuckler, M. A., & Scott-Clark, T. T. (2015). Tonal priming is resistant to changes in pitch height. *Attention, Perception, & Psychophysics*, *77*(6), 2011–2020. https://doi.org/10.3758/s13414-015-0904-7

Puchalla, J. L., Schneidman, E., Harris, R. A., & Berry, M. J. (2005). Redundancy in the Population Code of the Retina. *Neuron*, *46*(3), 493–504. https://doi.org/10.1016/j.neuron.2005.03.026

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235.

Rebuschat, P. (2022). Implicit learning and language acquisition: Three approaches, one phenomenon. In A. S. Reber & R. Allen (Eds.), *The cognitive unconscious*. Oxford University Press.

Ren, Z., Gerstner, W., & Rohrmeier, M. (2023). Music as flow: a formal representation of hierarchical processes in music. *Proceedings of the 24th International Society for Music Information Retrieval Conference*.

Resnik, P. (1992). Left-Corner Parsing and Psychological Plausibility. *Proceedings of the 14th Conference on Computational Linguistics*, 191–197. https://doi.org/https://dl.acm.org/doi/10.3115/992066.992098

Riemann, H. (1893). *Vereinfachte Harmonielehre oder die Lehre von den tonalen Funktionen der Akkorde*. Augner.

Rink, J. (2015). The (F)utility of Performance Analysis. In *Music in Profile: Twelve Performance Studies* (pp. 206–224). Oxford University Press.

Robinson, C. W., Moore, R. L., & Crook, T. A. (2018). Bimodal Presentation Speeds up Auditory Processing and Slows Down Visual Processing. *Frontiers in Psychology*, *9*.

Rogers, T. T., & McClelland, J. L. (2008). Précis of Semantic Cognition: A Parallel Distributed Processing Approach. *Behavioral and Brain Sciences*, *31*(6), 689–714. https://doi.org/10.1017/S0140525X0800589X

Rohrmeier, M. (2010). *Implicit learning of musical structure: experimental and computational modelling approaches* (Doctoral dissertation). University of Cambridge.

Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, *5*(1), 35–53. https://doi.org/10.1080/17459737.2011.573676

Rohrmeier, M. (2013). Musical Expectancy: Bridging Music Theory, Cognitive and Computational Approaches. *Zeitschrift der Gesellschaft für Musiktheorie [Journal of the German-speaking Society of Music Theory]*, *10*(2), 343–371. https://doi.org/10.31751/724

Rohrmeier, M. (2020a). Towards a formalization of musical rhythm. *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 621–629. https://doi.org/10.5281/zenodo.4245508

Rohrmeier, M. (2020b). The Syntax of Jazz Harmony: Diatonic Tonality, Phrase Structure, and Form. *Music Theory and Analysis (MTA)*, *7*(1), 1–63. https://doi.org/10.11116/MTA.7.1.1

Rohrmeier, M., & Cross, I. (2008). Statistical Properties of Tonal Harmony in Bach's Chorales. *Proceedings of the 10th International Conference on Music Perception and Cognition*, 619–627.

Rohrmeier, M., & Cross, I. (2009). Tacit tonality: Implicit learning of context-free harmonic structure. *ESCOM 2009 : 7th Triennial Conference of European Society for the Cognitive Sciences of Music.*

Rohrmeier, M., & Cross, I. (2013). Artificial Grammar Learning of Melody Is Constrained by Melodic Inconsistency: Narmour's Principles Affect Melodic Learning. *PLOS ONE*, *8*(7), e66174. https://doi.org/10.1371/journal.pone.0066174

Rohrmeier, M., Dienes, Z., Guo, X., & Fu, Q. (2014). Implicit learning and recursion. In F. Lowenthal & L. Lefebvre (Eds.), *Language and recursion* (pp. 67–85). Springer. https://doi.org/10.1007/978-1-4614-9414-0_6

Rohrmeier, M., Fu, Q., & Dienes, Z. (2012). Implicit Learning of Recursive Context-Free Grammars. *PLOS ONE*, *7*(10), e45885. https://doi.org/10.1371/journal.pone.0045885

Rohrmeier, M., & Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. *International Journal of Psychophysiology*, *83*(2), 164–175. https://doi.org/10.1016/j.ijpsycho.2011.12.010

Rohrmeier, M., & Moss, F. C. (2021). A formal model of extended-tonal harmony. *Proceedings of the 22nd International Society for Music Information Retrieval Conference.*

Rohrmeier, M., & Neuwirth, M. (2015). Towards a Syntax of the Classical Cadence. In P. Bergé & M. Neuwirth (Eds.), *What is a cadence? theoretical and analytical perspectives on cadences in the classical repertoire*. Leuven University Press.

Rohrmeier, M., & Pearce, M. T. (2018). Musical Syntax I: Theoretical Perspectives. In R. Bader (Ed.), *Springer Handbook of Systematic Musicology* (pp. 473–486). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-55004-5_25

Rohrmeier, M., & Rebuschat, P. (2012). Implicit Learning and Acquisition of Music. *Topics in Cognitive Science*, *4*(4), 525–553. https://doi.org/10.1111/j.1756-8765.2012.01223.x

Rohrmeier, M., Rebuschat, P., & Cross, I. (2011). Incidental and online learning of melodic structure. *Consciousness and Cognition*, *20*(2), 214–222. https://doi.org/10.1016/j.concog.2010.07.004

Rohrmeier, M., & Widdess, R. (2017). Incidental Learning of Melodic Structure of North Indian Music. *Cognitive Science*, *41*(5), 1299–1327. https://doi.org/10.1111/cogs.12404

Ross, B. (2014). *Music, language and syntactic integration* (Doctoral dissertation). University of Cambridge. Cambridge.

Rothstein, W. N. (1989). *Phrase Rhythm in Tonal Music.* Schirmer.

