Ann. N.Y. Acad. Sci. ISSN 0077-89

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *Speech Rhythm in Ontogenic, Phylogenetic, and Glossogenetic Development*
REVIEW

Space and time in models of speech rhythm

Sam Tilsen
Department of Linguistics, Cornell University, Ithaca, New York

Address for correspondence: Sam Tilsen, Department of Linguistics, Cornell University, 203 Morrill Hall, Ithaca, NY 14853. tilsen@cornell.edu

How do rhythmic patterns in speech arise? There are many representational frameworks for describing rhythmic patterns, but none of these directly connect representations to articulatory processes, which have physical manifestations in the acoustic signal. Here, a new model of speech rhythm is presented, one in which rhythmic patterns arise from spatial mechanisms that govern the organization of articulatory gestures. The roles of time and space in symbolic representations of the metrical structure are analyzed, and conventional understandings of stress and accent are called into question. One aspect of rhythmic patterns, in particular—the directionality of accentual patterns—is examined closely. A novel dynamical model is developed, which proposes a reinterpretation of directionality and other temporal phenomena in speech.

Keywords: speech rhythm; metrical theory; rhythmic typology; prosody; dynamical models

Introduction

In many languages, words conform to a pattern in which some syllables can occur with an accent while others cannot. These accents—often a change of pitch, loudness, and/or duration—may have the effect of grabbing the attention of a listener, facilitating word identification, and potentially creating a rhythm, i.e., a pattern that repeats in time. Curiously, the pattern is typically predictable only from the beginning or only from the end of the word. In conventional terms, there is a directionality parameter for stress assignment and stress is assigned "from the left edge" or "from the right edge" of the word. Schematic examples of stress patterns with left-to-right (L→R) and right-to-left (R→L) directionality are contrasted in Table 1.

The spatial vocabulary (i.e., *left*, *right*, and *edge*) is somewhat odd, because words do not really have edges, and because the mapping of left/right to earlier/later is arbitrary. To say that words have "edges" relies on a metaphor in which time is a linear space and syllables are objects arranged in that space. This makes sense from a naïve view of speech because graphemes are spatially arranged in our writing systems, in a way that corresponds to their temporal order in production. Hence, one can say that there are syllables in a word that are "at the left edge" or "at the right edge." The observation that a spatial vocabulary for describing rhythmic patterns is useful, despite the fact that words, as produced in speech, are not spatial entities, we refer to as *the puzzle of directionality*. The spatial vocabulary could certainly be motivated if there is a spatial mapping of the components of words to a physical space in the brain. But is there really a space of this sort? This article explores the idea that such a space indeed exists.

A curious aspect of directionality is that the distribution of L→R and R→L patterns across languages is relatively balanced.[1] Moreover, specific L→R patterns observed in one language typically have symmetric R→L counterparts in another language. These symmetries might be unexpected because time is asymmetric: causes precede effects, and entropy always increases. Many morphological, phonological, and phonetic patterns do indeed reflect an "arrow of time." Suffixation is more prevalent than affixation and suffixes tend to be more tightly bound to roots than prefixes.[2] Word-initial strengthening and word-final weakening are more common than their counterparts.[3–6] Lexical

doi: 10.1111/nyas.14102

Table 1. Schematic comparison of stress patterns with L→R and R→L directionality

Number of σ	L→R	R→L
1	**σ**	**σ**
2	**σ** σ	σ **σ**
3	**σ** σ **σ**	**σ** σ **σ**
4	**σ** σ **σ** σ	σ **σ** σ **σ**
5	**σ** σ **σ** σ **σ**	**σ** σ **σ** σ **σ**

Note: **σ**, stressed syllable; σ, unstressed syllable.

access/retrieval appears to privilege earlier sounds over later ones,[7] and in tip-of-the-tongue states speakers are often aware of just the first sound or first few sounds in a word.[8,9] Furthermore, aerodynamic effects lead to the decrease of fundamental frequency over the course of an utterance, so that pitch tends to be lower later on in utterances,[10,11] and the initiations of articulatory movements precede the achievements of movement targets, an obvious but nonetheless consequential fact.[12] Since many speech-related phenomena exhibit temporal asymmetries, one might wonder why there are not similar asymmetries in the possible directionality of stress assignment. As we will eventually see, if accentuation is understood to originate from spatial patterns in the brain, the symmetry of directionality is not unexpected.

The main aims of this article are (1) to analyze the role of spatiotemporal reasoning in phonological representations of rhythmic structure in words, and (2) to describe a novel dynamical model in which temporal patterns emerge from spatial patterns that govern the organization of articulatory gestures. The focus here is on rhythmic patterns *within words*, i.e., lexical stress/accent, rather than phrasal accentual patterns. Furthermore, a primarily motoric and developmental perspective is adopted; consequently, issues related to the perception of rhythm and phenomena emerging from interactions between agents are not discussed. These restrictions of scope are useful because a detailed model of the system that gives rise to rhythmic patterns on short timescales for one speaker may be a prerequisite to understanding the perceptual, social, and historical forces that influence rhythmic patterns on longer timescales for multiple speakers and in larger domains such as phrases. The novelty of the proposed model is that it views rhythm in spontaneous conversational speech as a consequence of a spatial organization of articulatory movements, thereby providing a missing link between directionality in representations of rhythmic patterns and mechanisms for controlling speech.

The paper is organized as follows. First, we clarify how the terms accent and stress are understood in the current approach and argue that stress should be viewed as a purely structural phenomenon that does not have direct articulatory or acoustic correlates. We then contrast various symbolic representations of accentual patterns with gestural scores, which are dynamic representations of forces that govern articulatory movements. A typological classification of word-level accentual patterns is presented, and we examine how a model of accentual patterns developed by Goldsmith[13] can be reinterpreted in the gestural score framework by introducing a wave/field model of gestural organization. The wave/field model is first shown to generate quantity-insensitive accentual patterns in the typological classification and subsequently is applied to quantity-sensitive patterns and additional phenomena such as the rhythm rule and durational lengthening associated with accent. The paper concludes with a discussion of the advantages of the proposed model and its relation to previous approaches.

What is accent and what is stress?

Any adequate understanding of speech rhythm requires a conception of how variation in the acoustic prominence of the speech signal is controlled by speakers. Variation in prominence is often associated with the terms *stress* and *accent*, but there is substantial diversity in the literature regarding how these terms are used and what they refer to. A current, influential view is that stress is a structural property of syllables, which determines where accents may occur, and that accents are manifested through articulatory control over variables such as pitch, intensity, or phonation quality. This view is consistent with the structural representation in Figure 1, where syllables are labeled as strong (σ_s) or weak (σ_w), and an H* accent (i.e., a high pitch gesture) is associated with a stressed syllable. The model that we ultimately develop is mostly consistent with this view but elaborates on both the nature of stress (i.e., the structural organization)

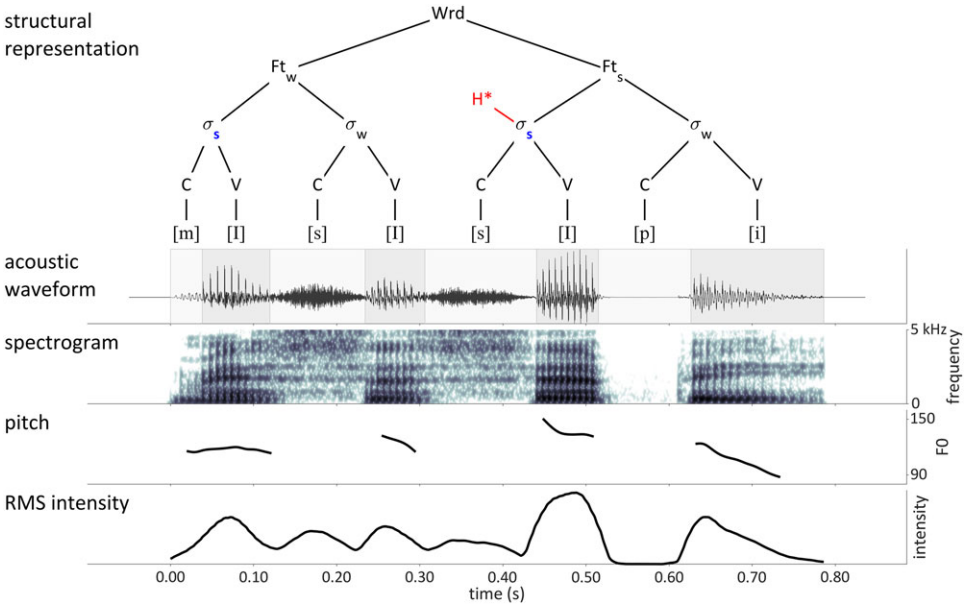

Figure 1. **Structural representation of stress and accent in the word Mississippi, along with various forms of acoustic information. From top to bottom: structural representation of a prosodic word with two feet and an accent associated with the penultimate syllable; acoustic waveform with segmentation; spectrogram (from 0 to 5000 Hz), fundamental frequency/pitch (F0); root-mean-square intensity.**

and the nature of accent, which relates to the control of pitch, intensity, phonation quality, and other characteristics of speech that can enhance or reduce acoustic prominence.

Accent (or accentuation) is understood here as the phenomenon whereby acoustic prominence is controlled, and accents are articulatory events, which accomplish that control. Accents are often observed to involve changes of pitch, and hence it is not uncommon for researchers to think of accents as only a matter of pitch. For example, the H* in Figure 1 is manifested as a higher F0 in the third syllable relative to other syllables in the word. However, a more general conception of accent is preferable, in which accents may effect variation not only in pitch but also in acoustic intensity (also shown in Fig. 1), segmental duration, spectral tilt (reflecting vocal fold tension), and articulatory kinematics (i.e., movement ranges/velocities/targets). We refer to these variables as acoustic/articulatory *correlates of accent*, but it is important to emphasize that the correlates of accent do not necessarily co-occur in a given language or utterance context: in some languages, only a subset of these variables may associate with accents, and from utterance to utterance, these associations are statistical rather than deterministic. Thus, there are no universal correlates of accent.

