Grouping in music and language

Jonah Katz

West Virginia University

katzlinguist@gmail.com

Keywords: music, language, grouping, prosody, modularity

Abstract. This paper reviews evidence concerning the nature of grouping in music and language, and attempts to draw from this topic some general lessons about music, language, and cognition more generally. The two domains both involve correspondence between auditory discontinuities and group boundaries, reflecting the Gestalt principles of proximity and similarity, as well as a nested, hierarchical organization of constituents. There are also obvious differences between musical and linguistic grouping. Grappling with those differences requires one to think in detail about modularity, information flow, levels of description, and the functional nature of cognitive domains.

1 Theoretical and methodological preliminaries

Similarities and differences between language and music are of interest in part because they bear on the question of *modularity*: to what extent do various cognitive processes apply to different sensory inputs, drive different kinds of behavior, and make use of different representations? There is no shortage of review literature on the music-language comparison, from a variety of different methodological and theoretical perspectives, and the current paper does not comprehensively summarize all available research. Instead, I explore one particular aspect of music-language similarity, the structure of *grouping*, and from this exercise suggest some general lessons about the theory and practice of comparative cognition. One crucial point is that *music* and *language* each constitute a complex and heterogeneous set of representations, processes, and behaviors. Each of these components may rely on information from other components internal or external to their cognitive domains. And each component may be described at different levels. Progress in understanding the relationship between music and language requires clarifying the details of implicit computational principles underlying specific processes, representations, and behaviors in the two domains.

One of the central questions in the history of cognitive science is the extent to which language makes use of information in a distinct manner compared to other cognitive domains. While the term *modularity* has several other implications (e.g. brain localization, automaticity), I use it here to single out this particular question. While it may seem straightforward to examine existing theories of language and music and compare their properties, there are a number of pitfalls inherent to the enterprise. Because neither language nor music can be rigidly defined by association with a specific

set of behaviors or computations, the only coherent way to think about either domain is as a collection of heterogeneous cognitive resources. If we find that many of these resources are common to music and language, it may imply that they are domain-general or it may imply that *music* and *language* are words used to describe overlapping sets of cognitive domains.

When comparative study suggests that music and language share some property, it raises several questions. One is what kind of a property it is (e.g. structural, behavioral, neural). A second question is why the property is the way it is, instead of being some other property. A third question is why the two domains might share the property. The variety of possible answers to such questions implies a variety of different forms of domain-generality that could be of interest to cognitive scientists. Two domains might share some resource for 'low-level' reasons related to some more basic domain, for reasons related to the internal structure of the domains themselves, or, possibly, by coincidence. Any such parallels may be the consequence of homology, where two domains share properties because they are deeply related at the level of human evolution; or analogy, where the property arises independently more than once because it is a 'good solution' to some problem.

Marr (1982) distinguishes three levels of description for any cognitive process. The *computational* level describes in abstract symbolic terms what is being processed, what the process itself is like, and why the process looks the way it does instead of some other way. The *algorithmic* level describes how that process is applied to input representations in real time. And the *implementational* level describes the machinery that performs the algorithm, generally part of the brain.

Marr emphasizes the fact that descriptions at the three levels are necessarily related to one another in intricate and complex ways, but the relationship is not deterministic: some analytical choices at one level are independent of choices at other levels. This is important because it also has implications for the study of shared resources between cognitive domains. Music and language may share computational properties, but may implement them with different algorithms in different areas of the brain, or with the same algorithm in different parts of the brain, or for all we know, with different algorithms in the same part of the brain. Similarly, they may share some processing or learning algorithm but apply it to completely different representations and thus generate systems with completely different computational properties. In this paper, I focus on the computational level, partly out of the conviction that if we are to understand parallels in processing or neural substrates, we first must understand *what* is being processed. Heffner & Slevc (2015) offer an overview more oriented towards processing and neuroimaging.

There are several related topics that I don't have space to discuss here. I briefly touch on meter and syntax, but only to the extent that they interact with grouping. Each of those components could be the subject of a book-length review on its own. For interesting overviews, see Fitch (2015) on meter and Rohrmeier *et al*. (2015) on syntax. I do not discuss textsetting here; Dell (2015) gives a highly relevant overview of one genre. While imposing a text on a piece of music (or vice versa) clearly involves shared linguistic and musical grouping, I've chosen to focus on evidence from independently motivated principles of music and language.

2 Grouping in language

Speech and music both involve sequences of auditory events in time. Some of those events pattern together to the exclusion of others with regard to perceived coherence, acoustic patterns, or other phenomena. I follow Lerdahl & Jackendoff (1983) in referring to this topic as *grouping*. Linguists generally call it *prosodic phrasing*, but I use *grouping* here for consistency. In this section, I briefly review the overarching theory and specific acoustic reflexes of grouping in language.

Linguistic theories of grouping tend to be concerned with relating two sets of empirical results. The first involves context-dependent phonetic realizations of particular sounds. Such patterns can only be described with the proviso that some sounds that pattern together to the exclusion of others, that is, with *constituents*. The second set of results involve these same grouping constituents; they can only be described with (among other things) reference to syntactic structure (Selkirk 1972, Pierrehumbert 1980). The dominant framework for unifying these results is the *prosodic hierarchy* (Selkirk 1984, Nespor & Vogel 1986). This approach characterizes the sound reflexes of grouping as resulting from: (1) a function that maps from syntactic structure to groups; and (2) a function that maps from mental representations of grouped sounds to the actual pronunciation of those sounds. As such, prosodic phonology requires a theory of groups, a theory of the function from groups to sounds, and a theory of the function from syntactic structure to groups.

2.1 Broad characteristics of linguistic grouping representations

A variety of factors go into determining the grouping structure of an utterance: at least syntax, pragmatics (including *information structure*), and affect play a role (Shattuck-Hufnagel & Turk 1996). The general 'shape' of grouping structures in language appears to be sets of constituents

nested hierarchically inside larger constituents (Selkirk 1984, Nespor & Vogel 1986, Hayes 1989). For example, some instance of the English utterance *the child in front discovered an object* might be grouped as in Figure 1. Grouping is shown in two formally equivalent visual forms here: the bracket notation generally used in linguistics (though without any notation of prominence) and the grouping notation introduced by Lerdahl & Jackendoff (1983) in their exposition of musical grouping.

[FIGURE 1 ABOUT HERE]

There are several important pieces of background to appreciate here. First, the mapping from syntax to grouping is not deterministic, so a range of slightly different grouping structures is possible for this utterance. Second, theories differ as to how many distinct levels of grouping are present in such an utterance, and in details of where the boundaries between groups fall. Third, representations of linguistic grouping generally also display headedness, which I omit here but discuss in the final section of this paper. Finally, most linguistic theories posit labels for the different levels shown in figure 1. That is, groups at the same level are posited to be the same *type* of group, and sound patterns are generally described in terms of those types. Typical labels here might be: (1) *syllable*, (2) *prosodic word*, (3) *phonological phrase*, and (4) *utterance*, though the number and nature of these labels differs between theories.

In this example, I leave out the traditional labels. This reflects an emerging consensus that prosodic groups are not as neatly layered as originally believed, instead closely mirroring syntactic structure (Ladd 1986, Ishihara 2003, Féry & Truckenbrodt 2005). The resultant theories either weaken (Féry

2010, Selkirk 2011) or eliminate (Wagner 2005) the distinctions between various levels of the prosodic hierarchy.