Ruiz, M. H., Koelsch, S., & Bhattacharya, J. (2009). Decrease in early right alpha band phase synchronization and late gamma band oscillations in processing syntax in music. *Human Brain Mapping*, *30*(4), 1207–1225. https://doi.org/10.1002/hbm.20584

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527. https://doi.org/10.1016/j.cogpsych.2022.101527

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., & McIntosh, A. R. (2015a). Predictions and the brain: how musical sounds become rewarding. *Trends in Cognitive Sciences*, *19*(2), 86–91. https://doi.org/10.1016/j.tics.2014.12.001

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., & McIntosh, A. R. (2015b). Predictions and the brain: how musical sounds become rewarding. *Trends in Cognitive Sciences*, *19*(2), 86–91. https://doi.org/10.1016/j.tics.2014.12.001

Salzer, F. (1962). *Structural Hearing: Tonal Coherence in Music* (Illustrated edition). Dover Publications.

Sammler, D., Novembre, G., Koelsch, S., & Keller, P. E. (2013). Syntax in a pianist's hand: ERP signatures of "embodied" syntax processing in music. *Cortex*, *49*(5), 1325–1339. https://doi.org/10.1016/j.cortex.2012.06.007

Sanguinetti, G. (2012). *The art of partimento: history, theory, and practice*. Oxford University Press.

Sauvé, S. A. (2018). *Prediction in polyphony: modelling musical auditory scene analysis* (Thesis). Queen Mary University of London
Accepted: 2018-10-18T11:43:39Z.

Sauvé, S. A., Sayed, A., Dean, R. T., & Pearce, M. T. (2018). Effects of pitch and timing expectancy on musical emotion. *Psychomusicology: Music, Mind, and Brain*, *28*(1), 17–39. https://doi.org/10.1037/pmu0000203

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, *44*. https://doi.org/10.1017/S0140525X20000333

Savitch, W. J., Bach, E., Marsh, W. E., & Safran-Naveh, G. (Eds.). (1987). *The Formal Complexity of Natural Language*. D. Reidel Publishing Company.

Schellenberg, E. G., & Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Memory & Cognition*, *43*(7), 1021–1031. https://doi.org/10.3758/s13421-015-0519-1

Schenker, H. (1929). Letter to Moritz Violin.

Schenker, H. (1935). *Der Freie Satz*. Universal Edition.

Schenker, H. (1987). *Counterpoint*. New York : Schirmer Books ; London : Collier Macmillan.

Schenker, H., & Esser, H. (2000). *The art of performance*. Oxford University Press.

Schiltknecht, D. (2011). ›Konstrukt‹ und ›Funktion‹: Eine Herleitung der Simonschen Tonfelder. *Zeitschrift der Gesellschaft für Musiktheorie [Journal of the German-speaking Society of Music Theory]*, *8*(2). https://doi.org/10.31751/632

Schmalfeldt, J. (1992). Cadential processes: The evaded cadence and the “one more time” technique. *Journal of Musicological Research, 12*(1-2), 1–52. https://doi.org/10.1080/01411899208574658

Schmalfeldt, J. (2017). *In the Process of Becoming: Analytic and Philosophical Perspectives on Form in Early Nineteenth-century Music.* Oxford University Press.

Schuiling, F. (2022). Music As Extended Agency: On Notation And Entextualization IN Improvised Music. *Music and Letters, 103*(2), 322–343. https://doi.org/10.1093/ml/gcab109

Sears, D. R. W., Caplin, W. E., & McAdams, S. (2014). Perceiving the Classical Cadence. *Music Perception, 31*(5), 397–417. https://doi.org/10.1525/mp.2014.31.5.397

Sears, D. R. W., Pearce, M. T., Caplin, W. E., & McAdams, S. (2018). Simulating melodic and harmonic expectations for tonal cadences using probabilistic models. *Journal of New Music Research, 47*(1), 29–52. https://doi.org/10.1080/09298215.2017.1367010

Sears, D. R. W., Spitzer, J., Caplin, W. E., & McAdams, S. (2018). Expecting the end: Continuous expectancy ratings for tonal cadences. *Psychology of Music,* 0305735618803676. https://doi.org/10.1177/0305735618803676

Sears, D. R. W., Spitzer, J., Caplin, W. E., & McAdams, S. (2020). Expecting the end: Continuous expectancy ratings for tonal cadences. *Psychology of Music, 48*(3), 358–375. https://doi.org/10.1177/0305735618803676

Sears, D. R. W., Verbeten, J., & Percival, H. M. (2021). Intonation discrimination for tonal chord sequences in a priming paradigm:Effects of target predictability and musical expertise. *Auditory Perception & Cognition, 0*(0), 1–16. https://doi.org/10.1080/25742442.2021.1972744

Sears, D. R. W., Verbeten, J. E., & Percival, H. M. (2023). Does order matter? Harmonic priming effects for scrambled tonal chord sequences. *Journal of Experimental Psychology: Human Perception and Performance,* No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/xhp0001103

Sears, D. R., Pearce, M. T., Spitzer, J., Caplin, W. E., & McAdams, S. (2019). Expectations for tonal cadences: Sensory and cognitive priming effects. *Quarterly Journal of Experimental Psychology, 72*(6), 1422–1438. https://doi.org/10.1177/1747021818814472

Serafine, M. L., Glassman, N., & Overbeeke, C. (1989). The Cognitive Reality of Hierarchic Structure in Music. *Music Perception, 6*(4), 397–430. https://doi.org/10.2307/40285440

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust Effects of Working Memory Demand during Naturalistic Language Comprehension in Language-Selective Cortex. *Journal of Neuroscience, 42*(39), 7412–7430. https://doi.org/10.1523/JNEUROSCI.1894-21.2022

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC),* 49–58.

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

Shea, N. (2014). Exploitable Isomorphism and Structural Representation. *Proceedings of the Aristotelian Society, 114,* 123–144.

Shelemay, K. K. (2008). Notation and Oral Tradition. In *The Garland Handbook of African Music* (2nd ed.). Routledge.