A highly relevant point about accent is that its acoustic/articulatory correlates are manifested gradiently, rather than categorically. For example, it is possible to produce an emphatic focus accent with very subtle or very drastic acoustic/articulatory effects, and all ranges in between. Phonological representations, which depict accents as present or absent, do not provide a direct account of such variation. Indeed, studies of prominence perception do not support a categorical view of accent: listeners appear to use a variety of acoustic cues as well as lexical information to assess syllable prominence, and the extent to which naïve listeners agree in their assessments of prominence is far below what would be expected if accents were always clearly present or absent.[14]

Although accents are events controlled through articulatory mechanisms, stress should not be understood as an articulatory phenomenon. Rather, stress relates solely to a structural organization of syllables, of the sort exemplified in Figure 1. To facilitate exposition of this point, it is helpful to

distinguish between stress in a structural sense and stress in a featural sense. In the featural sense, stress is viewed as a phenomenon that is independent of accent and has particular acoustic/articulatory manifestations, just as accent does. It is this featural sense of stress that we reject here. To motivate this perspective, we consider several facts about stress and accent.

First, it is uncontroversial that accents can only occur with syllables that are stressed in the structural sense,[10,11] and this is an important clue that stress and accent are not, in fact, independent phenomena. If stress and accent were distinct articulatory phenomena, then we would expect their distributions to be at least partly independent. Instead, accents only occur on structurally stressed syllables.

Second, research on the "phonetic correlates of stress" has found that, just as with accent, the magnitudes of acoustic/articulatory correlates vary gradiently and there are no universal correlates of stress;[1] pitch, intensity, segmental duration, spectral tilt, and articulatory kinematic variables appear to correlate with stress in a language-specific and contextually contingent manner. Thus, when it comes to acoustic and articulatory correlates, the "effects" of stress are no different from those of accent, as defined above.

Third, early empirical studies, which purportedly found "phonetic effects of stress,"[15,16] failed to deconfound accent from stress. Because stressed syllables can be produced with gradient degrees of accentuation, phonetic measurements of stressed syllables will reflect effects of accent. To avoid the aforementioned confound, a number of studies have purported to compare unstressed syllables to stressed, unaccented syllables.[17,18] Problematically, such approaches must assume a restrictive, categorical phenomenology of accent in which accents are either present or absent in association with stressed syllables, and in which it is possible to determine whether an accent is present or not from more abstract considerations (e.g., from lexical and/or pragmatic information, or from phrasal structure). If we reject these assumptions, it is not possible to control for accent in an investigation of stress. In other words, to measure "effects of stress" independently of accent, is it always necessary to presuppose that one can identify stressed syllables that have no degree of accentuation. This presupposition is unmotivated.

The alternative, simpler view of stress and accent adopted here is that all phonetic correlates of stress are really the correlates of accentuation, where accentuation is understood as a gradient phenomenon with a variety of articulatory and acoustic manifestations. Thus, studies that have claimed to dissociate the effects of stress and accent such as Sluijter and Van Heuven[17] are in fact identifying correlates of accentuation, which can be gradiently present in stressed syllables and which need not be associated with particular semantic/pragmatic information or phrasal structure. Stress in this simpler view is *purely* structural: there are acoustic and articulatory correlates of stress only because stressed syllables may be produced with gradient degrees of accentuation. This view is in line with operationalized determinations of stress such as in Hayes,[1] where all four of the proposed diagnostics of stress are reducible to the potential for a syllable to be produced with an accent.

Another complication in the phenomenology of stress and accent is that there appear to be two types of stress, in the structural sense. In many languages, it is possible to distinguish between primary stress and secondary stress. Syllables with primary stress are typically produced with more extreme accentuation than syllables with secondary stress; accents on syllables with secondary stress have weaker phonetic effects that may not be statistically distinguishable from unstressed syllables.[17] In the example of *Mississippi* in Figure 1, the third syllable has primary stress, and the first syllable has secondary stress. While primary stress intuitions tend to be robust and can be readily verified by tapping experiments,[18] secondary stress intuitions are not always robust across speakers of a language. We can infer in the current paradigm that the nonrobustness of secondary stress intuitions is a consequence of variation in the strength of accents that are associated with secondary stress.

A common representational approach to distinguishing primary and secondary stress involves positing (1) that syllables are grouped into feet, (2) that feet are grouped into a prosodic word, and (3) that one of the feet in the prosodic word is the strongest. As shown in Figure 1, the syllable with primary stress in *Mississippi* is the one that is associated with a strong foot, conceptualized as the "head" of the prosodic word. Applying the same logic within each foot, syllables with stress

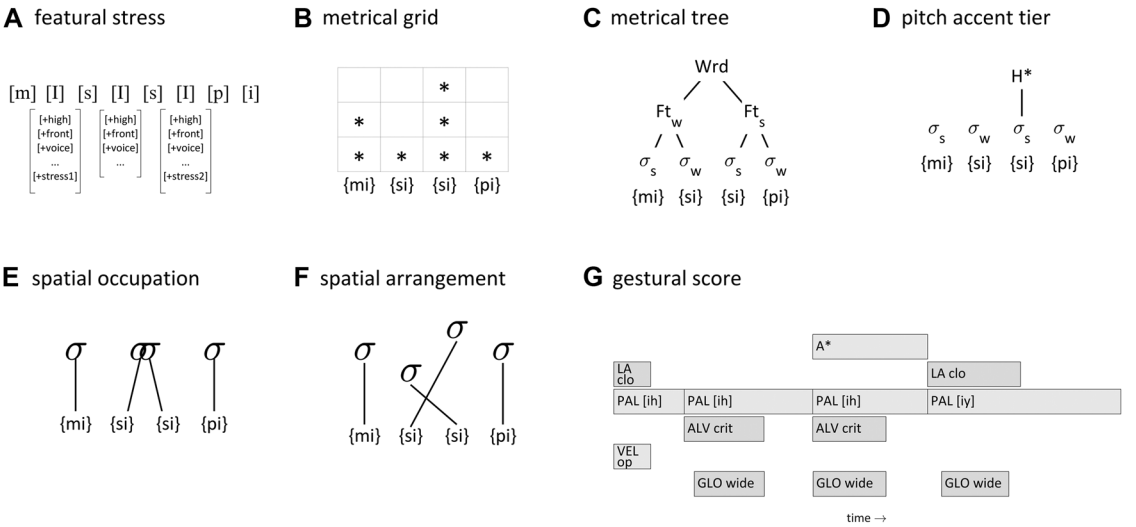

Figure 2. **Representations of stress and accent, and examples of violations of metaphoric inferences. (A) Featural representation in which primary and secondary stress features are distinguished. (B) Metrical grid in which prominence marks are associated with syllables. (C) Metrical tree, where stress is represented via a hierarchical grouping structure. (D) Representation of pitch accent on an accentual tier. (E and F) Examples of violations of spatial occupation and spatial arrangement mappings. (G) Gestural score, where gestures are periods of time during which forces drive changes in the state of the vocal tract.**

(whether primary or secondary) are the heads of feet. However, there are alternative representational approaches that employ different metaphors, which we examine below.

Conceptual metaphors in symbolic and dynamic representations of stress and accent

Formal, symbolic phonological representations of the sort in Figure 2A–D are grounded in the metaphor that linguistic units are physical objects. The metaphor provides a set of mappings from our experiences with the concrete domain of physical objects to the abstract domain of linguistic units.[19,20] These mappings are used to reason analogically about linguistic units, by drawing inferences from our experience with physical objects. For example, there is no *a priori* reason why units in representations might not overlap, as shown in Figure 2E. Indeed, it is well established that the articulatory movements within and between syllables typically *do* overlap. Why have no formal representational models ever been developed in which symbols overlap? Such representations have not even been discussed as a possibility. The reason is that in our typical experiences with physical objects, two distinct objects *do not occupy the same space*. This characteristic of our experience in the physical domain is transferred to how we represent and reason about the abstract domain, i.e., linguistic units.

Symbolic representations also universally employ the metaphor that temporal order is a spatial arrangement. In the conventional application of this metaphor, units are arranged horizontally and events that occur later in time are arranged to the right of events, which occur earlier in time. There is no *a priori* reason why symbolic representations might not be constructed with nonlinear spatial arrangements as in Figure 2F. Indeed, when units of different "types" are considered, nonlinear arrangement is used extensively.[21–23] Nonlinear arrangements of units of the same type are generally avoided because such depictions violate the conventional mapping of temporal order to a linear spatial arrangement. Furthermore, because units are conceptualized as discrete objects, it is natural to infer a discretization of time in such representations, that is, a temporal order.

Unlike symbolic representations, the gestural scores of articulatory phonology (AP),[24,25] which are based on the computational model of task dynamics (TD),[26,27] do not evoke the object metaphor. Instead, gestural scores are schematic representations of dynamics. An example score for

Mississippi is shown in Figure 2G. An articulatory gesture is a period of time in which there are forces acting upon a parameter of the vocal tract. These forces drive the state of the vocal tract toward the new target state. For example, the "LA clo" gesture in Figure 2G specifies a target value of the tract variable *lip aperture*, and activation of the gesture results in a bilabial closure (for the [m] sound in *Mississippi*). Likewise, the "PAL [ih]" and "PAL [i]" gestures specify palatal constriction targets for the vowels in *Mississippi*. The empirical correlates of gestures are trajectories in the state space of the vocal tract, i.e., movements. Early work in the AP framework focused on oral articulatory gestures; more recently gestural models of tones and intonational pitch accents have been developed.[28–31]

Building on these recent extensions of the framework to the control of tone and pitch, we reinterpret accents here as *accentual gestures*. Accentual gestures drive the state of the vocal tract toward pitch, intensity, or phonation quality targets, and thereby function to control the correlates of accent. In general, the entities that are conventionally understood as accents can be viewed as groups of coordinated accentual gestures. Cross-linguistic variation in the acoustic/articulatory correlates of accent is simply a consequence of the fact that in different languages, speakers make use of different groups of accentual gestures. Gradient variation in accent correlates between and within speakers is understood to arise from differences in the targets of accentual gestures, which may be modulated by contextual and paralinguistic factors such as effort.