Despite differences between theories, most researchers agree on several broad formal properties of grouping structure in language. Phonetic material is exhaustively partitioned into groups. If a group *X* contains some of the elements in another group *Y*, then either *X* or *Y* contains all of the elements in the other group (no partial overlap). Every utterance coincides with a single group that contains all other groups. And there is a strong tendency for groups to contain exactly two smaller groups. These general representational properties of linguistic grouping are inferred on the basis of sound patterns displayed by individual speakers of various languages. Next, we turn our attention to the nature of those sound patterns.

2.2 The phonetics of linguistic grouping

A wide variety of phonetic and/or phonological generalizations have been described with reference to grouping structure. In this section, I focus on three phonetic dimensions: duration, f0/pitch, and consonant manner. Details of all three dimensions depend on the context in which a sound is uttered, and stating those contexts requires reference to grouping structure.

2.2.1 Duration

The duration of a sound depends on many factors, including its inherent features, speech rate, the sounds that occur around it, and emphasis. Even in the face of this pervasive variation, however, it is possible to identify strong tendencies in how grouping affects duration.

Many languages lengthen the final element within a group. This *final lengthening* occurs at various levels of group boundary, and affects a variety of sounds located near such boundaries. Wightman *et al*. (1992) find successively longer vowels at the ends of English words, small prosodic groups, and larger groups. Lengthening at the ends of English words compared to word-medial vowels, on the other hand, is harder to detect and may be limited to vowels near pitch accents (Turk & Shattuck-Hufnagel 2000). Gordon & Munro (2007) provide evidence for final lengthening of Chickasaw vowels at the utterance, phrase, and (perhaps) word levels. They also review other languages where final lengthening is attested at one or more levels, including Arabic (de Jong & Zawaydeh 1999), Finnish (Oller 1979), Greenlandic Eskimo (Nagano-Madsen 1992), Mandarin (Duanmu 1996), Yoruba (Nagano-Madsen 1992), and Creek (Johnson & Martin 2001).

Many languages also lengthen the *initial* element within groups. This mainly affects consonants, and is true of both articulatory and acoustic measurements. Consonants show initial lengthening at one or more levels of grouping in Korean (Jun 1993), English (Fougeron & Keating 1997), French (Keating *et al*. 2003), Gurindji (Ennever *et al*. 2017), Taiwanese (Keating *et al*. 2003), and Japanese (Onaka *et al*. 2003). These initial duration differences are generally accompanied by a suite of articulatory effects referred to as *initial strengthening*; I leave this aside until the discussion of consonant manner in section 2.2.3.

In general, then, sounds at the initial and final edges of groups tend to be longer than their counterparts internal to groups. This is not true for every speaker, sound, level of grouping, and language, but is a fairly robust generalization. The reverse pattern, where initial or final elements in a group are *shorter* than their medial counterparts, is extremely rare.

2.2.2 Pitch

Linguistic pitch is used for many different purposes. The most relevant one for grouping is *edge tones*, pitch targets or movements that occur at the edges of groups. Linguistic theories include edge tones because many languages tend to display extreme f0 values and/or f0 movement at the beginnings and ends of groups.

In American English, f0 movement tends to occur at the ends of groups corresponding to syntactic phrases (e.g. Pierrehumbert 1980). These f0 changes do not correspond to stress or lexical tone, but to the illocutionary force or discourse function of constituents. Most researchers posit at least two levels of grouping in English that generate edge tones (e.g. Beckman & Pierrehumbert 1986; Ladd 1986), which means that higher-level group boundaries will tend to have more tonal targets and hence more f0 movement than lower-level boundaries. For instance, one analysis of the English 'continuation rise' posits a combination of edge tones at the intermediate- and intonational-phrase levels of grouping (Pierrehumbert & Hirschberg 1990).

Similar edge-tone systems exist in a variety of languages that differ from English in the presence and nature of stress, pitch accent, and lexical tone. Bengali, for instance, differs from English in having

almost entirely predictable word stress, but its intonational system also features edge tones at two levels of grouping (Hayes & Lahiri 1991). Tokyo Japanese lacks word stress but displays lexical (unpredictable) pitch accents in some words. Groups corresponding to single content words with adjacent function morphemes are generally realized with an initial high tone and a final low tone, resulting in a sharp f0 rise phrase-initially (McCawley 1968). This rise is larger ('pitch reset') at the beginnings of groups that correspond to larger syntactic constituents (Pierrehumbert & Beckman 1988). Seoul Korean plausibly lacks both stress and pitch accent; groups are delimited most consistently by a final rise in f0 (Jun 1993, Kim 2004). As in English, higher levels of grouping involve the agglutination of additional edge tones (Jun 1993). Sri Lankan Malay, which lacks stress and accent, also features groups demarcated by final f0 rises (Nordhoff 2012).

Some languages with lexical tone also tend to align more complex or dynamic f0 patterns near group boundaries, sometimes but not always involving edge tones. Thai has lexical tone and predictable stress (Tingsabadh & Abramson 1993). Many lexical tones take a complex and dynamic form at the ends of phrases but are simplified and flattened phrase-internally (Morén & Zsiga 2006). This is due not to edge tones, but rather to simplification of lexical contour tones everywhere *except* at the edges of groups. Bantu languages, on the other hand, generally display lexical tone but lack word stress. Many of them feature penultimate lengthening at the utterance, phrase, or word level of grouping. This lengthening is frequently accompanied by edge tones that produce more complex f0 contours on the second-to-last vowel in a group than in other prosodic contexts (Hyman 2013).

In sum, many languages display 'extra' tonal movement at the edges of groups, compared to group-internal contexts. While the amount of f0 movement in any given context can also be affected by

factors such as metrical prominence and lexical tone, edge tones are attested independently of these phenomena and can interact with them.

2.2.3 Consonantal manner

A third property dependent on grouping structure is *manner*, a set of phonetic features pertaining acoustically to the magnitude and velocity of changes in intensity. In articulatory terms, manner corresponds roughly to the degree of constriction in the vocal tract: vowels are associated with relatively open vocal tract configurations, and consonant configurations range from wide and vowel-like (approximants) to extremely narrow (obstruents).

Consonants are often longer and less vowel-like at group boundaries, shorter and more vowel-like within groups. The *initial strengthening* literature finds that consonants are longer at the beginnings of larger groups, and also that the articulatory gestures associated with their constrictions are more extreme (e.g. Byrd & Saltzman 1998; Keating *et al*. 2003; Onaka *et al*. 2003). In English, words that begin with vowels are more likely to display initial glottal constrictions at the beginnings of larger groups (Pierrehumbert & Talkin 1992; Dilley *et al*. 1996). English vowels are also more likely to occur with glottalization at the *ends* of larger groups (Redi & Shattuck-Hufnagel 2001).

Phonological theory describes group-dependent manner differences in terms of *lenition* and *fortition*. While these terms are used to describe a broad and heterogeneous set of phonetic patterns, there is a 'core' set of lenition patterns seen in many language families that tends to target medial consonants at one or more levels of grouping, making them less constricted and/or shorter than their initial

counterparts (see Kirchner 1998 and Lavoie 2001 for typological surveys). Two of the more common lenition processes, spirantization and voicing, are illustrated in examples 1 and 2, respectively.