Shepard, R. N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology, 62*(3), 302–309. https://doi.org/10.1037/h0048606

Shi, L., & Griffiths, T. (2009). Neural Implementation of Hierarchical Bayesian Inference by Importance Sampling. *Advances in Neural Information Processing Systems, 22.*

Shtyrov, Y., Kujala, T., & Pulvermüller, F. (2010). Interactions between Language and Attention Systems: Early Automatic Lexical Processing? *Journal of Cognitive Neuroscience, 22*(7), 1465–1478. https://doi.org/10.1162/jocn.2009.21292

Sioros, G., Davies, M. E. P., & Guedes, C. (2018). A generative model for the characterization of musical rhythms. *Journal of New Music Research, 47*(2), 114–128. https://doi.org/10.1080/09298215.2017.1409769

Sioros, G., Miron, M., Davies, M., Gouyon, F., & Madison, G. (2014). Syncopation creates the sensation of groove in synthesized music examples. *Frontiers in Psychology, 5.*

Sipser, M. (2012). *Introduction to the Theory of Computation* (3 edition). Cengage Learning.

Slevc, L. R., & Okada, B. M. (2015). Processing structure in language and music: a case for shared reliance on cognitive control. *Psychonomic Bulletin & Review, 22*(3), 637–652. https://doi.org/10.3758/s13423-014-0712-4

Slevc, L. R., Reitman, J. G., & Okada, B. M. (2013). Syntax in music and language: The role of cognitive control, 7.

Slevc, L. R., Rosenberg, J. C., & Patel, A. D. (2009). Making psycholinguistics musical: Self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychonomic Bulletin & Review, 16*(2), 374–381. https://doi.org/10.3758/16.2.374

Sloboda, J. A., & Gregory, A. H. (1980). The psychological reality of musical segments. *Canadian Journal of Psychology/Revue canadienne de psychologie, 34*(3), 274–280. https://doi.org/10.1037/h0081052

Smith, K. M. (2010). Skryabin's Revolving Harmonies, Lacanian Desire, and Riemannian Funktionstheorie. *Twentieth-Century Music, 7*(2), 167–194. https://doi.org/10.1017/S1478572211000156

Smith, K. M. (2020). *Desire in chromatic harmony: a psychodynamic exploration of fin de siècle tonality.* Oxford University Press.

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology, 107*, 102632. https://doi.org/10.1016/j.jmp.2021.102632

Soto-Faraco, S., & Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 55*(1), 23–40. https://doi.org/10.1080/02724980143000136

Spence, C., Nicholls, M. E. R., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics, 63*(2), 330–336. https://doi.org/10.3758/BF03194473

Sprouse, J., & Lau, E. F. (2013). Syntax and the brain. In M. den Dikken (Ed.), *Cambridge Handbook of Generative Syntax, the* (pp. 971–1005). Cambridge University Press. https://doi.org/10.1017/CBO9780511804571.033

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

Steedman, M. (1996). The blues and the abstract truth: Music and mental models. In A. Garnham & J. Oakhill (Eds.), *Mental Models in Cognitive Science. Essays in Honour of Phil Johnson Laird* (pp. 305–318). Psychology Press.

Steedman, M. (2000). *The syntactic process*. MIT Press.

Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The Role of Harmonic Expectancy Violations in Musical Emotions: Evidence from Subjective, Physiological, and Neural Responses. *Journal of Cognitive Neuroscience*, *18*(8), 1380–1393. https://doi.org/10.1162/jocn.2006.18.8.1380

Steinmetz, C., & Reiss, J. (2021). Pyloudnorm: A simple yet flexible loudness meter in Python. *150th Audio Engineering Society Convention*.

Stoffer, T. H. (1985). Representation of Phrase Structure in the Perception of Music. *Music Perception*, *3*(2), 191–220. https://doi.org/10.2307/40285332

Stupacher, J., Witte, M., Hove, M. J., & Wood, G. (2016). Neural Entrainment in Drum Rhythms with Silent Breaks: Evidence from Steady-state Evoked and Event-related Potentials. *Journal of Cognitive Neuroscience*, *28*(12), 1865–1877. https://doi.org/10.1162/jocn_a_01013

Sun, L., Feng, C., & Yang, Y. (2020). Tension Experience Induced By Nested Structures In Music. *Frontiers in Human Neuroscience*, *14*.

Tan, I., Lustig, E., & Temperley, D. (2019). Anticipatory Syncopation in Rock: A Corpus Study. *Music Perception*, *36*(4), 353–370. https://doi.org/10.1525/MP.2019.36.4.353

Tan, N., Aiello, R., & Bever, T. G. (1981). Harmonic structure as a determinant of melodic organization. *Memory & Cognition*, *9*(5), 533–539. https://doi.org/10.3758/BF03202347

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 217–262). Academic Press. https://doi.org/10.1016/B978-012497770-9.50009-1

Taruffi, L., & Küssner, M. B. (2019). A review of music-evoked visual mental imagery: Conceptual issues, relation to emotion, and functional outcome. *Psychomusicology: Music, Mind, and Brain*, *29*(2-3), 62–74. https://doi.org/10.1037/pmu0000226

Tekman, H., & Bharucha, J. (1998). Implicit Knowledge Versus Psychoacoustic Similarity in Priming of Chords. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 252–260. https://doi.org/10.1037/0096-1523.24.1.252

Temperley, D. (1999). Syncopation in Rock: A Perceptual Perspective. *Popular Music*, *18*(1), 19–40.

Temperley, D. (2001a). *The Cognition of Basic Musical Structures*. MIT Press.

Temperley, D. (2001b). The Question of Purpose in Music Theory: Description, Suggestion, and Explanation. *66*, 66–85. https://doi.org/10.7916/D8TT4PQZ

Temperley, D. (2004). Communicative Pressure and the Evolution of Musical Styles. *Music Perception*, *21*(3), 313–337. https://doi.org/10.1525/mp.2004.21.3.313

Temperley, D. (2007). *Music and Probability*. MIT Press.