Unlike pitch, intensity, and phonation quality, which can be readily associated with target states, durational effects of accent must be conceptualized differently in a gestural framework, because durations are not states of the vocal tract. We will discuss later on how durational effects of accent have a natural reinterpretation in the current approach and are mechanistically distinct from target-state parameters. The reader should note that time is conceptualized linearly in the gestural score, but crucially, gestural scores are not conceptualized as objects, so there is no intuition that gestures cannot occupy the same space. In other words, gestures can overlap in time. Furthermore, it is not sensible to refer to gestures as ordered in time, because there is not always a unique order of overlapping events. Thus, no temporal discretization is imposed by the score.

In addition to spatial occupation and linear ordering, some representations such as the metrical tree (Figs. 1 and 2C) employ an object connection schema to evoke grouping and containment relations. By convention, an object that is connected to another object, which is vertically higher in the tree, is *contained by* the higher level object. Containment schemas are implicit in metrical trees but are also depicted explicitly in bracketed grids.[32] These schemas have been used to conceptualize accentual patterns as the product of a "foot construction" algorithm, in which syllables are grouped into feet of some type (i.e., trochaic and iambic), beginning at a word edge. The reader should note that while containment is fundamental to phrase structure models of syntax, it is somewhat more contested in theoretical approaches in phonology: debates have arisen regarding whether an object can be connected (i.e., contained) by two distinct higher level objects (e.g., ambisyllabicity[33]), whether objects on a given level must be necessarily contained by objects on the next highest level[34,35] (exhaustivity), and whether an object can contain an object of the same type (recursivity). The metrical grid (Fig. 2B) and the gestural score (Fig. 2G) are examples of representations that lack containment/grouping relations altogether.

The above analysis of conceptual models of stress and accent reveals that there are two incompatible sets of metaphors. On one hand, the traditional symbolic conception views speech as spatially arranged linguistic objects and provides notions of temporal order (i.e., discretized time), and in many cases imposes grouping/containment of objects. On the other hand, the AP conception views speech as a state space trajectory driven by forces and lacks temporal discretization and grouping. Below we consider how these two different sets of conceptual metaphors fare in regard to the classification of accentual patterns across languages.

Classification of quantity-insensitive accentual patterns

In some languages, accentual patterns are predictable entirely from the position (i.e., temporal order) of syllables relative to the edges (beginning/end) of a word—these are called *quantity insensitive* patterns. In other languages, accentuation is partly predictable from the composition of the syllables in a word (e.g., the presence of a long

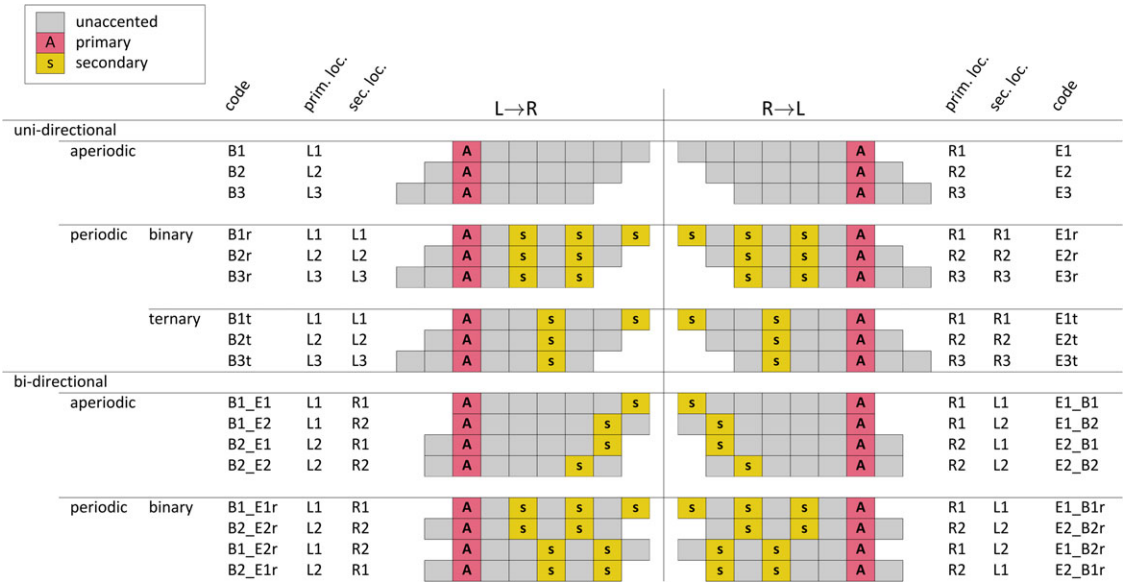

Figure 3. Classification of accentual systems. Directionality is represented by the vertical division of the table. Words are aligned according to the location of primary accent. Rx/Lx indicates syllable positions counting from the right/left edge of the word. Classification codes, where B/E indicates the beginning/end of the word, are included. Note that some logically possible patterns are omitted for brevity.

vowel or coda consonant in a syllable), or is unpredictable and must be determined from long-term (i.e., lexical) memory. We will address quantity sensitive and lexical patterns later; in this section, we review only the quantity insensitive patterns.

A classification scheme for quantity insensitive patterns is shown in Figure 3, derived from typologies in several sources.[1,36,37] Directionality of stress assignment is reflected by the two sides of the vertical division of the table and can be considered as a parameter of the typology. Another parameter is the relative location of the primary accent, which is generally the first, second, or third syllable from the edge associated with the directionality parameter. In unidirectional systems, primary accent and secondary accent (if present) are predictable from the same edge of the word; in bidirectional systems, primary accent and secondary accent are predictable from different edges of the word. Another parameter is whether secondary accent locations are periodic or aperiodic. In periodic patterns, secondary accents occur at regular intervals, either every other syllable (binary) or every third syllable (ternary). In aperiodic patterns, there is no secondary accent (unidirectional systems) or there is a single secondary accent (bidirectional systems). Note that Figure 3 also lists codes for each pattern used in the remainder of this paper where "B" refers to the beginning of the word and "E" to the end of the word.

Symbolic representations, which allow for notions of grouping and discretized time, provide a natural basis for describing accentuation patterns. Accents are uncontroversially associated with syllables, rather than individual segments, and symbolic representations readily allow for the grouping of segments into syllables. In contrast, the dynamic representations of gestural scores cannot achieve the same natural description of accentuation patterns because gestures are not grouped into syllables. The association of accents with syllables might be reinterpreted in a gestural framework as a constraint that accentual gestures can be coupled only to vocalic gestures. However, vocalic gestures cannot be substituted wholesale for syllables because (1) syllables may contain multiple vocalic gestures (as in diphthongs), and (2) in some languages there appear to be syllables that lack vowels.[38,39] Because gestural scores do not represent discretized time or grouping, there is no natural basis for counting syllables from the edge of the score. Symbolic representations thus have a

Table 2. Summary of key terms 1

Key terms	Summaries
Articulatory gestures	Dynamical specifications of a target state for an acoustic or geometric parameter of speech
Accents/accentual gestures	Articulatory gestures, which control fundamental frequency (F0), acoustic intensity, and spectral tilt (related to phonation quality)
Stress	A structural/organizational property of syllables in a word
Primary accent	The most prominent accent in a word
Secondary accents	Accents in a word that are not the primary accent
Accentual pattern	A pattern of association between accentual gestures and syllables in a word
Directionality	The specification of an accentual pattern relative to the left edge/beginning (B) or right edge/end (E) of a word
Periodic accent	A repeating pattern of accents on every other syllable (binary periodic accent) or every third syllable (ternary periodic accent)
Quantity sensitive accentual patterns	Accentual patterns in which the articulatory compositions of syllables influence the association of accentual gestures

considerable advantage over gestural scores when it comes to describing classes of accentual patterns. A summary of relevant terminology is provided in Table 2.

The selection–coordination framework and grouping of gestural selection

The selection–coordination (s/c) framework,[12,40,41] which is an extension of AP[24,25] and TD,[26,42] imposes grouping on gestures of the score and hence allows for a dynamic conception of speech that is more suitable for understanding accentual patterns. The s/c framework accomplishes this by integrating gestural scores with a competitive queuing mechanism.[43–46] The empirical motivation and details of the s/c framework have been discussed extensively in earlier work;[12,40,41] here, a brief overview is provided.

In the s/c model, prior to the production of a word, premotor systems associated with articulatory gestures in the word are organized into competitively selected sets $\{g\}_1 \ldots \{g\}_n$, which conform to a stable pattern of relative activation, as shown in the top panel of Figure 4. When production of the word is initiated, a competition process occurs in which the activations of the sets increase until one of them exceeds a selection threshold. At this point, the gestures in the above-threshold set are executed. Note that the precise timing of the execution of coselected gestures is governed by phasing mechanisms hypothesized in the AP framework.[12,47] During the epoch in which set $\{g\}_1$ is selected, competing sets $\{g\}_2$ and $\{g\}_3$ are *gated*, i.e., their activation is prevented from increasing. Eventually, feedback is received regarding the achievement of targets associated with the gestures in $\{g\}_1$. The feedback induces the suppression of this set and degates the competitors, allowing for the competition process to resume until the next most highly active set, $\{g\}_2$, surpasses the threshold and is selected. This cycle of selection and feedback-induced suppression iterates until all sets have been selected and suppressed.