(1) Spirantization in Spanish

#__	**[+approx]__[+approx.]**
[g̲oðo] ‘Goth’	[bisiɣ̲oðo] ‘Visigoth’
[b̲eso] ‘kiss’	[elβ̲eso] ‘the kiss’
[d̲ia] ‘day’	[ojð̲ia] ‘nowadays

(2) Voicing in Sanuma (Yanomaman, Borgman 1986)

#__	**[+approx]__[+approx.]**
[t̲elulu] ‘dance’	[hud̲e] ‘heavy’
[p̲aso] ‘spider monkey’	[ib̲a] ‘my’
[k̲ahi] ‘mouth’	[ãg̲a] ‘tongue’
[t͡s̲inimo] ‘corn’	[had͡z̲a] ‘deer’

In Spanish, some morphemes begin with voiced stops in isolation (left column). But when phrase-medial and flanked by approximants or vowels, the same sounds are pronounced as approximants (right column). More generally, the approximants resulting from lenition are absent phrase-initially in

Spanish, and voiced stops are absent phrase-medially when flanked by vowels or approximants. In Borgman's (1986) description of Sanuma, word-initial stops are voiceless (left column), but word-medial stops flanked by vowels are optionally voiced (right column). Kingston (2008) and Katz (2016) propose that these typologically ubiquitous patterns align larger changes in intensity with group boundaries and smaller changes with non-boundaries. On this view, voicelessness and stopping are favored at group boundaries because they create larger changes in intensity, while voicing and continuancy are favored medially because they entail smaller changes from surrounding sounds. While examples (1) and (2) are meant to be simple illustrations, the exact nature and contexts of lenition and fortition phenomena can be quite complex, often varying by the phonetic features of the affected sounds or adjacent sounds.

In sum, a common effect of group boundaries on acoustic patterning is that consonants tend to be less vowel-like adjacent to boundaries, more vowel-like medially. The result, if consonants are adjacent to vowels, is larger changes in intensity near group boundaries and smaller changes medially.

3 Grouping in music

While musical grouping has not been studied as thoroughly as its linguistic counterpart, there is some agreement as to how it works. Lerdahl & Jackendoff (1983), in their *Generative Theory of Tonal Music* (henceforth *GTTM*), lay out a theory of musical grouping in great detail and subsequent work provides a fair bit of support for their description. Other models (e.g. Narmour 1990, Cambouropoulos 2001) differ in details and orientation but tend to agree on the broad characteristics of grouping. The motivation for grouping as a level of representation in music, just as in language,

pertains to the relationship between sound structure and syntactic structure. In particular, *GTTM* argues that: (1) experienced listeners tend to parse musical events into certain kinds of constituents on the basis of auditory features; and (2) those constituents are relevant to interpreting the harmonic, rhythmic, and/or thematic information within a piece. The latter aspects of grouping are addressed separately in *GTTM* as *time-span* and *prolongational reduction*. The theory is couched in somewhat different terms than the linguistic theory of prosody and based on different kinds of evidence. But the general similarity of the resulting structures, as Lerdahl & Jackendoff note, is striking.

3.1 General properties of musical grouping structure

In *GTTM*, grouping structure is inferred on the basis of a listener's intuitions. Those intuitions in turn are based on auditory properties described in section 3.2. Before turning to those properties, however, I outline the general structure of musical groups. As in language, grouping appears to involve sets of constituents nested hierarchically inside larger constituents. Figure 2 shows a possible grouping structure for the traditional folk song 'Turkey in the Straw'. As in figure 1, the grid notation is shown above the example and *GTTM* notation below; the two are formally equivalent.

[FIGURE 2 ABOUT HERE]

Just as in the linguistic example, this grouping structure is not uniquely determined by the properties of the tune. Other listeners could disagree about the exact location of grouping boundaries. And it is possible that different performances of this tune could evoke different grouping structures. But all possible grouping structures tend to share some basic properties.

GTTM states these properties as Grouping Well-Formedness Rules. Several of them are instructive for comparative purposes. Occurring musical events are exhaustively partitioned into groups. If a group *X* contains some of the elements in another group *Y*, then either *X* or *Y* contains all of the elements in the other group (with one exception involving transformation rules). Every piece coincides with a single group that contains all other groups. And there is a fairly strong tendency for groups to contain exactly two smaller groups. These representational properties of musical grouping are inferred from intuitions about constituency and (indirectly) from harmonic interpretation. The reader may notice that they are exactly the same as the properties of linguistic grouping structure discussed in section 2.1. Next, we turn our attention to the nature of the sound patterns that motivate grouping.

3.2 The acoustics of musical grouping

Whereas the mapping between sound patterns and grouping in language tends to be approached as a function from groups to sounds, *GTTM* approaches musical grouping in precisely the opposite direction, as a function from sounds to inferred grouping structures. Despite this reversal, the actual content of the two functions is quite similar. Pearce (2008) offers a detailed and thorough review of grouping theories and empirical results, and some of the review here is based on that discussion.

We saw in section 2.2 that linguistic group boundaries often coincide with longer events and larger changes in f0 and intensity. The same is true in music. *GTTM* posits violable constraints calling for moments of auditory disruption in the musical surface to be aligned with group boundaries. These

constraints are inspired by the Gestalt principles of *proximity* and *similarity*, claimed to be domain-general principles that guide grouping in all modalities (Wertheimer 1938). The proximity principle states that elements that are closer to one another are more likely to be grouped together. The similarity principle states that elements that are more similar to one another are more likely to be grouped together.

The proximity principle entails that notes whose onsets (and to a lesser extent, offsets) are less temporally distant from one another are less likely to be perceived as spanning a group boundary. Similarity principles entail that notes whose pitch, timbre, or intensity (among other properties) are less distinct from one another are less likely to be perceived as spanning a group boundary. In figure 2, for instance, the proximity principle predicts the boundary between the sixth and seventh groups, which falls after a note longer than surrounding ones. Pitch similarity predicts the boundaries between the first, second, and third groups, which occur at local maxima for pitch change. Assessing the effects of intensity and timbre requires an actual performance, rather than a written representation. There are other grouping principles at work in this excerpt, involving inherent properties of groups (binarity, parallelism) and possibly alignment with metrical positions. The claim of *GTTM* is that the groups singled out here are also the constituents relevant to a harmonic or motivic analysis of the excerpt. For instance, all of the groups here end on local *consonances*, notes contained in the local harmony implied by the piece. At a higher level of structure those notes outline a tonic triad progressing to a dominant chord in the first half of the excerpt, then again a tonic triad progressing to a V-I cadence to close the excerpt.

Basic grouping principles have been robustly confirmed for musicians and non-musicians through explicit grouping tasks (Deliège 1987, Peretz 1989) and implicit tasks examining the influence of grouping on memory (Deutsch 1980, Dowling 1973, Tillmann & McAdams 2004). There is evidence that infants use the proximity principle (Jusczyk & Krumhansl 1993, Krumhansl & Jusczyk 1990). And higher-level group boundaries have a cumulative, hierarchical effect on production, perception, or recall (Large *et al*. 1985; Stoffer 1985; Todd 1985).

The edges of perceived musical groups, then, tend to be marked by disruptions in pitch and intensity, and by temporal disjuncture. This system is plausibly a consequence of domain-general Gestalt principles. And it bears an obvious resemblance to the marking of group boundaries in linguistic grouping. In the final section, I compare the two systems in more detail.