Temperley, D. (2011). The Cadential IV in Rock. *Music Theory Online*, *17*(1). https://doi.org/10.30535/mto.17.1.8

Temperley, D., & Marvin, E. W. (2008). Pitch-Class Distribution and the Identification of Key. *Music Perception*, *25*(3), 193–212. https://doi.org/10.1525/mp.2008.25.3.193

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/S0140525X01000061

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, *55*(5), 1061–1069. https://doi.org/10.1121/1.1914648

Terhardt, E. (1984). The Concept of Musical Consonance: A Link between Music and Psychoacoustics. *Music Perception: An Interdisciplinary Journal*, *1*(3), 276–295. https://doi.org/10.2307/40285261

Terhardt, E., Stoll, G., & Seewann, M. (1982). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, *71*(3), 679–688. https://doi.org/10.1121/1.387544

Tillman, B., Bharucha, J. J., & Bigand, E. (2003). Learning and perceiving musical structures: Further insights from artificial neural networks. In *The cognitive neuroscience of music* (pp. 109–123). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198525202.003.0008

Tillmann, B., & Bigand, E. (2001). Global context effect in normal and scrambled musical sequences. *Journal of Experimental Psychology. Human Perception and Performance*, *27*(5), 1185–1196. https://doi.org/10.1037//0096-1523.27.5.1185

Tillmann, B., Koelsch, S., Escoffier, N., Bigand, E., Lalitte, P., Friederici, A. D., & von Cramon, D. Y. (2006). Cognitive priming in sung and instrumental music: Activation of inferior frontal cortex. *NeuroImage*, *31*(4), 1771–1782. https://doi.org/10.1016/j.neuroimage.2006.02.028

Tillmann, B. (2005). Implicit Investigations of Tonal Knowledge in Nonmusician Listeners. In *The neurosciences and music II: From perception to performance* (pp. 100–110). New York Academy of Sciences.

Tillmann, B. (2012). Music and Language Perception: Expectations, Structural Integration, and Cognitive Sequencing. *Topics in Cognitive Science*, *4*(4), 568–584. https://doi.org/10.1111/j.1756-8765.2012.01209.x

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*(4), 885–913. https://doi.org/10.1037/0033-295X.107.4.885

Tillmann, B., & Bigand, E. (2004). The Relative Importance of Local and Global Structures in Music Perception. *The Journal of Aesthetics and Art Criticism*, *62*(2), 211–222.

Tillmann, B., & Dowling, J. W. (2007). Memory decreases for prose, but not for poetry. *Memory & Cognition*, *35*(4), 628–639. https://doi.org/10.3758/BF03193301

Tillmann, B., Janata, P., & Bharucha, J. J. (2003). Activation of the inferior frontal cortex in musical priming. *Cognitive Brain Research*, *16*(2), 145–161. https://doi.org/10.1016/S0926-6410(02)00245-8

Toiviainen, P., & Eerola, T. (2004). The Role of Accent Periodicities in Meter Induction: A Classification Study.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic Influences On Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *33*(3), 285–318. https://doi.org/10.1006/jmla.1994.1014

Tymoczko, D. (2011). *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. OUP USA.

Tymoczko, D. (2020). Review-Essay on Fred Lerdahl's *Composition and Cognition* (University of California Press, 2019). *Music Theory Online*, *26*(1).

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian Models of Conceptual Development: Learning as Building Models of the World. *Annual Review of Developmental Psychology*, *2*(1), 533–558. https://doi.org/10.1146/annurev-devpsych-121318-084833

Ulrich, R., & Miller, J. (1993). Information Processing Models Generating Lognormally Distributed Reaction Times. *Journal of Mathematical Psychology*, *37*(4), 513–525. https://doi.org/10.1006/jmps.1993.1032

Van de Cavey, J. (2016). *Syntax across domains: overlap in global and local structure processing* (Doctoral dissertation). Ghent University.

Van de Cavey, J., & Hartsuiker, R. J. (2016). Is there a domain-general cognitive structuring system? Evidence from structural priming across music, math, action descriptions, and language. *Cognition*, *146*, 172–184. https://doi.org/10.1016/j.cognition.2015.09.013

van Ooyen, B., Cutler, A., & Bertinetto, P. M. (1993). Click Detection in Italian and English. *Eurospeech 93. 3rd European Conference on Speech Communication and Technology*, 5.

van Schijndel, M., Exley, A., & Schuler, W. (2013). A Model of Language Processing as Hierarchic Sequential Prediction. *Topics in Cognitive Science*, *5*(3), 522–540. https://doi.org/10.1111/tops.12034

Van Vugt, F., Jabusch, H.-C., & Altenmüller, E. (2012). Fingers Phrase Music Differently: Trial-to-Trial Variability in Piano Scale Playing and Auditory Perception Reveal Motor Chunking. *Frontiers in Psychology*, *3*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vogelzang, M., Mills, A. C., Reitter, D., Van Rij, J., Hendriks, P., & Van Rijn, H. (2017). Toward Cognitively Constrained Models of Language Processing: A Review. *Frontiers in Communication*, *2*. https://doi.org/10.3389/fcomm.2017.00011

von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leopold Voss.

Vuust, P., Dietz, M. J., Witek, M., & Kringelbach, M. L. (2018). Now you hear it: a predictive coding model for understanding rhythmic incongruity. *Annals of the New York Academy of Sciences*, *1423*(1), 19–29. https://doi.org/10.1111/nyas.13622

Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews Neuroscience*, *23*(5), 287–305. https://doi.org/10.1038/s41583-022-00578-5

Vuust, P., & Witek, M. A. G. (2014). Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Frontiers in Psychology*, *5*, 1111. https://doi.org/10.3389/fpsyg.2014.01111

Vuvan, D. T., & Hughes, B. (2019). Musical Style Affects the Strength of Harmonic Expectancy. *Music & Science*, *2*, 2059204318816066. https://doi.org/10.1177/2059204318816066

Wagner, N. (1995). No Crossing Branches? The Overlapping Technique in Schenkerian Analysis. *Theory and Practice*, *20*, 149–175.