To conceptualize the stability of activation patterns, the s/c framework employs a quantal potential function,[12,48] in which energy barriers maintain the relative activation pattern prior to production and during steady state epochs of production. As shown in the bottom panels of Figure 4, the competitive queuing dynamics can be described more phenomenologically as a sequence of relatively steady-state epochs (i–v) and intermittent, abrupt reorganizations (i′–iv′). In the initial organization (i), the highest level of the potential (the selection level) is unoccupied. When the response is initiated, a fast-timescale reorganization of the potential occurs in which each set is promoted one level (i′). Subsequently, a new stable pattern emerges (ii) in which the selection level is occupied, inducing execution. Feedback regarding target achievement eventually induces another abrupt reorganization (ii′), in which the selected system is demoted and the competitors are promoted. Alternating steady states and abrupt reorganizations continue until all systems have been demoted to the ground level.

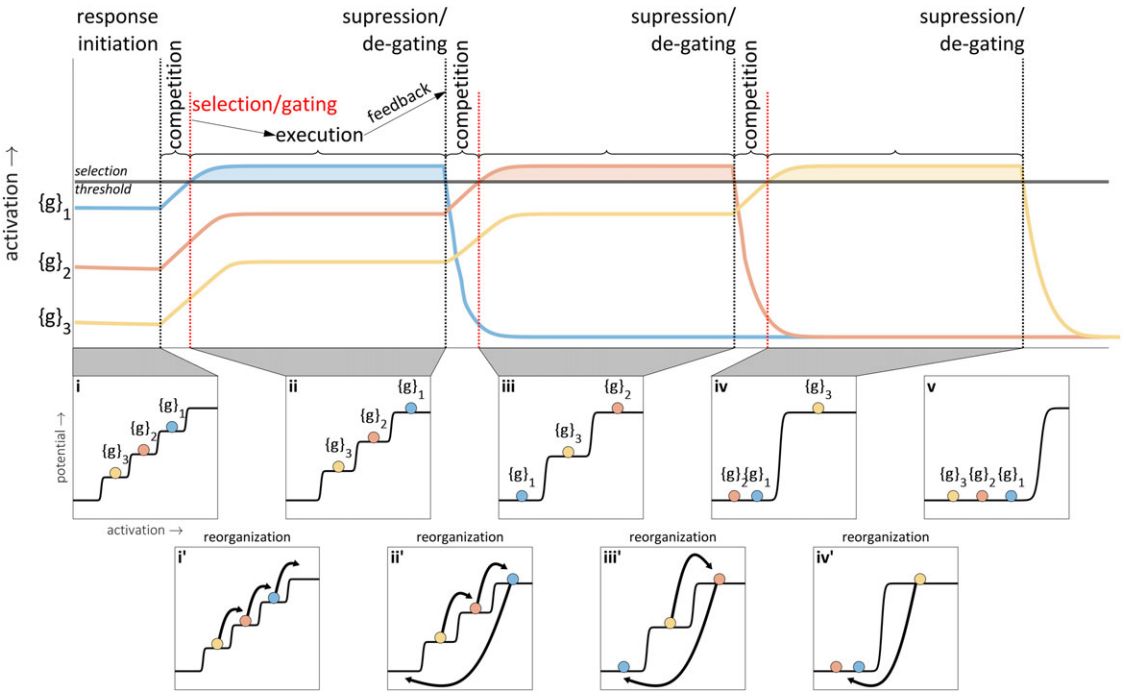

Figure 4. **Competitive queuing model of sequencing and quantal potential functions describing steady state relative activation patterns and reorganizations. Three sets of articulatory gestures, $\{g\}_1$, $\{g\}_2$, and $\{g\}_3$ are initially organized in a stable pattern of relative activation (i). When the sequence is initiated, a rapid competition process occurs, corresponding to an abrupt reorganization of the potential (i′). The gestures in the first set to reach the selection threshold are selected (ii), while the competing sets are suppressed. Subsequently, feedback drives the suppression of the selected set and the competition process resumes (ii′). The selection–feedback–suppression cycle iterates (iii, iii′, …) until all sets have been suppressed.**

In the s/c framework, the association of accentual gestures with syllables is reinterpreted as the coselection of accentual gestures with a set of oral articulatory gestures. This conception is only possible because the framework incorporates a selection mechanism in which gestures are organized into competitively selected sets. Such a mechanism is not available in the standard model of AP. Importantly, the s/c framework does not require that there exists a spatial mapping of systems to their order of selection; the order of selection may be determined solely by an initial relative activation pattern. However, to account for directionality in accentual systems, it will be useful to impose a spatiotemporal correspondence between sets of gestures and their order of selection. Below we extend the s/c model to accomplish this, but first, we consider a unique connectionist approach developed in Goldsmith,[13] which, to a large extent, inspired the current approach.

The Goldsmith model: a dynamical computational theory of accentual systems

The Goldsmith model[13] is a connectionist network in which each node corresponds to a syllable in a word. The nodes are linearly arranged reflecting their order in the word. Figure 5A shows a network for an eight-syllable word form. Each node has a real-valued activation state, and in each time step, the nodes transmit a portion of their activation to their nearest neighbors. The leftward and rightward transmission coefficients are α and β. The initial and final nodes can receive an external source of activation, and a separate external source can be uniformly applied to influence the internal activation of all nodes. All external sources are held constant throughout a simulation.

To understand the temporal evolution of the model, consider the time course of node activation shown in Figure 5C. In the initial

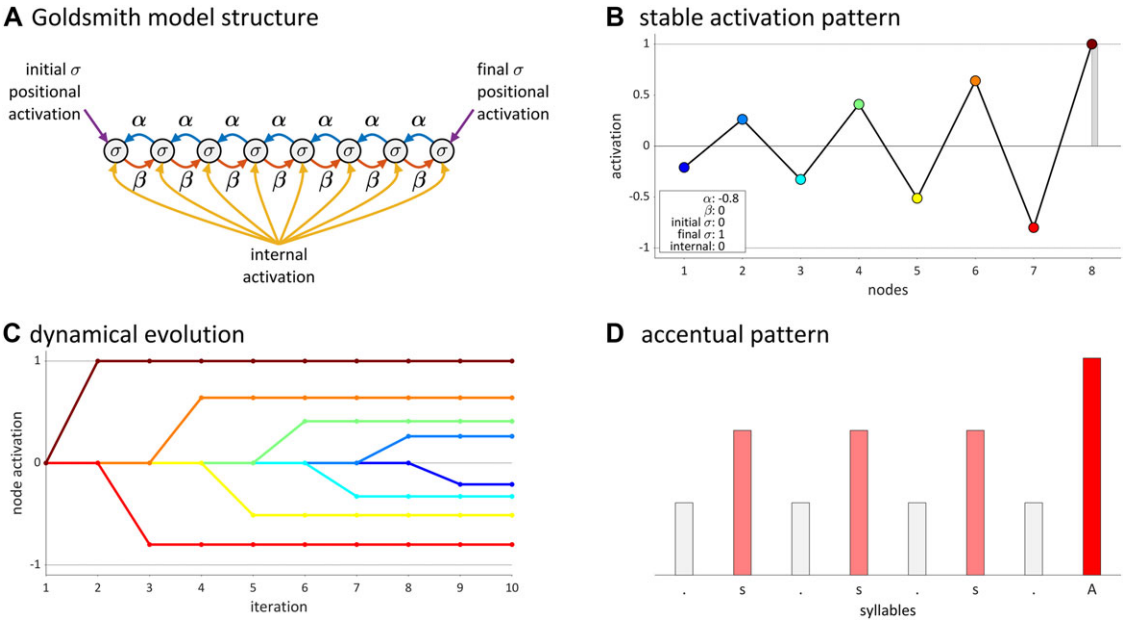

Figure 5. Illustration of the Goldsmith connectionist model, showing a periodic E1r pattern. (A) Model structure for a word form with eight syllables. (B) A stable pattern of activation that emerges after iteration of the model with parameters $\alpha = -0.8$, $\beta = 0$, and final-σ activation of 1. (C) Dynamical evolution of the model. (D) Accentual pattern: "A" primary accent; "s" secondary accent; and "." unaccented.

condition, all nodes have 0 activation. At the first time step, positional activation causes the final syllable node to become activated. In the second step, the final syllable node transmits a portion of its activation to the penultimate syllable—in this case, negative activation (or perhaps, inhibition), according to the leftward transmission coefficient (here $\alpha = -0.8$). At each time step, activation is transmitted further leftward, and the final node continues to receive 1 unit of positional action. After some number of iterations, the activation function over nodes stabilizes, exhibiting a pattern of activation peaks and valleys (Fig. 5B). As shown in Figure 5D, the location of primary accent is the highest peak, and all other peaks are potential locations for secondary accents. In this example, the stable activation pattern corresponds to pattern E1r, i.e., R→L iambs.

The model is able to generate many of the patterns shown in the classification in Figure 3 (e.g., see Fig. S1, online only). When $\alpha > \beta$ and the positional activation is final, the system generates patterns with R→L directionality. Depending on whether the positional activation is positive or negative, an E1r or E2r pattern is generated. Patterns with L→R directionality are generated with initial positional activation and $\alpha < \beta$. When the dominant transmission coefficient is positive rather than negative, the pattern will have only a single peak anchored to a word edge, thereby generating an aperiodic accentuation pattern. Bidirectional patterns can be generated by combining initial and final activation with rightward or leftward dominance, and lexical or quantity sensitive patterns—which we address later on—can be generated by imposing node-specific (nonuniform) internal activation.