4 Parallels and non-parallels between musical and linguistic grouping

The similarities between musical and linguistic grouping outlined in the preceding sections are relatively clear: in both domains, constituents that are relevant to some external domain (syntax, memory, harmonic or semantic interpretation) display relative acoustic continuity internally and disruption at their edges. In both domains, the relevant constituents appear to be hierarchically nested. The types of evidence used to support these conclusions, however, are somewhat different in the two domains, and so are the resulting theories. Theories of linguistic grouping are based on the distribution of sounds in speech production, which are taken to be a product of structural factors. Theories of musical grouping are based on intuitions or chunking in memory, and take groups to be a product of auditory continuity and disruption along broadly Gestalt lines. The precise details of which

acoustic parameters contribute to grouping and how they do so also differ in the two domains. In this section, I examine these differences in greater detail and attempt to draw some general conclusions about the two systems.

4.1 Directionality, production, and perception

In generative linguistics, the *grammar* is taken to be a body of implicit knowledge that includes the principles governing linguistic structure-building. This means that while the grammar influences all aspects of linguistic behavior, it is described not in terms of algorithms that drive specific production and processing activities, but in terms of the 'final state' of information that algorithms for specific processes may draw upon. One common model is a function from arrays of *lexical items* to the set of well-formed utterances in the language in question, including both their sound patterns and truth-conditional semantic interpretations. Lexical items are tuplets stored in long-term memory that capture arbitrary associations between sound, meaning, and syntactic features; at a first pass, they're similar to the everyday notion of *word*. There are various intermediate steps in this function, including a function from syntactic structure to prosodic grouping, and one from lexical representations of sounds in a grouping structure to the pronunciation of those sounds. The *GTTM* model of grouping is different because the entire form of the musical grammar in this theory is different. In *GTTM*, the grammar is also a final-state theory of the information that specific musical processes may call upon. But it is described as a function from (representations of) the auditory stream of musical pieces to a set of metrical, prosodic, and harmonic analyses of those pieces. This is in some ways the opposite of its linguistic counterpart.

One question that immediately arises is whether the directionality in these descriptions is actually necessary. In linguistics, it is generally believed that there is information loss in the mapping from syntactic structures to prosodic ones: grouping is insensitive to some of the distinctions that matter for syntactic constituency. Although theories differ, examples could include some distinctions between syntactic arguments and modifiers, or between structures with more or fewer levels of embedding under function morphemes. In other words, most theories predict that more than one syntactic structure can be mapped to the same grouping structure. That said, the more recent 'recursive' approaches to prosody discussed in section 2.1 entail less information loss between syntactic and grouping structures, possibly none. And at the level of linguistic behavior it is clear that listeners recover the syntactic structure of an utterance in part from information in the auditory stream; acoustic reflexes of grouping are an important part of that process, in precisely the way that Gestalt grouping principles would predict.

At lower levels of syntactic constituency, such as the morpheme or word, there is abundant experimental evidence bearing on the segmentation of both existing and novel lexical items, where 'novel lexical items' are recurring strings of sounds in artificial, unfamiliar pseudo-languages. The detection of such constituents for infant and adult listeners depends in part on transitional probabilities between speech sounds (e.g. McQueen 1998, Mattys & Jusczyk 2001) and between syllables (e.g. Saffran *et al*. 1996a, b): all else being equal, lower-probability transitions are more likely to be inferred to span a group boundary than higher-probability transitions. Related research provides evidence, from artificial and natural languages and from listeners of all ages, that all of the acoustic cues to grouping discussed in section 2.2 can induce segmentation and/or reinforce statistical

cues to improve segmentation (Saffran *et al*. 1996a; Bagou *et al*. 2002; Nakatani & Dukes 1977; Christophe *et al*. 2003; Kim 2004; Millotte *et al*. 2011; Katz & Fricke 2018).

At higher levels of syntactic constituency, cues to prosodic grouping help listeners adjudicate between competing syntactic analyses of similar strings of words (e.g. Price *et al*. 1991; Schafer *et al*. 2000). The precise nature of the cues involved and the details of online processing are a matter of debate (see Carlson *et al*. 2001), but the general fact that a description of sentence processing must make reference to grouping is not in doubt.

So while theories of linguistic grammar generally map from syntactic structure through grouping structure to sound patterns, there is no doubt that listeners 'reverse-engineer' this mapping to recover constituency from the acoustic stream. Beyond this, several theories within the broad tent of generative linguistics propose a more symmetrical relationship between syntactic structure and grouping. Richards (2010, 2016) argues that the description of syntactic computations across languages must make reference to language-specific principles of grouping. Jackendoff (2002) proposes that syntactic and grouping structures are independently generated and relations between them are enforced by correspondence constraints. Steedman (2000) argues that grouping directly reflects the constituency of information structure, with highly flexible relationships between each of these domains and syntactic structure. Taken together, these considerations suggest that the difference in directionality between traditional linguistic theories and the *GTTM* may be largely a matter of convenience, orientation, and/or methodology, rather than reflecting deep differences in the flow of information within grammar.

If the relations between syntax and grouping and between grouping and sound patterns in language are bi-directional, is the same true of music? In *GTTM*, there is a fully transparent relationship between auditory disruption and grouping, and there is no reason that the mapping could not be reversed to derive probabilities of various types of disruption from grouping structure. In terms of grouping and syntax, the theory entails that more than one grouping structure can map to the same syntactic (reductional) structure, but the converse is equally true. And while the majority of the *GTTM* principles concern the mapping from musical surface to grouping and from grouping to syntax, the authors are careful to note that information from the syntactic reduction components is also necessary to adequately describe the grouping system. More generally, *GTTM* differs from most linguistic approaches in that it takes 'the set of well-formed musical pieces' to be either a given or an incoherent concept. As such, the theory describes how pieces of music are assigned structure, but does not attempt to describe why some pieces are more or less likely to exist than others. Steedman (1984), Rohrmeier (2011), and Katz (2017) differ from *GTTM* in this regard, but have little to say about grouping. One major question, then, is whether syntactic constituents in music could be described as generating a probability distribution over possible grouping implementations, as the relationship is generally described in language. Katz & Pesetsky (2009) show that the *GTTM* algorithm governing the relation between time-span reduction (reflecting grouping constituency) and prolongational reduction (reflecting syntactic relations) can be reversed in this way with little or no loss of information. That said, empirical support is somewhat limited for *both* directions of mapping.

Evidence that some independently motivated notion of 'syntactic constituent' in music corresponds in a systematic way to musical grouping is much patchier than for other generalizations discussed here. This is in part because syntactic constituency is far less clear in music than it is in language, and

because musical grouping is rarely approached from the perspective of syntactic structure. There is a body of research showing (sometimes implicitly) that performers tend to elongate the ends of major harmonic sections (e.g. Todd 1985, Repp 1992, 1998). The most substantial collection of relevant analyses may well be the corpus of excerpts analyzed in *GTTM* and Lerdahl's (2001) extension of the theory. These analyses show that, in the default case, the domains relevant to computing syntactic dependencies (prolongational reduction) are *the same* domains relevant to computing rhythmic prominence (time-span reduction). This is despite the fact that the authors frequently choose examples meant to illustrate the complexity of the mapping (in these cases, of course, there are limited degrees of mismatch). It should be understood that the *GTTM* theory of syntax is not universally accepted even for Western tonal music, and there are questions as to whether harmony is even appropriate for syntactic analysis (some genres of music have weak or nonexistent notions of harmony). That said, within harmonic traditions, major structural markers (e.g. cadences, tonic returns, the beginnings of major harmonic sections) tend to be realized in ways that set them apart from surrounding material in terms of grouping. This is such a basic aspect of such genres that it is more likely to be presupposed than actively investigated.