Wall, L., Lieck, R., Neuwirth, M., & Rohrmeier, M. (2020). The Impact of Voice Leading and Harmony on Musical Expectancy. *Scientific Reports*, *10*(1), 5933. https://doi.org/10.1038/s41598-020-61645-4

Walsh, S. (1984). Musical Analysis: Hearing Is Believing? *Music Perception*, *2*(2), 237–244. https://doi.org/10.2307/40285293

Wason, R. (2002). Musica practica: music theory as pedagogy. In T. Christensen (Ed.), *The Cambridge History of Western Music Theory* (pp. 46–77). Cambridge University Press. https://doi.org/10.1017/CHOL9780521623711.004

Waters, K. (2005). Modes, Scales, Functional Harmony, and Nonfunctional Harmony in the Compositions of Herbie Hancock. *Journal of Music Theory*, *49*(2), 333–357.

Webber, B. L. (1988). Tense as Discourse Anaphor. *Computational Linguistics*, *14*(2), 61–73.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. https://doi.org/10.1038/nn0602-858

White, C. (2022). *The Music in the Data*. Routledge.

White, C., & Quinn, I. (2018). Chord Context and Harmonic Function in Tonal Music. *Music Theory Spectrum*, *40*(2), 314–335O. https://doi.org/10.1093/mts/mty021

Wiggins, G. A. (2007). Models of musical similarity. *Musicae Scientiae*, *11*(1), 315–338. https://doi.org/10.1177/102986490701100112

Witek, M. A. G., Clarke, E. F., Kringelbach, M. L., & Vuust, P. (2014). Effects of Polyphonic Context, Instrumentation, and Metrical Location on Syncopation in Music. *Music Perception*, *32*(2), 201–217. https://doi.org/10.1525/mp.2014.32.2.201

Witek, M. A. G., Clarke, E. F., Wallentin, M., Kringelbach, M. L., & Vuust, P. (2014). Syncopation, Body-Movement and Pleasure in Groove Music. *PLOS ONE*, *9*(4), e94446. https://doi.org/10.1371/journal.pone.0094446

Woolhouse, M., Cross, I., & Horton, T. (2016). Perception of nonadjacent tonic-key relationships. *Psychology of Music*, *44*(4), 802–815. https://doi.org/10.1177/0305735615593409

Yust, J. (2006). *Formal Models of Prolongation* (Doctoral dissertation). University of Washington.

Yust, J. (2015). Voice-Leading Transformation and Generative Theories of Tonal Structure. *Music Theory Online*, *21*(4).

Yust, J. (2018). *Organized Time: Rhythm, Tonality, and Form*. Oxford University Press.

Zhang, J., Jiang, C., Zhou, L., & Yang, Y. (2016). Perception of hierarchical boundaries in music and its modulation by expertise. *Neuropsychologia*, *91*, 490–498. https://doi.org/10.1016/j.neuropsychologia.2016.09.013

Zhang, J., Zhou, X., Chang, R., & Yang, Y. (2018). Effects of global and local contexts on chord processing: An ERP study. *Neuropsychologia*, *109*, 149–154. https://doi.org/10.1016/j.neuropsychologia.2017.12.016

Gabriele Cecchetti

PhD, MPhil (Cantab.), MMus, BSc

Music cognition | Digital musicology | Music theory and performance

EPFL CDH DHI DCML
Station 14, INN 115
CH – 1015 Lausanne
☎ *+39 339 8058123*
✉ *gabriele.cecchetti@epfl.ch*
iD *0000–0002–1486–1886*

Appointments

2024–present **Post-doctoral Research Fellow**.
The MARCS Institute for Brain, Behaviour, and Development, Western Sydney University. Sydney (AU).

2019–2024 **Doctoral Assistant**.
Digital and Cognitive Musicology Lab, EPFL. Lausanne (CH).

oct-dec 2022 **Visiting Fellow**.
The MARCS Institute for Brain, Behaviour, and Development, Western Sydney University. Sydney (AU).

2018 **Maestro collaboratore**.
Accademia Nazionale di Santa Cecilia (IT).

Education

2024 **PhD in Systematic Musicology**.
Digital and Cognitive Musicology Lab, EPFL. Lausanne (CH).

2019 **MPhil in Music Studies**, *with Distinction*.
Centre for Music and Science, University of Cambridge. Cambridge (UK).

2018 **Diploma Accademico di II Livello (MMus) in Violoncello**, *cum laude*.
Conservatorio "S. Cecilia". Rome (RM).

2015 **Laurea (BSc) in Physics**.
Università "La Sapienza". Rome (RM).

since 2010 **Instrumental training (Violoncello)**, *Francesco Storino (Orchestra dell'Accademia Nazionale di Santa Cecilia, Rome), Orfeo Mandozzi and Marcin Seniavsky (International Piano Academy, Imola), Enrico Bronzi (Fondazione Musicale "S. Cecilia", Portogruaro), Francesco Pepicelli (Accademia Sherazade, Rome), Marco Fiorini (String Quartet, Rome).*

Funding, Scholarships, and Awards

2024 **"Marica de Vincenzi" Post-doctoral Fellowship**, Department of Cognitive Science, University of Trento, 45.350€ (declined).

2023 **ICMPC Travel Award**, 17th International Conference of Music Perception and Cognition, 30.000JPY.

2022 **Best student paper award [23]**, *Institute of Mathematics and its Applications*, 1st Mathematics in Music Conference, Royal College of Music.

2022 **International Visiting Scholar Award**, *The MARCS Institute for Brain, Behaviour and Development*, Western Sydney Unversity, 2500$.

2020 **SNF Spark Grant (Co-Investigator with P.I. Steffen A. Herff)**, *"Wanderful music: A systematic investigation into music-induced mind wandering"*, Swiss National Science Foundation (CH), 95.775CHF.

2019 **Master's Award**, *Homerton College*, University of Cambridge (UK).