An important feature of the Goldsmith model is that, just like object-based symbolic representations, a spatial arrangement of units is imposed. This arrangement provides a basis for the nearest-neighbor constraint on interactions and for differentiating leftward and rightward transmission of activation. If we take the spatial arrangement and object-metaphors somewhat literally, there are a number of problems that arise (see Fig. S2, online only). One problem is rearrangement: for different words, different spatial arrangements of nodes are required. How is this variation in spatial arrangement accomplished? Another problem is multiplicity: without any further constraints, words could

be associated with an arbitrarily large number of nodes and hence an arbitrarily large space. On the other hand, if the space is constrained to be finite (so that there is a maximum node capacity) a *void space* problem arises: for words whose number of syllables is less than the maximum capacity, some of the space is "unused," assuming that unit "size" (i.e., how much space a unit occupies) is constant. Rearrangement, multiplicity, and void space may not seem problematic from a symbolic perspective, but if we are to really embrace the idea that rhythmic patterns emerge from interactions in a *physical space*, such issues must be addressed.

The motor sequencing field and sets of coupled articulatory gestures

Here, we conjecture that there is a physical space, in the brain, in which there is a spatial arrangement of systems that organize articulatory gestures into sets. To ground this conjecture, let us imagine that space contains a large population of interacting microscopic units (e.g., a network of excitatory and inhibitory neurons, or perhaps cortical microcircuits). This population is the *set organization population* (see Fig. S3, online only). We assume, on the basis of empirical and theoretical studies,[49–54] that the microscopic units can enter into a regime of collective oscillation. We then posit that the full population has the ability to self-organize into subpopulations and that these subpopulations are spatially arranged in a manner that corresponds to the *initial* organization of sets of gestures in the s/c potential. Hence, one subpopulation occupies a region of the space that is associated with a set of gestures that will be selected first/earliest in time; a different subpopulation in a neighboring region of the space is associated with gestures that will be selected next, and so on. For the word *Mississippi*, there are four subpopulations of the set organization field, corresponding to the four syllables in the word.

In addition to the spatially arranged population of microscopic units that encodes set organization, we posit a second population that is composed of subpopulations that encode articulatory gestures. The spatial organization of this *gestural population* does not reflect a spatiotemporal mapping; instead, its topology relates to a somatotopic organization associated with the targets of articulatory gestures in relevant sensorimotor coordinates. The microscopic units in the gestural population and the set organization population interact bidirectionally via synaptic projections. Via a positive feedback or resonance mechanism, gestural and set organization subpopulations are able to transiently couple when they are in the collective oscillation regime. This mechanism has the effect of temporarily "binding" gestural subpopulations. In a sense, this picture is a mechanistic, microscale elaboration of more abstract slot-filler models[55] that describe the organization of segments into syllables.

Starting from the microscale picture, we zoom out to a more macroscopic perspective and refer to a collectively oscillating subpopulation of microscopic units as a *system*. Each system has a time-varying activation state, which is derived from a short-time integration of a function of all of the states of the microscopic units in the corresponding subpopulation. Furthermore, we think of the entire population of microscopic units as a field, so that the systems are associated with distinct regions of a *motor sequencing field*.

Next, we envision that the organization of contemporaneously active sets of gestures is accomplished via a *set organization standing wave* in the motor sequencing field, which leads to the picture in Figure 6A. This particular standing wave pattern is generated by imposing zero amplitude variation (i.e., node) boundary conditions on the spatial edges of the field, which receives a periodic external input. The set organization standing wave self-organizes such that there will be one antinode (local maximum in amplitude variation) for each set of coselected gestures (cf. the vertical axis labels in Fig. 6B).

To classify accentual systems, two additional dynamical mechanisms are attributed to the motor sequencing field. One is a *metrical* standing wave that may have symmetric boundary conditions (node—node; antinode—antinode) or asymmetric boundary conditions (node—antinode; antinode—node), and a wave number that corresponds to a half-integer multiple of the number of sets (in the case of symmetric boundary conditions) or a quarter-integer multiple of the number of sets (in the case of asymmetric boundary conditions). We refer to a combination of boundary conditions and wave number as a *mode*; the collection of all possible metrical wave modes for up to five sets is shown in Figure 6D.

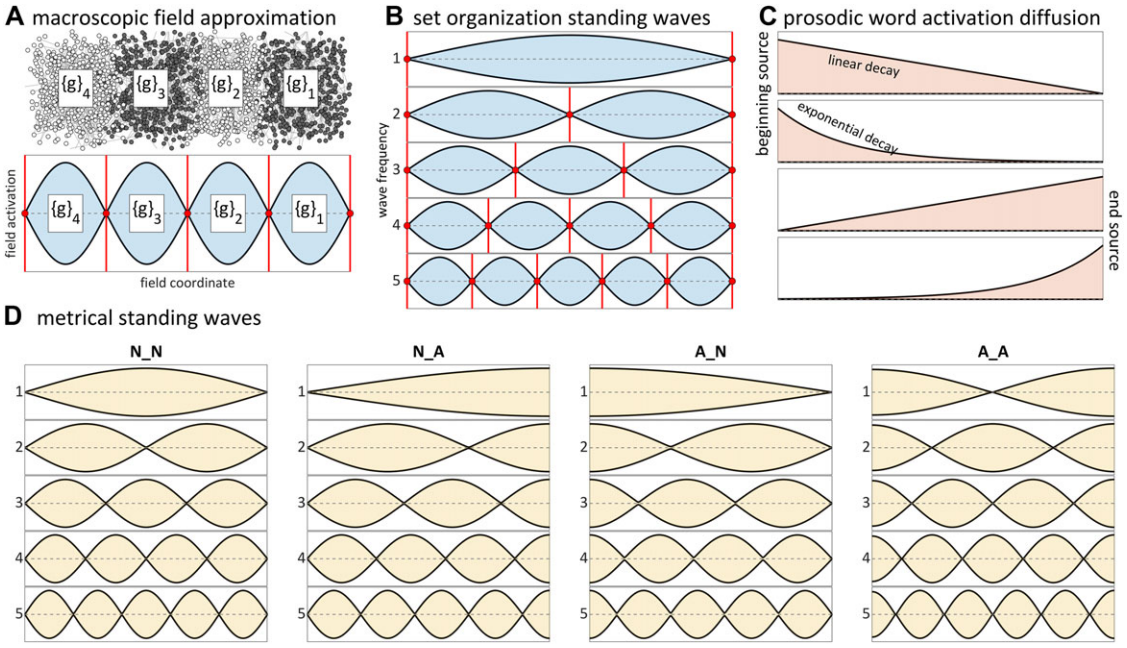

Figure 6. Dynamics in the wave/field model. (A) Macroscopic conceptualization of the set organization population as a field. (B) Set organization standing waves for words of one to five syllables. (C) Prosodic word standing waves. (D) Metrical standing waves: each combination of wave number and boundary condition is a mode of the metrical subfield.

The set organization and metrical standing waves are excited by a periodic source that is located at the beginning or end (i.e., left or right edge) of the field. The frequency of the source varies in order to excite different modes of the field. All simulations of standing waves were conducted numerically using a finite difference method applied to the one-dimensional damped wave equation (see the model description in File S1, online only).

The other dynamical mechanism in the motor sequencing field is *prosodic word activation diffusion*. This is modeled by the one-dimensional diffusion equation (see the model description in File S1, online only), where a source of excitation is implemented as a nonzero activation boundary condition. Depending on the value of the diffusion coefficient in the equation, the prosodic word activation diffusion pattern exhibits either a linear change in activation density over the field or an exponentially decaying density, as contrasted in Figure 6C.

The activation of the motor sequencing field is derived from the interactions of the three subfields described above: (1) the set organization subfield, (2) the metrical subfield, and (3) the prosodic word subfield. There are a number of ways in which these interactions could be modeled; for current purposes, we adopt a relatively simple approach in which motor sequencing field activation is the product of the set organization subfield with a weighted sum of the metrical and prosodic word subfields. By integrating the motor sequencing field activation over the regions of space associated with each partition/set of gestures, activation values are obtained for each set. As in the Goldsmith model, peaks in the activation pattern are associated with accents. In other words, accentual gestures can be coselected with sets in a region of the space where there is a peak in the activation pattern. The strongest accent is assumed to couple with the most highly active set of gestures; hence, primary accent is the highest peak. Mechanistically, these assumptions are sensible if the activation of a set influences its propensity to couple with accentual gestures; on the microscale, this implies that if more neurons are spiking in a subpopulation, its interactions with gestural populations are stronger. A summary of the key terms in the wave/field model is provided in Table 3.