To summarize, the differences between *GTTM* and standard linguistic theory in characterizing syntax-grouping correspondence do not necessarily reflect any deep differences between the two underlying cognitive systems. In language, there is evidence that the mapping from syntax to grouping is less uni-directional than appreciated in the early stages of prosodic theory. And in music, many or all of the mappings from sounds to groups and groups to syntactic constituents are fully reversible. The difference in presentation has more to do with *GTTM*'s overarching goals, which are

different from most linguistic approaches, and with the fact that grouping in music is much easier to intuit (and more widely agreed upon) than syntactic structure.

4.2 Acoustics and prominence

Similarities in the acoustic reflexes of musical and linguistic grouping are striking enough to raise the question of whether the same grouping algorithm could be applied to auditory objects in the two domains. I think the answer is probably not, and the reasons why tell us something interesting about basic 'design principles' of music and language.

While theorists disagree on precisely which aspects of musical events are crucial to computing combinatoric (syntactic) dependencies, all theories share one broad commonality: the categories relevant to syntax severely underdetermine actual acoustic realizations. For instance, in a theory where chords are the basic building blocks of syntax, there is an infinite number of ways that a given chord can be performed: more or fewer notes, longer or shorter duration, higher or lower pitch and intensity, faster or slower attack, etc. As all of these acoustic parameters are freely varied by composers and performers to demarcate groups, grouping algorithms do fairly well by simply scanning a representation of musical notes and locating acoustic discontinuities. For instance, Thom *et al*. (2002) show that a variety of simple grouping algorithms based on very few acoustic parameters produce good agreement with human annotators, even using a measure that ignores a large portion of the agreeing cases (where neither humans nor models infer a group boundary).

To simplify somewhat, a listener who hears an acoustic discontinuity in music can be relatively confident that it marks a group boundary. In language, on the other hand, there are a number of sources for auditory discontinuity besides grouping *per se*. Perhaps the most obvious is metrical prominence, generally referred to as *stress* at the level of words and *accent* at higher levels. The presence and absence, position, and acoustic implementation of stress and accent all vary across languages. But the most frequent acoustic parameters of prominence are precisely those used for marking group boundaries: pitch extrema, changes in intensity (including changes in specific frequency bands), and longer duration (see Ortega-Llebaria & Prieto 2010 for a concise review). This means that acoustic discontinuities in the speech stream may correspond to prosodic group boundaries or they may correspond to a prominent syllable (among other things). So at a bare minimum, any grouping algorithm for a language will need to be supplemented with language-specific principles that help relativize acoustic disruption to the metrical prominence of the material being parsed.

Why doesn't this issue arise as much in music? One reason is that metrical prominence in music tends to be highly regular. While patterns of metrical prominence must be inferred from the acoustic stream, once a local metrical pattern is established it is unlikely to change during the course of a musical piece. As such, there is less need to mark metrical prominence with acoustic changes (though there is some tendency to do so, e.g. Palmer & Krumhansl 1990 on the occurrence of musical events by level of metrical prominence). In language, on the other hand, metrical prominence is less predictable in the general case. Many languages have unpredictable stress placement within a word (see van der Hulst & Gordon to appear for a general review of stress). Even in languages where stress has been described as fully predictable and alternating in a regular pattern, closer inspection often

reveals that the morphological composition of a word can introduce departures from regularity (see Baker 2014 for an overview of Australian languages). Regardless of the level of metrical regularity within words, the fact that different words contain different numbers of syllables in and of itself entails that metrical regularity will be weakened or absent at the level of phrases and utterances. As such, languages with stress and/or accent virtually always mark its location using one of the acoustic parameters discussed here. Or at the very least, if a language didn't mark prominence in one of these ways, it is unlikely that linguists (or infants learning the language) would notice the prominence.

Another reason why acoustic discontinuity doesn't necessarily entail group boundaries in language pertains to a basic property of linguistic sound systems: many languages use duration, pitch, and intensity to mark contrasts in lexical meaning. In English, for example, intensity is one of the most obvious differences between obstruents like /k/, /b/, and /z/ and sonorants like /m/, /l/, and /w/ (Ladefoged & Johnson 2011). Duration is a strong cue listeners use to discriminate voiced and voiceless fricatives (Cole & Cooper 1975). And the perception of voicing contrasts for consonants is affected by f0 values at the beginning of a following vowel (Haggard *et al*. 1970). Another way of putting this is that in English, an abrupt drop in intensity, raised F0, or long duration of noise could just as easily be caused by a voiceless fricative as a group boundary. This is despite the fact that English duration and pitch are not generally considered to be primary dimensions of contrast; in other languages, they clearly are primary and would be expected to play an even larger role in discriminating speech sounds.

This means that the amount of acoustic disruption or change relevant to inferring group boundaries must be relativized not only to the metrical prominence of the linguistic material in question but also

to its lexical segmental makeup. These two properties suffice to make the mapping between acoustics and grouping a fair bit more complex in language than in music. And they both stem ultimately from one of the most profound differences between music and language: the presence of a lexicon. Speakers possess implicit knowledge about the meaningful parts (morphemes) of their languages, involving at least arbitrary pairings of sounds and meanings in long-term memory. The sounds corresponding to any lexical item in any particular language in the general case have some internal temporal structure: *cat* is one syllable long in English but *feline* is two; *cat* is pronounced with two tongue body gestures, one rising to create a constriction at the velum and another associated with a low front vowel, and these two gestures occur in a fixed order. This means that in addition to grouping meaningful morphemes into words, phrases, and sentences, human languages also group meaningless sound features into those meaningful units. The property is referred to as *duality of patterning*; it has been characterized as a basic 'design feature' of human language (Hockett 1960; see Ladd 2012 for an overview and some complications), and something that sets humans apart from other animals. If nothing else, music may show that duality of patterning is not necessary to generate extremely complex symbolic systems that unfold in time.

In sum, while the relationship between acoustics and grouping is similar in music and language, the ways in which this relationship guides processing must be somewhat different in the two domains. The basic building blocks of linguistic syntax are themselves temporally complex with regard to the number and nature of the speech sounds they contain, properties which are memorized in the lexicon. This duality of patterning means that there are more independent factors contributing to acoustic continuity and disruption in language than there are in music. The question then arises: why does language displays a rich lexicon and dual patterning while music does not? There is no firm answer,

but there is a common intuition that a rich lexicon is necessary to express truth-conditional meanings with any degree of specificity, and that recombination of meaningless elements is necessary for a lexicon of any substantial size. On this view, some rather intricate and complex differences between music and language can be traced to the lexicon and, ultimately, differences in communicative function.

4.3 Conclusion

There are several areas of similarity between grouping in music and language. With regard to acoustics, both domains make use of something like Gestalt principles of proximity and similarity. With regard to grouping structure itself, both domains involve hierarchically nested constituents. And with regard to information-exchange with other cognitive systems, both display a systematic (if noisy) correspondence with syntactic or semantic constituency. The final question I ask here is *why* such similarities exist, and what they mean for our theories of music, language, and cognition more generally.