2019 **Victoria Brahm-Schild Internship Grant**, *Homerton College*, University of Cambridge (UK).

2018 **Cambridge European Scholarship**, *Cambridge Trust*, Cambridge (UK), 10.000£.

2018 **Torno Subito**, *Regione Lazio*, Rome (RM), 15.055€.

2018 **Berliner Symphoniker Orchestra, traineeship**, *Berliner Symphoniker Orchestra*, Berlin (DE), 1st awardee.

2016–2017 **Bursary (Chamber music)**, *Conservatorio "S. Cecilia"*, Rome (RM), 1.500€.

2016 **Concert Award (Chamber Music)**, *Campus delle Arti*, Bassano del Grappa (VE).

Professional Experience

Academic Service

2024–present **XXI International Conference of Music Theory and Analysis**, *Scientific Committee*, Conservatorio G. Martucci, Salerno (IT).

2021–present **Italian Society for Music Theory and Analysis (GATM)**, *Scientific Board.*

2021–present **Analitica. Online Journal of Music Studies**, *Editorial Board.*

2021–present **Empirical Musicology Review**, *Copy Editor.*

2021–present **Reviewer**, *Cognition, Musicae Scientiae, Rivista di Analisi e Teoria Musicale, RiLUnE.*

2023 **XX International Conference of Music Theory and Analysis**, *Scientific Committee*, Conservatorio G. Martucci, Salerno (IT), 19–22 October 2023.

2023 **XX International Conference of Music Theory and Analysis**, *Session Chair*, Conservatorio G. Martucci, Salerno (IT), 19–22 October 2023.

2022 **XIX International Conference of Music Theory and Analysis**, *Session Chair*, Conservatorio G. Martucci, Salerno (IT), 20–23 October 2022.

2020–2021 **Italian Society for Music Theory and Analysis (GATM)**, *Research Committee.*

2020–2021 **Italian Society for Music Theory and Analysis (GATM)**, *Conferences and Seminars Committee.*

2019 **Society for Education and Music Psychology Research**, *Conference organizer*, University of Cambridge, Cambridge (UK), 22 March 2019.

Consultancy

2023 **Conservatorio "G. Giacomantonio", Cosenza**, *Music Theory in the XXI Century*.

2022 **Hochschule für Musik Mannheim**, *Digital Humanities in Music Theory*.

Teaching

2021–2022 **EPFL**, *Lausanne (CH)*.
Teaching Assistant: *Digital Musicology* (MSc).

2019–present **Villa Pennisi in Musica (Summer School)**, *Acireale (CT)*.
Yearly seminars in *Psychoacoustics* for architects and engineers.

2021 **AVOS Project Academy**, *Rome*.
Seminars in Advanced Music Theory and Analysis.

2021 **EPFL**, *Lausanne (CH)*.
Teaching Assistant: *Musical Theory and Creativity – Algorithmic Composition* (BSc).

Supervision

2023 **Léo Bruneau**, *6ECTS Bachelor Project*.

2022-2023 **Siyi Wang**, *12ECTS Semester Project*.

2021-2022 **Cédric Tomasini**, *MSc Thesis*.

Professional Musical Experience

2010–present Cellist in Orchestra Filarmonica di Benevento (Pappano, Rana, Biondi, Piovano, Rizzari, Ciampa), "Roma Tre" Symphony Orchestra (Piovano), Orchestra del Conservatorio "S. Cecilia" (Renzetti, Aprea, Lucantoni), Orchestra del Conservatorio di Potenza, Accademia Ducale, Prometheus Chamber Orchestra (Ferranti), string quartet (Rome, Bologna), piano duo (Rome, Venice), piano trio (Rome, Imola).

2018 **Accademia Nazionale "S. Cecilia"**, *Rome (RM)*, Maestro Collaboratore.

2012–2015 **Prometheus Chamber Orchestra**, Rome (RM), Founder and organizer.

2012–2016 **Accademia Nazionale "S. Cecilia"**, *Rome (RM)*, Youth Orchestra tutor.

Professional Memberships

2021–present **Cognitive Science Society**.

2018–present **Italian Society for Music Theory and Analysis (GATM)**.

Doctoral Training

2023 Advanced Music Theory Seminars (Dres Schiltknecht – Michael Polth, Hochschule für Musik Mannheim)

2022 Electroencephalography (Steffen Herff, EPFL)

2020 Neuroscience: behavior and cognition (Olaf Blanke – Michael Herzog – Maria Sandi Perez, EPFL)

2020 Winter School on AI and ethics (Christoph Benzmüller – Bertrand Lomfeld, FU Berlin)

2019 Schenkerian and Tonfeld Theory for Music Analysis (Oliver Schwab-Felisch, TU Berlin – Johannes Schild, ZHdK)

2019 Applied Data Analysis (Robert West, EPFL)

Languages

Italian:	native	English:	proficient (C2)
German:	advanced (C1)	French:	basic

Computing Skills

Coding: Python (Data Analysis, Machine Learning), R (Bayesian Data Analysis), PsychoPy (Experimental Design), C, Mathematica

Editing: LaTeX, Open Journal System

Music: LilyPond, MuseScore, Finale, Sibelius, music21, Audacity

Publications

Preprints and Forthcoming Publications

[1] Steffen A. Herff, Leonardo Bonetti, **Gabriele Cecchetti**, Peter Vuust, Morten L. Kringelbach, and Martin A. Rohrmeier. Hierarchical syntax models of music predict theta oscillations during music listening. *bioRxiv*, in review.

[2] Steffen A. Herff, **Gabriele Cecchetti**, Petter Ericson, and Estefania Cano. Solitary Silence and Social Sounds: Music influences mental imagery, inducing thoughts of social interactions. *bioRxiv*, in review.

[3] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Priming of abstract harmonic structure in music. *Journal of Experimental Psychology: Human Perception and Performance*, In review.