Table 3. Summary of key terms 2

Key terms	Summaries
Selection–coordination (s/c) theory	An extension of the articulatory phonology model in which gestures are organized into sets, which are competitively selected
Set of gestures	A group of articulatory gestures—often corresponding to a syllable—which are contemporaneously selected (i.e., coselected). The influences of coselected gestures on the vocal tract co-occur in time
Competitive selection	Mutually exclusive selection of gestures, which results in their influences on the vocal tract being ordered sequentially in time
Quantal potential	A function that describes forces that stabilize the relative activations of sets of gestures
Reorganization	Abrupt changes in relative activations of sets of gestures, which govern the sequencing of syllables in a word form
Promotion/demotion	Reorganizations that increase/decrease relative activation
Gestural population	Microscale conception of neural populations that correspond to individual gestures
Set organization population	Microscale conception of neural populations that correspond to sets of gestures
Motor sequencing field	Macroscale interpretation of the set organization population that defines an activation function over a one-dimensional space
Set organization subfield	Component of the motor sequencing field activation that is a standing wave with node boundary conditions and a spatial frequency that corresponds to the number of sets of gestures in a word form
Metrical subfield	Component of the motor sequencing field activation that is a standing wave whose mode determines the secondary accent pattern of a word form
Standing wave mode	The combination of boundary conditions and wavenumber of a standing wave
Prosodic word subfield	Component of the motor sequencing field activation that is modeled as a diffusion pattern and that determines the location of primary accent pattern in a word form

Generating quantity-insensitive patterns in the wave/field model

Given the above constructs, all of the periodic quantity insensitive patterns can be generated by choice of metrical field modes, prosodic word diffusion pattern, and excitation source locations (see Table S1, online only). Some examples are shown in Figure 7A–E, in each case for words composed of two to five syllables. Figure 7A shows pattern B1r (L→R trochees) and Figure 7B shows B2r (L→R iambs). The reader should observe that within a given pattern, different metrical modes are used, depending on the number of sets that are organized in a word (or equivalently, the number of field partitions, which often corresponds to the number of syllables). Taken together, we refer to the modes employed for a given pattern as a *progression* of modes because the wave number of the mode increases with the number of organized sets. The reader should also observe that a different progression of metrical modes is used for B2r than for B1r. Ternary periodic patterns as in Figure 7D are similar to binary ones but require a different progression of metrical modes.

Variation in directionality is modeled by varying the location of excitation sources, which can be at the beginning or end of the field. Patterns E1r and E2r, which are the R→L counterparts of B1r and B2r, can be generated with an excitation source at the end of the field; E1r and E2r employ different progressions of metrical modes than B1r and B2r. Because prosodic word activation is strongest at the edge where the source is located, the primary accent (i.e., the highest activation peak) will be associated with the activation peak closest to the source edge.

For the generation of aperiodic patterns, there are two reasonable approaches. One is to impose zero weight on the metrical field, as shown for pattern B1 in Figure 7C. In this case, patterns B2 and E2 require an additional mechanism, a "clamp" that inhibits the edge of the field associated with the prosodic word source. The clamping mechanism may be useful for generating patterns in which edge-units are "extrametrical." An alternative is to posit an accentual gesture competition mechanism—specific to aperiodic systems—which allows only one accentual gesture to be selected with a group of coorganized sets. In that case, B1 can be derived from

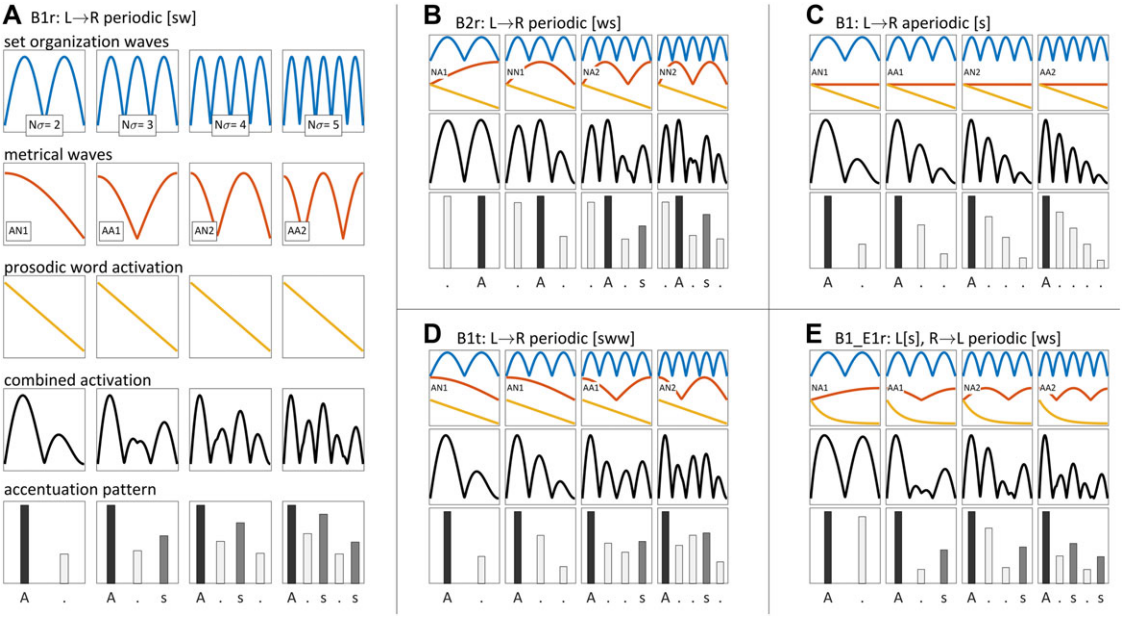

Figure 7. **Examples of quantity insensitive systems generated by the model. (A) B1r, i.e., an L→R periodic binary pattern. (B) B2r is the same as B1r except a different set of metrical modes is selected. (C) An aperiodic pattern generated with the zero-weighted metrical field. (D) Ternary pattern. (E) Bidirectional pattern, where metrical and prosodic word sources are at different edges.**

B1r, B2 from B2r, etc. The accentual gesture competition mechanism has the advantage that fewer model parameters are needed to generate the full range of quantity insensitive systems.

For unidirectional patterns, the excitation source location is the same for the metrical and prosodic word fields. To generate bidirectional patterns, metrical and prosodic word source locations differ. One particular example is shown in Figure 7E, where the prosodic word subfield has a beginning source, while the metrical field has an end source. Explorations of the model indicate that substantial differences in the relative weighting of metrical and prosodic fields are required for generating bidirectional patterns (see File S1. Model details). This may be related to the fact that such patterns are relatively rare, and many of the logically possible bidirectional patterns in Figure 3 are not attested in any languages.

In general, periodic accentual patterns differ with regard to the progression of metrical modes that are employed. It is reasonable to ask if there is any systematicity within or between these progressions. To address this, recall that different modes require different frequencies of source excitation. On the basis of the fact that physical energy of an emitted wave is proportional to the square of the frequency, the metrical modes can be arranged in a hierarchy, according to the square of the source frequency. The hierarchy consists of distinct energy levels, each occupied by two modes. The levels alternate between pairs of asymmetric and symmetric boundary conditions and increase as wave number increases (see Fig. S4, online only). The mode progressions for the four periodic binary patterns (B1r, B2r, E1r, and E2r) are related through a small set of symmetries involving the field boundary conditions. The modes employed for ternary patterns (B1t and E1t) are the same as those of binary pattern modes, except that each asymmetric mode in the progression is used twice. Although it is an open question how speakers of a language learn to employ one progression of modes and not others, the fact that the progressions are systematic makes the learning problem potentially more tractable.

Importantly, the wave/field model incorporates a physical space, which maps *indirectly* to temporal order; sets are mapped to space such that the most highly active set in the initial organization of a word form is located at the beginning of the field, with successively less active sets located farther toward the end of the field. Because the space

is finite, the multiplicity problem is avoided; indeed, the model holds that as the number of contemporaneously organized sets increases, the space devoted to each becomes smaller. With further elaboration of the model, this could be used to predict instability that gives rise to a bound on cardinality (e.g., 7 ± 2 sets[56]). There is also no void space problem: in all circumstances, the entirety of the space in the wave/field model is "used" for the purpose of organizing sets of articulatory and accentual gestures; hence, there is no issue with what happens in "unused" space.

Quantity-sensitive patterns

In quantity-sensitive accentual patterns, the locations of accents are influenced by syllable "weight." Weight is a phenomenon in which syllables can be classified as "heavy" or "light," according to their composition. In some quantity-sensitive languages, only syllables that contain a diphthong or long/tense vowel are heavy; in others, syllables that contain a coda consonant are also heavy.[57,58] Such patterns may also be regular, i.e., fully predictable from syllable composition, or irregular, i.e., derived from lexical long-term memory. English is an example of the latter class. The typology of quantity sensitive accentual patterns is complicated, and it is beyond the scope of the current article to provide a comprehensive analysis. Our focus here is on the conditions that motivate such patterns, which are predicted by applying the wave/field model to hypothesized developmental changes in the organization.

Useful examples for our purposes are provided by geographical names of Native American origin, which are often morphologically opaque to speakers. The majority of such names conform to a periodic quantity-insensitive pattern (E2r), such as *Mississippi*, *Tallahassee*, and *Massachusetts*. However, some words in this class exhibit a quantity-sensitive primary accent on the final syllable, as in *Kalamazoo, Manitowoc*, *Mattamuskeet*, and *Saxapahaw*. In these forms, the final syllable is accented and heavy: it contains a long/tense vowel or a rime with a coda consonant.

The s/c framework provides a new way of reasoning about how quantity-sensitive patterns of this sort emerge. The developmental hypothesis of the s/c theory holds that speakers transition from competitive to coordinative control regimes in early development. Specifically, in early development gestures associated with postvocalic consonants or the second vocalic gesture in diphthongs are organized into separate competitively selected sets, as shown under the *prototypical competitive control* end point of the continuum in Figure 8A. Subsequently, via increasing reliance on internal feedback for degating gestures,[40] children transition to a coordinative regime in which gestures are organized into the same set, labeled as *prototypical coordinative control* in Figure 8A. This distinction makes use of the term "syllable" inappropriate for a general model of articulatory organization: for adults, gestures may typically be selected in syllable-sized sets, but for children in the early word stage (1- to 2.5-year olds), the sets correspond more closely to moras.