There is a fairly clear intuition about why Gestalt principles work the way they do. In general, grouping inferences based on proximity and similarity are relatively likely to accurately reflect the sources of sounds in the environment (e.g. Deutsch 1999). For instance, two sequences of sounds separated by a pause are more likely to come from two different sources than two sequences not separated by a pause. The same is ostensibly true for sequences separated by a discontinuity in pitch, intensity, etc. This general form of reasoning is not limited to the auditory modality, and neither are the proposed principles of Gestalt grouping. Wertheimer (1938) famously applied them to visual

arrays of shapes. Signed languages implement prosodic groups visually using the proximity rule (mainly for manual signs) and possibly similarity rules (for non-manual signs; see Fenlon & Brentari to appear for an overview). Charnavel (2016) argues that the structure of dance also makes use of Gestalt visual grouping. And Spelke (1994) explains why certain principles of spatial cognition and object recognition, some of which are related to proximity and similarity, are likely to give rise to ecologically valid inferences about objects in the world.

All of these considerations suggest that grouping principles should be widely applicable across a variety of perceptual domains and modalities. So are grouping principles 'the same' in music and language? There is a sense in which they are and a sense in which they aren't. The basic principles that guide grouping in the two domains are based on the same types of information and may ultimately be rooted in properties of the environment in which human perception occurs. That said, how the principles are *deployed* in the two domains is rather different. We noted in section 4.2 that the likelihood that any given acoustic disruption marks a group boundary in language must be compared to the likelihood that it marks something else, such as a distinction in metrical prominence or segmental features. This is less likely to matter in music (though not entirely irrelevant). Beyond this, while the *form* of grouping principles may be 'given' by general Gestalt principles, learning the grouping conventions of any genre of music or language necessarily involves assigning different weights to different acoustic cues. These weights differ quite a bit between languages, and there is no reason they shouldn't vary between genres of music as well. So a second sense in which musical and linguistic grouping might be 'the same' is if the weighting of cues in one domain affects the weighting of cues in the other. Iverson *et al*. (2008) argue that linguistic grouping affects non-linguistic auditory grouping in precisely this way, although that argument is based on questionable

claims about grouping and acoustics in Japanese and English. Some subsequent research replicates the finding that language experience affects non-linguistic grouping (Bhatara *et al*. 2016; Molnar *et al*. 2016), although these effects are not robust across different stimuli and tasks (Frost *et al*. 2017; Langus *et al*. 2016). The majority of these experiments concern only repeating binary patterns, and it would be imprudent to draw from them broad conclusions about the relationship between music and language, but they do highlight an interesting type of question.

There are additional questions about the domain-specificity or domain-generality of grouping. On one view, the fact that grouping involves principles of audition independent of language means that it is not part of the narrow language faculty. While researchers are free to define technical terms as they see fit, it does seem to me that this notion of 'language faculty' is so narrow that it will fail to include the vast majority of all interesting cognitive processes involved in language, and may not include much of *anything* in the end. On the other hand, it is undoubtedly important, when asking about perceptual resources that music and language share, to bear in mind that those resources may also be shared with an array of perceptual processes in other domains and even other modalities. So there is good evidence that language and music are deeply similar with regard to grouping, but not that they are deeply different from other cognitive domains in this regard.

The presence of hierarchical structure is arguably less general and more difficult to explain than Gestalt grouping. One common view in language is that grouping 'inherits' its hierarchical structure from syntax, although there is significant disagreement on this point (and even less agreement about why *syntactic* structure is hierarchical). If the same explanation is to be extended to music, it entails that *GTTM* must be 'reversed' along the lines discussed in section 4.1, and also that we accept the

model's characterization of musical syntax as hierarchically structured (see Katz 2017 and Temperley 2011 for arguments *pro* and *contra*, respectively). On this view, the structural similarities between musical and linguistic grouping emerge from the fact that both involve translating between the hierarchical graph structure that represents an utterance or piece of music and the temporal string of events that must be used to convey that structure from one organism to another. Hierarchically nested grouping should be shared with any cognitive domain that involves communication of hierarchically-structured representations through some sensory modality over time, including signed language and possibly dance in the visual modality. If some temporally complex cognitive activities lack hierarchical syntactic structure, there is no particular reason they should display hierarchical grouping. That said, it is quite difficult in practice to identify temporally complex cognitive activities that demonstrably lack hierarchical syntax. In principle, this view makes grouping relatively domain-specific, but with little evidence from external domains or even specifications of which domains *are* external.

Another possibility is that grouping hierarchy arises for reasons intrinsic to meter and rhythm. The study of temporal regularities at multiple timescales in language (e.g. Cummins & Port 1998, Tilsen 2009) and music (e.g. Jones & Boltz 1989) has led to independent suggestions in the two domains that production and perception can be described with systems of hierarchically coupled oscillators. On this view, grouping is hierarchical because it is instantiated in individual brains, and brains are organized in terms of hierarchically coupled oscillators (see Hauk *et al*. 2017 for an overview oriented towards language). As such, hierarchical grouping should be shared with all forms of motor control and temporally-modulated attending (see Tilsen 2009 for review of some relevant motor-

control literature). In principle, this view would make grouping relatively domain-general, but again, evidence from grouping structure in a broad array of cognitive domains is not easy to find.

In the end, then what do we gain from the comparative study of computational-level musical and linguistic cognition? One answer is that simply attempting to align theories in the two domains helps clarify our thinking about each of them, especially at the 'architectural' level of information flow between components and the functional level of explaining why information in these domains is structure the way it is instead of some other way. A second answer is that, to the extent we can isolate particular similarities and differences in the information states that underlie musical and linguistic cognition, those properties point to potentially fruitful areas of inquiry in other cognitive domains. And a final, optimistic answer is that any similarities may reveal deep cognitive properties rooted in evolution that distinguish human beings from other species. Regardless of whether music and language turn out to be closely related at the level of computation, behavior, brain, or evolution, however, it is surely worth asking these 'big-picture' questions in the most informed way possible.

Literature cited

Bagou O, Fougeron C, Frauenfelder U. 2002. Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. Presented at Speech Prosody, Aix-En-Provence.

Baker B. 2014. Word structure in Australian languages. In *The Languages and Linguistics of Australia: A comprehensive guide*, ed. H. Koch, R. Nordlinger, pp. 139-213 . Berlin: De Gruyter.

Beckman M, Pierrehumbert J. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3: 255-309.

Bhatara A, Boll-Avetisyan N, Agus T, Höhle B, Nazzi T. 2016. Language experience affects grouping of musical instrument sounds. *Cognitive Science* 40: 1816-1830.

Borgman, Donald. 1990. Sanuma. In Desmond Derbyshire & Geoff Pullum (eds.), *Handbook of Amazonian Languages, Vol. II*, 15-248. Berlin: Mouton de Gruyter.

Byrd D, Saltzman E. 1998. Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics* 26: 173-199.

Cambouropoulos E. 2001. The local boundary detection model and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference*, pp. 17-22. San Francisco: ICMA.

Carlson K, Clifton C, Frazier L. 2001. Prosodic boundaries in adjunct attachment. *Journal of Memory and Language* 45: 58-81.

Charnavel I. 2016. Steps towards a Generative Theory of Dance Cognition. LingBuzz: lingbuzz/003137

Christophe A, Gout A, Peperkamp S, Morgan J. 2003. Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics* 31: 585-598.