Peer-reviewed Publications

[4] **Gabriele Cecchetti**, Cédric A. Tomasini, Steffen A. Herff, and Martin A. Rohrmeier. Interpreting rhythm as parsing: Syntactic-processing operations predict the migration of visual flashes as perceived during listening to musical rhythms. *Cognitive Science*, 47(12), 2023.

[5] Ludovica Schaerf, Sabrina Laneve, **Gabriele Cecchetti**, Johannes Hentschel, and Martin A. Rohrmeier. The evolution of style in Debussys piano music: A DFT-based diachronic corpus study. *Humanities and Social Science Communications*, 10, 2023.

[6] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Hearing functional harmony in jazz: a perceptual study on music-theoretical accounts of extended tonality. *Musicae Scientiae*, 27(3):672–697, 2022.

[7] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Musical garden paths: Evidence for syntactic revision beyond the linguistic domain. *Cognitive Science*, 46(7), 2022.

[8] Steffen A. Herff, **Gabriele Cecchetti**, Liila Taruffi, and Ken Deguernel. Music influences vividness and content of imagined journeys in a directed visual imagery task. *Scientific Reports*, 11, 2021.

[9] Steffen A. Herff, Daniel Harasim, **Gabriele Cecchetti**, Christoph Finkensiep, and Martin A. Rohrmeier. Hierarchical syntactic structure predicts listeners' sequence completion in music. In *Proceedings of the 53 Annual Conference of the Cognitive Science Society*, 2021.

[10] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Musical syntactic structure improves memory for melody: evidence from the processing of ambiguous melodies. In *Proceedings of the 53 Annual Conference of the Cognitive Science Society*, 2021.

[11] **Gabriele Cecchetti**, Steffen A. Herff, Christoph Finkensiep, and Martin A. Rohrmeier. The experience of musical structure as computation: what can we learn? *Rivista di Analisi e Teoria Musicale*, 26(2):91–127, 2020.

Peer-reviewed Conferences

[12] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Towards an empirical model of structural hearing as representation and processing. In *XX International Conference of Music Theory and Analysis*, Conservatory "G. Martucci", Salerno, 19–22 October 2023.

[13] **Gabriele Cecchetti**, Christoph Finkensiep, Xinyi Guan, Steffen A. Herff, and Martin A. Rohrmeier. A generative framework for modelling music processing beyond the computational level. In *45th Annual Meeting of the Cognitive Science Society (CogSci 2023)*, Sydney, 26-29 July 2023.

[14] **Gabriele Cecchetti**, Christoph Finkensiep, Xinyi Guan, Steffen A. Herff, Anna Fiveash, Claire Pelofi, John E. Drury, and Martin A. Rohrmeier. Music Cognition between Theory and Experiment (Symposium). In *45th Annual Meeting of the Cognitive Science Society (CogSci 2023)*, Sydney, 26-29 July 2023.

[15] **Gabriele Cecchetti**, Cédric A. Tomasini, Steffen A. Herff, and Martin A. Rohrmeier. Parsing rhythm: a behavioural correlate of syntactic-parsing computations in the perception of musical rhythm. In *17th International Conference of Music Perception and Cognition ICMPC17-APSCOM7*, Nihon University, Tokyo, 24-28 August 2023.

[16] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Abstract structural priming in idiomatic musical harmony. In *17th International Conference of Music Perception and Cognition ICMPC17-APSCOM7*, Nihon University, Tokyo, 24-28 August 2023.

[17] Steffen A. Herff, **Gabriele Cecchetti**, Petter Ericson, and Estefanía Cano. Musics social presence: Inducing imagined social interactions through background music. In *17th International Conference of Music Perception and Cognition ICMPC17-APSCOM7*, Nihon University, Tokyo, 24-28 August 2023.

[18] Ken Deguernel, **Gabriele Cecchetti**, and Steffen A. Herff. Emotion, Motion, and Abstract Notions: Insights in the role of imagination in professional musicians practices from semi-guided interviews. In *17th International Conference of Music Perception and Cognition ICMPC17-APSCOM7*, Nihon University, Tokyo, 24-28 August 2023.

[19] **Gabriele Cecchetti**. Musical garden paths: syntactic ambiguity and retrospective revision in music processing. In *MARCS Conference Series 2022: Music Science*, MARCS Institute for Brain, Behaviour and Development (Western Sydney University), 11 November 2022.

[20] **Gabriele Cecchetti**. Challenges and prospects for the empirical investigation of analytical interpretation: the case of rhythm. In *XIX International Conference of Music Theory and Analysis*, Conservatorio G. Martucci, Salerno, 20–23 october 2022.

[21] Francesco Maschio, Simonetta Sargenti, Matteo Farné, and **Gabriele Cecchetti**. Un contributo per la costruzione di un sistema di analisi basata sullascolto per la musica elettronica del XXI secolo. In *XIX International Conference of Music Theory and Analysis*, Conservatorio G. Martucci, Salerno, 20–23 october 2022.

[22] Sabrina Laneve, Ludovica Schaerf, Johannes Hentschel, **Gabriele Cecchetti**, and Martin A. Rohrmeier. On ambiguity and fragmentation of tonal structure in Debussys piano music: a DFT approach. In *XIX International Conference*

of Music Theory and Analysis, Conservatorio G. Martucci, Salerno, 20–23 october 2022.

[23] Ludovica Schaerf, Sabrina Laneve, Johannes Hentschel, **Gabriele Cecchetti**, and Martin A. Rohrmeier. Discrete Fourier Transform unveils decreasing diatonicity and increasing fragmentation in Debussys piano music: a diachronic corpus study*. In *Mathematics in Music Conference*, Institute of Mathematics and its Applications and Royal College of Music, London, 13–15 july 2022. ***Best student paper award**.

[24] Steffen A. Herff, Liila Taruffi, **Gabriele Cecchetti**, and Ken Deguernel. Empirical characterisation of the effect of music on imagination. In *16th International Conference on Music Perception and Cognition & 11th triennial conference of the European Society for the Cognitive Sciences Of Music*, 2021.