The hypothesized developmental transition predicts that in some circumstances, there is an early stage in which gestures are organized in a way that is consistent with the quantity sensitive pattern. Specifically, consider a CV.CVC word form. Figure 8B shows the developmentally earlier, moraic organization where the coda consonant of the final syllable is organized as a distinct set of gestures. In this case, an E2r pattern generates accent on a second-to-last set, which is the final syllable—this is consistent with quantity-sensitive accentuation. The developmentally later, syllabic organization should—according to the E2r pattern—have accent on the initial syllable, but in quantity-sensitive systems, it may exhibit the deviant pattern shown in Figure 8C. This can be attributed to a lexicalization of the selection of the accentual gesture: in the earlier stage, children learn to coselect an accentual gesture with some specific set of gestures in a word form, and this bias on coselection becomes part of the long-term memory of the word form. In the wave/field model, this can be implemented by introducing a set-specific excitation source, which is directly analogous to the nonuniform internal activation employed by Goldsmith[13] to generate quantity-sensitive patterns. The "lexicalization" mechanism employed here is presumably very general and can be applied to generating accentuation patterns in so-called "free stress" languages (like English), where in some word forms, learned patterns of accentual gesture coselection override effects of the metrical field.

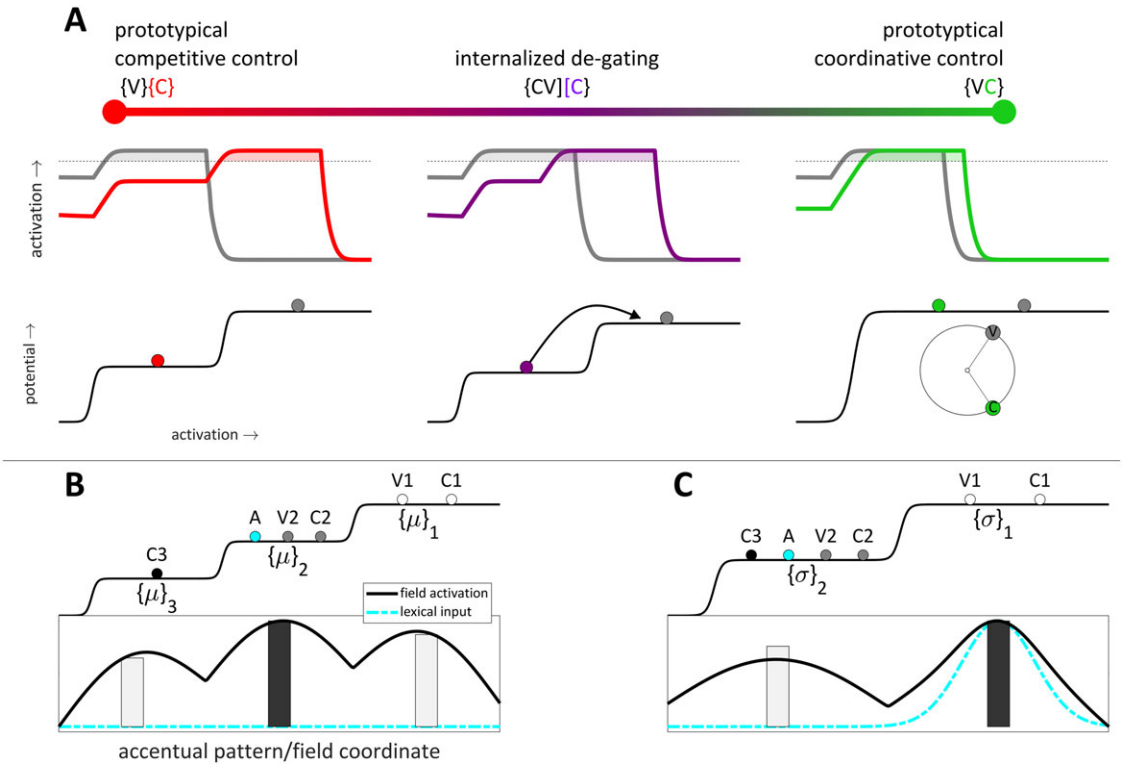

Figure 8. Application of developmental changes in gestural organization to understand quantity sensitivity. (A) Hypothesized developmental trajectory from competitive to coordinative control: children learn to coselect gestures, which were previously organized into separate sets. (B) Left: moraic organization in which postvocalic gesture is a separate partition of the motor sequencing field. Right: syllabic organization with lexicalization of accentual gesture selection.

Additional phenomena: durational lengthening and the rhythm rule

Duration is different from other correlates of accent, such as pitch, intensity, and phonation quality. The latter can be readily understood to involve control of a time-varying state of the vocal tract (e.g., F0, vocal fold tension, etc.), and are well suited to being modeled as gestures in the task dynamic framework. In contrast, accent-related durational variation must be understood differently, because durations are not state variables of the vocal tract (by definition, state variables evolve in time). It is hypothesized here that accentual gestures induce lengthening because they increase attention to external sensory feedback. This increased attention delays the time course of feedback-induced suppression, thereby prolonging the period of time during which selected gestures exert forces on the vocal tract (see Fig. S5A, online only). This concords with the s/c analysis of duration in early development: young children tend to produce words that are longer in duration because they rely to a greater degree than adults on external sensory feedback for the reorganization of gestural activation. An even more general prediction of the above hypothesis is that the degree of regularity in the timing of syllables in adult speech is determined by the regularity in the time course of feedback-induced suppression.

An important constraint on the wave/field model is that metrical/prosodic word influences on accentuation emerge only for systems, which are contemporaneously organized, i.e., organized at the same time in the motor sequencing field. By hypothesis, the scope of this domain tends to correspond to the prosodic word. Crucially, the imposition of the constraint does not entail that there is a one-to-one mapping of utterances to prosodic words. Consider an analysis of the so-called "rhythm−rule" pattern (Fig. S5B, online only): in a noun–noun

compound, the primary accent on the first member of the compound is reduced, as in *Mississippi Michael*. The reduction of the accent on Mississippi is predicted when the sets of gestures in *Mississippi* and *Michael* are contemporaneously organized, i.e., produced as a single prosodic word. As the cardinality of the second member of the compound increases, for example, *Mississippi Mikaela, Mississippi Michaelina*, and *Mississippi Michaelangelo*, the likelihood that the forms will be organized as a single prosodic word decreases. When the forms are produced as a sequence of two prosodic words, each of the two organizations is predicted to have one primary accent, and hence no "deletion" of primary accent is expected. Note that some mechanism is required for resetting the motor sequencing field when a sequence of prosodic words is produced, and this likely involves syntactic−conceptual mechanisms of the sort described in a recently developed model.[48] Rather than viewing the rhythm rule as a consequence of proximity of primary accents, as has been the traditional approach, the phenomenon is thus reinterpreted as a consequence of whether sets of gestures associated with a pair of words are organized at the same time or in a sequence.

Discussion and conclusion

The wave/field model provides a new understanding of how accentual patterns arise, one which is missing from traditional symbolic representations. At issue is the question of how temporal patterns of accentuation in speech are generated, and in particular why there is typological variation in the directionality and periodicity of accentuation. Traditional approaches can describe the variation, but they merely stipulate that the typological variation occurs, failing to provide a mechanism for its emergence.

The proposed model provides this missing piece of the puzzle by connecting a model of articulatory control—i.e., the s/c model—to an understanding of how sets of gestures are organized. The current model, as well as its inspiration, the Goldsmith model,[13] hold that rhythmic patterns arise from a spatial organization. By integrating a spatial model with the s/c framework, regularity in the timing of accentual gestures is derived from the coselection of accentual gestures with articulatory gestures. These coselection patterns are biased by standing wave and diffusion components of activation in a motor sequencing field. Crucially, this field is conceptualized as an approximation of the dynamics of a population of neurons that occupy a physical space in the brain. The directionality of accentual patterns arises naturally from the model via the hypothesis that external activation sources excite one or the other edge of the field.

The wave/field model can in some ways be viewed as a reinterpretation or elaboration of the Goldsmith model, and there are a number of similarities: the positional activation parameters parallel the location of source excitation, a mapping of units to a physical space is used, and in this space there are spatial waves (although the mechanisms that give rise to the waves in the two models differ). Also, the positive/negative sign of positional activation in the Goldsmith model corresponds to the antinode/node boundary conditions in the wave/field model. However, the wave/field model is not simply an alternative vocabulary. By integrating the model into the s/c framework, the model allows for the partitioning of the organizing space to vary in the course of development. This provides a natural basis for understanding quantity sensitive patterns through lexicalization of patterns, which arose in earlier stages of development.

The focus of this paper has been on accentuation in spontaneous conversational speech, but it is evident that periodicity of accentuation, i.e., the rhythmicity of speech, may be substantially enhanced in certain contexts or genres such as poetry, chant, lyrical music, and even prepared speech. A sensible account of periodicity enhancement in such contexts involves the entrainment of selection and suppression events to an external periodic signal (as in lyrical music) or an internally generated signal (as in composition and production of poetry). However, it is also evident that the scope of organization can be adjusted to promote rhythmicity. For example, spontaneous conversational production of the phrase *twinkle twinkle little star* may be organized quite differently from the lyrical production, where each syllable is a separate prosodic word and has a primary accent.

Finally, the directionality parameter of accentual patterns can be fully reinterpreted: it is unnecessary to impose the *temporal order is a spatial arrangement* metaphor on our conceptualization of rhythm,

because we have posited a real space in which the subsystems of a word form (i.e., sets of gestures) are organized. There are no "temporal edges" in this view. Instead, there is a spatially finite field, which is dynamically partitioned. This partitioning corresponds to the count of competitively selected sets in a word form, that is, cardinality, and cardinality determines which metrical standing wave mode is dominant in periodic systems. The claim that stress is "purely structural," thus gains a more detailed meaning: stress is the side effect of diffusive prosodic activation and metrical standing waves that interact to create biases on the selection of accentual gestures. More careful attention to the use of spatiotemporal metaphors in our theories is what makes this new understanding possible.

Acknowledgments

I would like to thank Abby Cohn, John Coleman Kevin Ryan, and Lucian Wong for helpful discussions in the course of developing this manuscript.