Cole R, Cooper W. 1975. Perception of voicing in English affricates and fricatives. *Journal of the Acoustical Society of America* 58: 1280-1287.

Cummins F, Port R. 1998. Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26: 145-171.

De Jong K, Zawaydeh BA. 1999. Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics* 27: 3-22.

Deliège I. 1987. Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's grouping preference rules. *Music Perception* 4: 325–360.

Dell F. 2015. Text-to-tune alignment and lineation in traditional French songs. In *Text and Tune*, ed. T Proto, P Canettieri, G Valenti, pp. 183-234. Bern: Peter Lang.

Deutsch D. 1980. The processing of structured and unstructured tonal sequences. *Perception and Psychophysics* 28: 381–389.

Deutsch D. 1999. Grouping mechanisms in music. In *Psychology of Music*, ed. D Deutsch, pp. 299-348. San Diego: Academic Press.

Dilley L, Shattuck-Hufnagel S, Ostendorf M. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24: 423-444.

Dowling WJ. 1973. Rhythmic groups and subjective chunks in memory for melodies. *Per- ception and Psychophysics* 14: 37–40.

Duanmu S. 1996. Pre-juncture lengthening and foot binarity. *Studies in Linguistic Sciences* 26: 95-115.

Ennever T, Meakins F, Round E. 2017. A replicable acoustic measure of lenition and the nature of variability in Gurindji stops. *Laboratory Phonology* 8: 1-32.

Fenlon J, Brentari D. To appear. Sign language prosody. In *Routledge Handbook of Theoretical and Experimental Sign Language Research*, ed. J. Quer, R. Pfau, A. Herrmann. Abingdon, UK: Routledge.

Féry C, Truckenbrodt H. 2005. Sisterhood and tonal scaling. *Studia Linguistica* 59: 223-243.

Féry C. 2010. Recursion in prosodic structure. *Phonological Studies* 13: 51-60.

Fitch T. 2015. The biology and evolution of musical rhythm: an update. In *Structures in the Mind: Essays on Language, Music, and Cognition in Honor of Ray Jackendoff*, ed. I Toivonen, P Csúri, E Van Der Zee, pp. 293-324. Cambridge, Mass.: MIT Press.

Fougeron C, Keating P. 1997. Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America* 101: 3728-3740.

Frost R, Monaghan P, Tatsumi T. 2017. Domain-general mechanisms for speech segmentation: The role of duration information in language learning. *Journal of Experimental Psychology: Human Perception and Performance* 43: 466-476.

Gordon M, Munro P. 2007. A phonetic study of final vowel lengthening in Chickasaw. *International Journal of American Linguistics* 73: 293-330.

Gurevich N. 2003. *Functional constraints on phonetically conditioned sound changes*. PhD thesis University of Illinois, Urbana-Champaign.

Haggard M, Ambler S, Callow M. 1970. Pitch as a voicing cue. *Journal of the Acoustical Society of America* 47: 613-617.

Hauk O, Giraud A, Clarke A. Brain oscillations in language comprehension. *Language, Cognition and Neuroscience* 32: 533-535.

Hayes B, Lahiri A. 1991. Bengali intonational phonology. *Natural Language and Linguistic Theory* 9: 47-96.

Hayes B. 1989. The prosodic hierarchy in meter. *Phonetics and Phonology* 1: 201-260.

Heffner C, Slevc R. 2015. Prosodic structure as a parallel to musical structure. *Frontiers in Psychology* 6: 1962.

Hockett C. 1960. The origin of speech. *Scientific American* 203: 88-111.

Hyman L. 2013. Penultimate lengthening in Bantu. In *Language Typology and Historical Contingency*, ed. B. Bickel, L. Grenoble, D. Peterson, A. Timberlake, pp. 309-330. Amsterdam: John Benjamins.

Ishihara S. 2003. *Intonation and Interface Conditions*. PhD thesis, MIT, Cambridge, Mass.

Iverson J, Patel A, Ohgushi K. Perception of rhythmic grouping depends on auditory experience. *Journal of the Acoustical Society of America* 124: 2263-2271.

Jackendoff R. 2002. *Foundations of Language*. Oxford: Oxford University Press.

Johnson K, Martin J. 2001. Acoustic vowel reduction in Creek: Effects of distinctive length and position in the word. *Phonetica* 58: 81-102.

Jones MR, Boltz M. 1989. Dynamic attending and responses to time. *Psychological Review* 96: 459-491.

Jun S. 1993. *The phonetics and phonology of Korean prosody*. PhD thesis, Ohio State University, Columbus.

Jusczyk P, Krumhansl C. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of Experimental Psychology: Human Perception and Performance* 19: 627–640.

Katz J. 2016. Lenition, perception, and neutralisation. *Phonology* 33: 43-85.

Katz J. 2017. Harmonic syntax of the 12-bar blues: A corpus study. *Music Perception* 35: 165-192

Katz J, Fricke M. 2018. Auditory disruption improves word segmentation: A functional basis for lenition phenomena. To appear in *Glossa*.

Katz J, Pesetsky D. 2009. The Identity Thesis for language and music. *LingBuzz*: lingbuzz/000959

Keating P, Cho T, Fougeron C, Hsu C. 2003. Domain-initial strengthening in four languages. In *Phonetic interpretation: Papers in laboratory phonology VI*, ed. J Local, R Ogden, R Temple, pp. 143-161. Cambridge, UK: Cambridge University Press.

Kim S. 2004. *The Role of Prosodic Phrasing in Korean Word Segmentation*. PhD thesis, University of California, Los Angeles.

Kingston J. 2008. Lenition. In *Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology,* ed. L Colantoni, J Steele, pp. 1-31. Somerville: Cascadilla.

Kirchner R. 1998. *An effort-based approach to consonant lenition*. PhD thesis, University of California, Los Angeles.

Krumhansl C, Jusczyk P. 1990. Infants' perception of phrase structure in music. *Psychological Science* 1: 70–73.

Ladd DR. 1986. Intonational phrasing: the case for recursive prosodic structure. *Phonology Yearbook* 3: 311-340.

Ladd DR. 2012. What *is* duality of patterning, anyway? *Language and Cognition* 4: 261-273.

Ladefoged K, Johnson K. 2011. *A Course in Phonetics*. Boston: Wadsworth.

Langus A, Seyed-Allaei S, Uysal E, Pirmoradian S, Marino C, Nespor M. 2016. Listening natively across perceptual domains? *Journal of Experimental Psychology: Learning, Memory and Cognition* 42: 1127-1139.

Large EW, Palmer C, Pollack JB. 1995. Reduced memory representations for music. *Cognitive Science* 19: 53–96.

Lavoie L. 2001. *Consonant Strength: Phonological Patterns and Phonetic Manifestations*. New York: Garland.

Lerdahl F, Jackendoff R. 1983. *A Generative Theory of Tonal Music*. Cambridge, Mass.: MIT Press.

Lerdahl F. 2001. *Tonal Pitch Space*. Oxford: Oxford University Press.

Marr D. 1982. *Vision*. Cambridge, Mass.: MIT Press.

Mattys SL, Jusczyk P. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78: 91-121.

McCawley J. 1968. *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.

McQueen JM. 1998. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language* 39: 21-46.

Millotte S, Morgan J, Margules S, Bernal S, Dutat M, Christophe A. 2011. Phrasal prosody constrains word segmentation in French 16-month-olds. *Journal of Portuguese Linguistics* 9: 67-86.

