To appear in the *JOURNAL OF PHONETICS* (accepted JUNE 24, 2015).

Unifying Speech and Language in a Developmentally Sensitive Model of Production

Melissa A. Redford

University of Oregon

Running title: Unifying speech and language production

Please address correspondence to: Melissa Redford, Department of Linguistics, University of Oregon, Eugene, OR 97405-1290; tel. 541-346-3818; email. redford@uoregon.edu

ABSTRACT

Speaking is an intentional activity. It is also a complex motor skill; one that exhibits protracted development and the fully automatic character of an overlearned behavior. Together these observations suggest an analogy with skilled behavior in the non-language domain. This analogy is used here to argue for a model of production that is grounded in the activity of speaking and structured during language acquisition. The focus is on the plan that controls the execution of fluent speech; specifically, on the units that are activated during the production of an intonational phrase. These units are schemas: temporally structured sequences of remembered actions and their sensory outcomes. Schemas are activated and inhibited via associated goals, which are linked to specific meanings. Schemas may fuse together over developmental time with repeated use to form larger units, thereby affecting the relative timing of sequential action in participating schemas. In this way, the hierarchical structure of the speech plan and ensuing rhythm patterns of speech are a product of development. Individual schemas may also become differentiated during development, but only if subsequences are associated with meaning. The necessary association of action and meaning gives rise to assumptions about the primacy of certain linguistic forms in the production process. Overall, schema representations connect usage-based theories of language to the action of speaking.

Keywords: speaking, schema theory, speech acquisition, speech planning, relative timing, temporal patterns.

1. Introduction

This special issue on the cognitive nature of speech sound systems provides an opportunity to focus on the indivisibility of cognition and action in language production that is so salient in development. For example, the child's first attempt at referential communication demonstrates a *naming insight* and thus the influence of conceptual development on language acquisition. At the same time, the child's early attempts at words clearly demonstrate the influence of speech motor skills on the linguistic forms produced. Although the conceptual and speech motor domains are often studied separately, here I embrace their indivisibility in speaking to argue for a schema-based theory of language production. The argument takes shape as the outline of a model. The goal is to provide a framework for understanding the details of how we deliver language through speech at any age in a manner that (a) is consistent with usage-based approaches to grammar and language acquisition (e.g., Bybee, 2001; Goldberg, 2006; Tomasello, 2009; Vihman, 2014) and (b) offers continuity in representations across levels of analysis.

The informal model outlined here is motivated by several assumptions: (1) complex behaviors are best understood in terms of their development; (2) fluent speech production requires a plan; and (3) the structure of the plan emerges with production during language acquisition. The last assumption presupposes that production and perception are separate processes. Production is tuned to the individual's anatomy and goals, perception to accommodating variable input in relation to meaning. Phonology emerges at the intersection of production and perception from the abstraction and categorization of sound patterns across less abstract motor and perceptual forms that are themselves connected to quasi-bounded conceptual (semantic, pragmatic) information in the lexicon. While the

phonology likely facilitates both word form acquisition and novel word creation across the lifespan, adults typically speak using words and phrases that we have uttered hundreds if not thousands of times before. This observation suggests that speech production need not rely on our most abstract knowledge of sound systems, but instead can be modeled as the activation of remembered forms. This paper focuses on the nature and organization of these forms. The central thesis is that the plan that guides fluent speech production (i.e., production at the level of the intonational phrase) emerges with language acquisition and with the extensive speech practice that acquisition entails. The units of execution in the plan are temporally structured chunks of remembered action and their sensory outcomes, called *schemas*. These are acquired and deployed in service of communicative goals. Individual schemas may combine with others to form larger units over developmental time, depending both on the frequency with which they have been deployed together and on the emergence of new or extended meanings. Schemas may also be differentiated to create smaller units of execution, provided these are associated with meaning and so can serve some specific communicative goal.

Although our focus is on the structure of the speech plan and how it evolves to reflect the complexity of adult language over developmental time, the point is to provide a framework that unifies higher- and lower-level processes in production. Let me therefore provide a higher- and lower-level context at the outset so that it is clear what types of representations are addressed in this paper.

With regards to the immediately higher-level context, speech planning is viewed as the sequential activation of conceptually-linked schemas within a temporal window that is defined by a domain-general constraint on attention or working memory. The order in

which schemas are activated is dependent on the abstract construction activated as a function of the propositional content the speaker intends to deliver (i.e., on language planning). Constructions are understood not only in the traditional manner as form-meaning pairings (Goldberg, 2006), but also as habitual or routine trajectories through a lexical-semantic space. As for the lower-level context, speech plan execution involves setting one or more control parameters that will affect overall time and force, determining both the global rate at which action will unfold and movement amplitude. The control parameter(s) can be reset with the activation of each schema in a sequence. This allows production to be modulated in response to the communicative context. Once parameters are set, schemas are executed. Coordinated articulatory movement is triggered at the relative timing intervals specified by the schema. Movement accuracy is controlled by efferent copy, generated from the sensory information stored within the schema. Between the activation of a particular construction and the details of motor control are the structured representations that guide speaking, intonational phrase by intonational phrase. These representations are the focus of the current paper.