Supporting information

Additional supporting information may be found in the online version of this article.

File S1. Model details.

Table S1. Wave/field model parameters for quantity insensitive patterns

Figure S1. Examples of accentual patterns generated by the Goldsmith model and their locations in α–β parameter space. (A, B) Final excitation/inhibition with leftward dominance produces periodic R1/R2 patterns. (A′, B′) Initial excitation/inhibition with rightward dominance produces periodic L1/L2 patterns. (C) Initial excitation with right-dominance $\beta > 0$ produces an aperiodic R1 system. (D) Combining initial and final positional activation generates a bidirectional system. (E) Nonuniform (node-specific) internal activation generates a pattern with primary stress on an arbitrary syllable.

Figure S2. Conceptual problems arising from the spatial occupation of units. (A) The rearrangement problem: how are nodes spatially arranged in a word-specific manner? (B) The multiplicity problem: can an arbitrarily large number of nodes be arranged? (C) The void space problem: what happens in unused space?

Figure S3. The relation between macroscale and microscale conception of set organizing systems and gestural systems. A set organization population differentiates into subpopulations, which encode sets of gestures. These sets arise from transient coupling of gestural populations to the set organization population. Each subpopulation of the set organization population corresponds to a different set of gestures whose initial activation is organized in the sequencing potential.

Figure S4. Energy hierarchy of metrical standing wave modes and progressions of modes used for binary and ternary periodic systems. (A) Energy hierarchy of metrical modes based on the squared frequency of the source excitation required to excite a given mode. (B) Mode progressions for binary periodic patterns. (C) Mode progressions for ternary periodic patterns; loops indicate that a mode is used twice consecutively in the progression.

Figure S5. Hypotheses regarding accentual influences on duration and the rhythm rule. (A) Accentual gestures increase duration by delaying the suppression of selected gestures. (B) The rhythm rule as a consequence of whether there is one contemporaneous organization of gestures or a sequence of two organizations.

Competing interests

The author declares no competing interests.

References

1. Hayes, B. 1995. *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press.
2. Hyman, L.M. 2008. Directional asymmetries in the morphology and phonology of words, with special reference to Bantu. *Linguistics* **46:** 309–350.
3. Beckman, J.N. 2013. *Positional Faithfulness: An Optimality Theoretic Treatment of Phonological Asymmetries*. Routledge.
4. Lavoie, L.M. 2001. *Consonant Strength: Phonological Patterns and Phonetic Manifestations*. Routledge.
5. Barnes, J. 2008. *Strength and Weakness at the Interface: Positional Neutralization in Phonetics and Phonology*. Walter de Gruyter.
6. Keating, P.A. 2006. Phonetic encoding of prosodic structure. In *Speech Production: Models, Phonetic Processes and Techniques*. J. Harrington & M. Tabain, Eds.: 167–186. New York, Hove: Psychology Press.
7. Nooteboom, S.G. 1981. Lexical retrieval from fragments of spoken words: beginnings vs. endings. *J. Phon.* **9:** 407–424.
8. Schwartz, B.L. 2001. *Tip-of-the-Tongue States: Phenomenology, Mechanism, and Lexical Retrieval*. Psychology Press.

9. James, L.E. & D.M. Burke. 2000. Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *J. Exp. Psychol. Learn. Mem. Cogn.* **26:** 1378.
10. Ladd, D.R. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.
11. Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.
12. Tilsen, S. 2018. Three mechanisms for modeling articulation: selection, coordination, and intention. Cornell Working Papers in Phonetics and Phonology. 1–50.
13. Goldsmith, J. 1994. A dynamic computational theory of accent systems. In *Perspectives in Phonology*. J. Cole & C. Kisseberth, Eds.: 1–28. Stanford, CA: CSLI.
14. Cole, J., Y. Mo & M. Hasegawa-Johnson. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Lab. Phonol.* **1:** 425–452.
15. Fry, D.B. 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* **27:** 765.
16. Fant, G., A. Kruckenberg & L. Nord. 1991. Durational correlates of stress in Swedish, French, and English. *J. Phon.* **19:** 351–365.
17. Sluijter, A.M. & V.J. Van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Am.* **100:** 2471–2485.
18. Ortega-Llebaria, M. & P. Prieto. 2011. Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Lang. Speech* **54:** 73–97.
19. Lakoff, G. 2008. *Women, Fire, and Dangerous Things*. University of Chicago Press.
20. Lakoff, G. & M. Johnson. 1980. The metaphorical structure of the human conceptual system. *Cogn. Sci.* **4:** 195–208.
21. Goldsmith, John A. 1976. Autosegmental phonology. London: MIT Press.
22. Goldsmith, J.A. 1990. *Autosegmental and Metrical Phonology*. Basil Blackwell.
23. Coleman, J. & J. Local. 1991. The "no crossing constraint" in autosegmental phonology. *Linguist. Philos.* **14:** 295–338.
24. Browman, C. & L. Goldstein. 1992. Articulatory phonology: an overview. *Phonetica* **49:** 155–180.
25. Goldstein, L. & C.A. Fowler. 2003. Articulatory phonology: A phonology for public language use. In *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. N.O. Schiller & A. Meyer, Eds.: 159–207. Berlin: Mouton de Gruyter.
26. Saltzman, E. & K. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1:** 333–382.
27. Kelso, J.A.S. & E. Saltzman. 1987. Skilled actions: A task dynamic approach. *Psychol. Rev.* **94:** 84–106.
28. Gao, M. 2008. *Tonal alignment in Mandarin Chinese: an articulatory phonology account*. Doctoral dissertation, Yale University, New Haven, CT.
29. Yi, H. 2017. Lexical tone gestures. Ph.D. dissertation, Cornell University.
30. Niemann, H., D. Mücke, H. Nam, *et al*. 2011. Tones as gestures: the case of Italian and German. In *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China.
31. Tilsen, S., D. Burgess & E. Lantz. 2013. Imitation of intonational gestures: a preliminary report. Cornell Working Papers in Phonetics and Phonology 2013, Ithaca, NY.
32. Halle, M. & J.-R. Vergnaud. 1987. *An Essay on Stress*. MIT Press.
33. Jensen, J.T. 2000. Against ambisyllabicity. *Phonology* **17:** 187–235.
34. Peperkamp, S.A. 1997. *Prosodic Words*. The Hague: Holland Academic Graphics.
35. Nespor, M. & I. Vogel. 1986. *Prosodic Phonology*. *Dordrecht: Foris Publications*.
36. Kager, R.W.J. 1995. The metrical theory of word stress. *Blackwell Handb. Linguist.* **1:** 367–402.
37. Gordon, M. 2002. A factorial typology of quantity-insensitive stress. *Nat. Lang. Linguist. Theory* **20:** 491–552.
38. Ridouane, R. 2008. Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology* **25:** 321–359.
39. Dell, F. & M. Elmedlaoui. 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *J. Afr. Lang. Linguist.* **7:** 105–130.
40. Tilsen, S. 2016. Selection and coordination: the articulatory basis for the emergence of phonological structure. *J. Phon.* **55:** 53–77.
41. Tilsen, S. 2013. A dynamical model of hierarchical selection and coordination in speech planning. *PLoS One* **8:** e62800.
42. Saltzman, E., L. Goldstein, C. Browman, *et al*. 1988. Modeling speech production using dynamic gestural structures. *J. Acoust. Soc. Am.* **84:** S146.
43. Grossberg, S. 1978. A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. *Prog. Theor. Biol.* **5:** 233–374.
44. Grossberg, S. 1987. The adaptive self-organization of serial order in behavior: speech, language, and motor control. *Adv. Psychol.* **43:** 313–400.
45. Bullock, D. & B. Rhodes. 2002. Competitive queuing for planning and serial performance. *CASCNS Tech. Rep. Ser.* **3:** 1–9.
46. Bullock, D. 2004. Adaptive neural models of queuing and timing in fluent action. *Trends Cogn. Sci.* **8:** 426–433.
47. Saltzman, E., H. Nam, J. Krivokapic, *et al*. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th International Conference on Speech Prosody*, Brazil.
48. Tilsen, S. 2018. Syntax with oscillators and energy levels. Cornell Working Papers in Phonetics and Phonology 2018, Ithaca, NY.
49. Acebrón, J.A., L.L. Bonilla, C.J.P. Vicente, *et al*. 2005. The Kuramoto model: a simple paradigm for synchronization phenomena. *Rev. Mod. Phys.* **77:** 137–185.
50. Breakspear, M., S. Heitmann & A. Daffertshofer. 2010. Generative models of cortical oscillations: neurobiological implications of the Kuramoto model. *Front. Hum. Neurosci.* **4:** 190.
51. Hong, H. & S.H. Strogatz. 2011. Kuramoto model of coupled oscillators with positive and negative coupling parameters:

an example of conformist and contrarian oscillators. *Phys. Rev. Lett.* **106:** 054102.
52. Kelso, J.A.S. 1997. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press.
53. Strogatz, S.H. 2000. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Phys. Nonlinear Phenom.* **143:** 1–20.
54. Buzsáki, G. & A. Draguhn. 2004. Neuronal oscillations in cortical networks. *Science* **304:** 1926–1929.
55. Shattuck-Hufnagel, S. 1979. Speech errors as evidence for a serial order mechanism in sentence production. In *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. W.E. Cooper & E.C.T. Walker, Eds.: 295–342. L. Erlbaum Associates.
56. Miller, G.A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63:** 81.
57. Zec, D. 2007. The syllable. In *The Cambridge Handbook of Phonology*. P. de Lacy, Ed.: 161–194. Cambridge University Press.
58. Hyman, L.M. 1985. *A Theory of Phonological Weight*. Dordrecht: Foris Publications.