Molnar M, Carreiras M, Gervain J. 2016. Language dominance shapes non-linguistic rhythmic grouping in bilinguals. *Cognition* 152: 150-159.

Morén B, Zsiga E. 2006. The lexical and post-lexical phonology of Thai tones. *Natural Language & Linguistic Theory* 24: 113-178.

Nagano-Madsen Y. 1992. Mora and prosodic coordination: A phonetic study of Japanese, Eskimo, and Yoruba. Lund: Lund University Press.

Nakatani L, Dukes K. 1977. Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America* 62: 714-719.

Narmour E. 1990. *The Analysis and Cognition of Basic Melodic Structures: The Implication-realisation Model*. Chicago: University of Chicago Press.

Nespor M, Vogel I. 1986. *Prosodic Phonology*. Dordrecht: Foris.

Nordhoff S. 2012. Synchronic grammar of Sri Lanka Malay. In *The Genesis of Sri Lanka Malay: A case of extreme language contact*, ed. S Nordhoff, pp. 13-52. Leiden: Brill.

Oller D. 1979. Syllable timing in Spanish, English, and Finnish. In *Current Issues in the Phonetic Sciences*, ed. P. Macneilage, pp. 189-216. New York: Springer.

Onaka A, Palethorpe S, Watson C, Harrington J. 2003. Acoustic and articulatory difference of speech segments at different prosodic positions. In *Proceedings of the 9th Australian International Conference on Speech Science and Technology*.

Ortega-Llebaria M, Prieto P. 2010. Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Language and Speech* 54: 73-97.

Palmer C, Krumhansl C. 1990. Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance* 16: 728-741.

Pearce M. 2008. The perception of grouping boundaries in music. Unpublished manuscript, Queen Mary University, London.

Peretz I. 1989. Clustering in music: An appraisal of task factors. *International Journal of Psychology* 24: 157–178.

Pierrehumbert J, Beckman M. 1988. *Japanese tone structure*. Cambridge, Mass.: MIT Press.

Pierrehumbert J, Hirschberg J. 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*, ed. P Cohen, J Morgan, M Pollack, pp. 271-311. Cambridge, Mass.: MIT Press.

Pierrehumbert J, Talkin D. 1992. Lenition of /h/ and glottal stop. In *Papers in Laboratory Phonology II*, ed. G Docherty, DR Ladd, pp. 90-117. Cambridge, UK: Cambridge University Press.

Pierrehumbert J. 1980. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Price PJ, Ostendorf M, Shattuck-Hufnagel S, Fong C. 1991. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America* 90: 2956–2970.

Redi L, Shattuck-Hufnagel S. 2001. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29: 407–429.

Repp BH. 1992. Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America* 92: 2546-2568.

Repp BH. 1998. A Microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America* 104: 1085-1100.

Richards N. 2010. *Uttering Trees*. Cambridge, Mass.: MIT Press.

Richards N. 2016. *Contiguity Theory*. Cambridge, Mass.: MIT Press.

Rohrmeier M. 2011. Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music* 5: 35-53.

Rohrmeier M, Zuidema W, Wiggins G, Scharff C. 2015. Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B* 370: 20140097.

Saffran J, Newport E, Aslin R. 1996a. Word segmentation: The role of distributional cues. *Journal of Memory & Language* 35: 606-621.

Saffran J, Aslin R, Newport E. 1996b. Statistical learning by 8-month-old infants. *Science* 274: 1926-1928.

Schafer A, Speer S, Warren P, White SD. 2000. Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research* 29: 169–182

Selkirk E. 1972. *The phrase phonology of English and French.* New York: Garland.

Selkirk E. 1984. *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, Mass.: MIT Press.

Selkirk E. 2011. The syntax-phonology interface. In *The Handbook of Phonological Theory*, ed. J Goldsmith, J Riggle, A Yu, pp. 435-483. Oxford: Blackwell.

Shattuck-Hufnagel S, Turk A. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25: 193-247.

Spelke E. 1994. Initial knowledge: six suggestions. *Cognition* 50: 431-445.

Steedman M. 1984. A generative grammar for jazz chord sequences. *Music Perception* 2: 52-77.

Steedman M. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry* 31: 649-689.

Stoffer TH. 1985. Representation of phrase structure in the perception of music. *Music Perception* 3: 191–220.

Temperley D. 2011. Composition, perception, and Schenkerian theory. *Music Theory Spectrum* 33: 146-168.

Thom B, Spevak C, Höthker K. 2002. Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proceedings of the 2002 International Computer Music Conference*. San Francisco: ICMA.

Tillmann B, McAdams S. 2004. Implicit learning of musical timbre sequences: Statistical regularities confronted with acoustic (dis)similarities. *Journal of Experimental Psychology: Learning, Memory and Cognition* 30: 1131–1142.

Tilsen S. 2009. Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science* 33: 839-879.

Tingsabadh MRK, Abramson A. 1993. Thai. *Journal of the International Phonetic Association* 23: 24-28.

Todd N. 1985. A model of expressive timing in tonal music. *Music Perception* 3: 33-57.

Turk A, Shattuck-Hufnagel S. 2000. Word-boundary-related duration patterns in English. *Journal of Phonetics* 28: 397-440.

Van der Hulst H, Gordon M. To appear. Word stress. In *The Oxford Handbook of Language Prosody*, ed. C Gussenhoven, A Chen. Oxford: Oxford University Press.

Wagner M. 2005. *Prosody and Recursion*. PhD Thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Wertheimer M. 1938 [English translation of 1923 essay]. Laws of Organization in Perceptual Forms. In *A source book of Gestalt psychology*, ed. W Ellis, pp. 71-88. London: Routledge.

Wightman C, Shattuck-Hufnagel S, Ostendorf M, Price P. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91: 1707-1717.

Figure Captions

Figure 1. A possible prosodic (grouping) realization of the English utterance *the child in front discovered an object*. The same grouping structure is shown in bracket notation (above the text) and Lerdahl & Jackendoff's (1983) grouping notation (below the text).

Figure 2. A possible grouping structure for the folk song *Turkey in the Straw*. The same grouping structure is shown in bracket notation (above the music) and Lerdahl & Jackendoff's (1983) grouping notation (below the music).

Figures

Figure 1

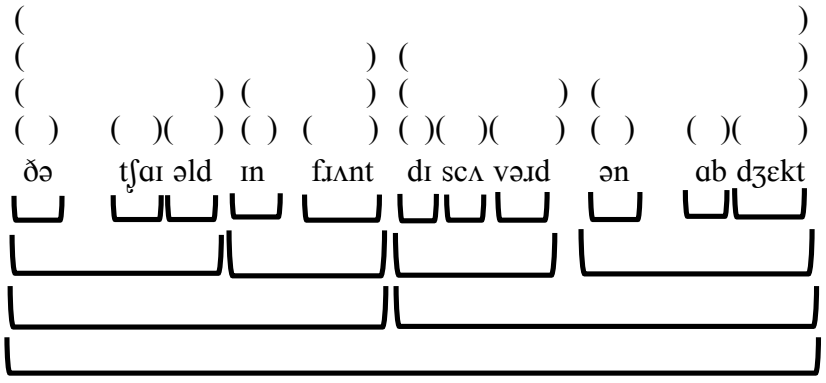

Figure 2