[25] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Robustness to interference in memory performance and syntactic representations of melodies. In *16th International Conference on Music Perception and Cognition & 11th triennial conference of the European Society for the Cognitive Sciences Of Music*, 2021.

[26] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Musical syntactic revision in the perception of melodies. In *16th International Conference on Music Perception and Cognition & 11th triennial conference of the European Society for the Cognitive Sciences Of Music*, 2021.

[27] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Hearing harmonic functionality in the idiom of extended tonality: Perceptual manifestation of octatonic substitutions. In *16th International Conference on Music Perception and Cognition & 11th triennial conference of the European Society for the Cognitive Sciences Of Music*, University of Sheffield, Sheffield, 2021.

[28] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Functional equivalence in chromatic harmony: a perceptual account. In *XVII International Conference of Music Theory and Analysis*, Istituto Superiore di Studi Musicali G. Lettimi, Rimini, 26–29 november 2020.

[29] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Dealing with ambiguity during online processing of tonal melodies: evidence for syntactic revision in music. In *XVII International Conference of Music Theory and Analysis*, Istituto Superiore di Studi Musicali G. Lettimi, Rimini, 26–29 november 2020.

[30] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Revision of musical structure as a perceptual phenomenon. In *Music and Psychology Research Conference*. Australian Music Psychology Society, 9 October 2020.

[31] **Gabriele Cecchetti**, Steffen A. Herff, and Martin A. Rohrmeier. Perceptual equivalence in preparing global harmonic closure in the jazz idiom. In *Brain, Cognition, Emotion, Music*, Canterbury, 20–21 may 2020. University of Kent.

[32] **Gabriele Cecchetti**. "Tacendo dicea": the early modern sound of Petrarchan speaking silence. In *XXVI Convegno annuale SIdM*, Conservatorio E.R. Duni, Matera, 18–20 october 2019.

[33] **Gabriele Cecchetti**. Exploring tonal hierarchies with an information-theoretic approach to cognitive similarity. In *XVI International Conference of Music Theory and Analysis*, Istituto Superiore di Studi Musicali G. Lettimi, Rimini, 10–13 october 2019.

Invited Talks and Workshops

[34] Claire Arthur, David Baker, John Ashley Burgoyne, **Gabriele Cecchetti**, Tuomas Eerola, Mary Farbood, Christoph Finkensiep, Peter Harrison, Stephan Koelsch, Elizabeth Margulis, Fabian Moss, Markus Neuwirth, Marcus Pearce, Claire Pelofi, Yannis Rammos, Martin Alois Rohrmeier, and Anja Volk. Decoding Musical Structure: Theory, Computation, and Neuroscience (Workshop). Monte Verità: Congressi Stefano Franscini, 5-9 February 2023.

[35] **Gabriele Cecchetti**. The emergence of structural representations in music "as" language. MARCS Research Meetings, MARCS Institute for Brain, Behaviour, and Development, 6 December 2022.

[36] **Gabriele Cecchetti**. Modelling the interpretation of musical rhythm as syntactic parsing. Music Science Seminars, MARCS Institute for Brain, Behaviour, and Development, 22 November 2022.

[37] Claire Arthur, David Baker, John Ashley Burgoyne, **Gabriele Cecchetti**, Johanna Devaney, Christoph Finkensiep, Klaus Frieler, Mathieu Giraud, Mark Gotham, Johannes Hentschel, Ana Llorens, Anna Matuszewska, Fabian Moss, Nestor Nápolez López, Markus Neuwirth, Yannis Rammos, Martin Alois Rohrmeier, David Sears, and Dmitri Tymoczko. Representing Harmony: Goals and Challenges (Workshop). École Polytechnique Fédérale de Lausanne, 13-16 September 2022.

[38] **Gabriele Cecchetti**. Computational tools for research and teaching in music theory. Hochschule für Musik und Darstellende Kunst Mannheim, 26 May 2022.

[39] **Gabriele Cecchetti**. Perceptual correlates of functional equivalence in chromatic harmony. Music Theory Research Group, University of Liverpool, 16 December 2020.

[40] **Gabriele Cecchetti**. Exploring the perceptual manifestations of syntactic processing in music. In *Perception, Cognition & Aesthetics Seminars*. Centre for Digital Music, Queen Mary University of London, 1 December 2020.

Outreach Publications, Talks, and Activities

[41] **Gabriele Cecchetti**. Tra teoria e percezione: un orecchio alla ricerca. Conservatory "G. Giacomantonio", Cosenza, 21 September 2023.

[42] **Gabriele Cecchetti**. Musical garden paths: making sense of structure in music listening. In *Digital Humanities Madness*. Digital Humanities Institute, EPFL, 18 November 2021.

[43] **Gabriele Cecchetti** and Lorenzo Dello Schiavo. Nature of language vs. language of nature — the limits of human languages in describing nature. In A. Altamore and G. Antonini, editors, *Galileo and the Renaissance Scientific Discourse*, Rome, 2010. Edizioni Nuova Cultura.

2014–2016 **Accademia Nazionale "S. Cecilia"**, *Rome (RM)*, Cellist and speaker in the educational and outreach programs.

Theses

PhD, 2024 Hearing structure in music: An empirical inquiry into listening as representation and processing.
Digital and Cognitive Musicology Lab, École Polytechnique Fédérale de Lausanne
Supervisors: Prof. Martin A. Rohrmeier, Dr. Steffen A. Herff

MPhil, 2019 Exploring tonal hierarchies with an information-theoretic approach to cognitive similarity.
Centre for Music and Science, University of Cambridge
Supervisor: Prof. Ian Cross

MMus, 2018 Voice without words. Auditory imagination as instrumental poetics in Robert Schumann's *Hausmusik*.
Conservatorio *Santa Cecilia*, Roma
Supervisor: M° Dante Cianferra

BSc, 2015 Finite symmetry groups: a semiclassical algebraic model for the molecular spectroscopy of C_{60}.
Department of Physics, Università *La Sapienza*, Roma
Supervisor: Prof. Antonio D. Polosa