2. Speaking as the implementation of language

A process-oriented model requires the identification of a starting point. The choice is important because it determines the type of description that will follow. Models of fluent speech production typically start with idealized adult language (e.g., Cooper & Paccia-Cooper, 1980; Garrett, 1980; Dell, 1986; Levelt, 1989; Wheeldon, 2000; Turk & Shattuck-Hufnagel, 2014; Shattuck-Hufnagel, 2015). Because adult language is so complex, theorists who start here have adopted linguistic representations derived by others from a 'pure' linguistic analysis; specifically, from the transcription-based analysis of linguistic structure

isolated from context. Although this type of analysis makes the problem of (adult) language description tractable for the linguist, adopting abstract linguistic representations to model fluent speech production means positing units of action that make no reference to speech. To effect speech, the theorist is left to assume that these abstract representations are translated into ones that the motor system is able to reference. This translation process has come to be known as phonological and phonetic encoding in the psycholinguistics literature, and it can result in incongruent marriages between theories.

Consider, for example, the marriage between Metrical Theory and Articulatory Phonology in Levelt's (1989) influential model, which proceeds in stages from lexical access and syllabification to morphological and metrical spellout. After metrical spellout, resyllabification is required. The resulting syllable structures are used to ensure appropriate serial ordering during segmental spellout. Once phonological plans are specified in this way, they are "*enriched* by prosodic information" (p. 409, emphasis in the original) and then matched, syllable by syllable, to motor programs that take the shape of Articulatory Phonology representations (Levelt cites Browman and Goldstein, 1986). It is the reference to Articulatory Phonology in the context of Levelt's information processing model that is jarring since the aim of that particular research program is to unify speech and language representation:

> Gestures are characterizations of discrete, physically real events that unfold during the speech production process. Articulatory phonology attempts to describe lexical units in terms of these events and their interrelations, which means that gestures are basic units of contrast among lexical items as well as units of articulatory action. From our perspective, phonology is a set of relations among physically real events, a

characterization of the systems and patterns that these events, the gestures, enter into (Browman & Goldstein, 1992:23).

Like Browman and Goldstein (1986; 1992), the model I propose grounds language in speech to avoid the problem of translation that occurs in current, psycholinguistic models of speech-language production. Translation is a problem because it requires postulating additional mechanisms for mapping one type of representation onto another in the process of rendering meaning into action (i.e., phonological and phonetic encoding). This kind of labor is at odds with the sheer speed and automaticity of speaking, where automaticity in behavior is defined as a process that requires no conscious attention or working memory resources dedicated to the selection and manipulation of information[1,2].

The model proposed here may be seen to parallel work in Articulatory Phonology; and I have absorbed many ideas and insights that have been offered by researchers who work within that framework. A difference between what is proposed here and a model cast from within Articulatory Phonology is that I emphasize the importance of development and

---

[1] The definition of automatic versus controlled processes that I adhere to here is taken from Shiffrin and Schneider (1977) who define these processes as follows:

> An automatic process can be defined… as a sequence of nodes that nearly always becomes active in response to a particular input configuration, where the inputs may be externally or internally generated and include the general situational context, and where the sequence is activated without *the necessity* of active control or attention by the subject (155-6; emphasis added).
>
> A controlled process utilizes a temporary sequence of nodes under the control of, and through attention by, the subject; the sequence is temporary in the sense that each activation of the sequence of nodes requires anew the attention of the subject (156).

[2] Gathercole and Baddeley's (1993:Ch. 4) conclusions support the contention that phonological and phonetic encoding are at odds with the automaticity of speech production. In particular, they find that the evidence from working memory studies on speech production fails to support specific predictions that follow from the hypothesis of encoding. The predictions they consider follow from Garrett's (1980) model.

the centrality of meaning for understanding the representations that guide fluent speech. More specifically, I take child language as the starting point for theorizing about production following the assumption that the complexity of adult language is best understood with reference to its acquisition. I hypothesize that the basic units that guide speech action are the remembered action patterns that first served communicative goals in development. These are differentiated over developmental time with use and the acquisition of more abstract meaning to form smaller units that are also linked to specific communicative goals. At the same time, basic level schemas may be combined to create more complex messages. If the same sequence of schemas is practiced repeatedly, they may fuse together to form a larger unit linked to its own specific meaning / communicative goal. The result is a hierarchically structured speech plan built up from functionally-motivated units of speech action linked to meaning via communicative goals.

2.1. Evidence for a speech plan

Fluent speech requires the coordination of multiple articulators through extended time in order to achieve a communicative goal. Speech therefore represents complex sequential action, and, like any such action, it requires an abstract plan to guide fluent output. Evidence for a guiding representation can be seen in speech at the local level in the form of coarticulation, which is especially salient for adjacent items within a syllable. But anticipatory coarticulation is also observed across multiple segments, and across syllable and word boundaries (Öhman, 1966; Daniloff & Moll, 1968; Recasens, 1989; Whalen, 1990; Magen, 1997). For example, Daniloff and Moll (1968) found almost half a century ago that anticipatory lip rounding for a subsequent rounded vowel precedes the target by up to 4 segments, whether or not there is an intervening word boundary. This well-established

finding of long distance coarticulation suggests that speech execution is guided by a "look-ahead" of at least one lexical item.

Anticipatory effects have also been observed at a more global level. For example, a number of studies have demonstrated that speakers accommodate breath control to the length of an utterance (Winkworth, Davis, Ellis, & Adams, 1994; Huber, 2008; Fuchs, Petrone, Krivokapić, & Hoole, 2013). Pitch settings are also adjusted to the length of a phrase; for example, the global declination in F0 across an utterance is at least partly time/utterance length dependent (Prieto, Shih, & Nibert, 1996; see also Ladd, 2008:75-80). Findings like these suggest that speakers have a rough idea of how long they plan to speak. A number of prosody researchers have argued that this reflects planning with reference to prosodic structure (Byrd & Saltzman, 2003; Krivokapić, 2007; Turk & Shattuck-Hufnagel, 2014; Shattuck-Hufnagel, 2015), an argument that is bolstered by the finding that pause duration varies systematically with the length and complexity of the preceding and following intonational phrases (Ferreira, 1993; Krivokapić, 2007), and the finding that intonational phrase boundaries are well marked by the distribution of speech errors across these units even when embedded within pause-delimited utterances (Choe & Redford, 2012).

What it means to plan speech at the level of the intonational phrase is an open question. Shattuck-Hufnagel (2015) proposes a model where incremental phonological and phonetic encoding occurs within the prosodic hierarchy. In particular, phonetic specification of phonological form becomes more and more complete as planning iterates through the hierarchy until—at the level of the metrical foot—the forms are sufficiently well specified to be passed to the motor system for execution. Because phonetic

specification is handled within the prosodic hierarchy, the representations passed to the motor system include prosodic settings that reference global structure. Shattuck-Hufnagel's model thus accommodates the evidence for speech planning across an extended domain in a way that, say, Levelt's (1989) model, cannot. However, like Levelt, Shattuck-Hufnagel assumes translation from the phonological to the phonetic, and these domains are both viewed as distant from the details of speech motor control.

The model I propose avoids translation and seeks instead to establish continuity of representation across levels of analysis. Shattuck-Hufnagel's (2015) prosody-first view of production is embraced, but the schema representations that guide fluent speech production are imagined to be abstracted from remembered action sequences that accomplished specific communicative goals. As in Articulatory Phonology, I assume that relative timing information is fundamental to these representations. Traditional phonological units, defined as discrete and atemporal, are viewed as epiphenomena. Other phonological units emerge from the practice of speaking. Specifically, phonemes and syllables describe patterns of articulatory coordination that are represented within the basic level schema; prosodic words emerge when specific schemas are regularly used together, and so fuse together through practice; and, the intonational phrase emerges from a domain-general temporal constraint on the speech plan that results in a planning window equivalent to the length of a maximal argument structure construction.

2.2. A plan that references goals and remembered action patterns

Speaking is a linguistic activity. More fundamentally, it is an intentional activity and a complex motor skill, one that exhibits protracted development (Smith & Zelaznik, 2004) and the fully automatic character of an overlearned behavior (Shiffrin & Schneider, 1977).

Together these observations suggest an analogy with skilled behavior in the non-language domain. This analogy structures the specifics of the proposed model, which borrows from many sources, including from Norman and Shallice's (1986) theory of hierarchical schema for the control of routine skilled action.

A fundamental assumption of Norman and Shallice's (1986) theory and the theories on which it draws is that skilled action is goal oriented. For example, Cooper and Shallice (2000) implement Norman and Shallice's theory as a computational model of the behavior "prepare instant coffee." The goal of the behavior is simply *to make coffee*. This high level goal is accessible to conscious inspection by the supervisory attentional system[3], which means that the goal is amenable to deliberate activation or inhibition. The goal *to make coffee* has subgoals that can also be inspected consciously, but need not be. In Cooper and Shallice's implementation these are "sugar into coffee," "milk into coffee," and "grinds into coffee." The subgoals in turn have subgoals; for "sugar into coffee," these include "hold," "discard," "open," and "transfer."

Every goal and subgoal is associated with at least one schema. The basic level schemas in the hierarchy are like the motor schemas first defined by Schmidt (1975) for discrete, skilled movements (see also Arbib, 1992; Shea & Wolf, 2005). Higher level schemas in Cooper and Shallice's (2000) model are simply ordered sets of lower level schemas (see also *schema assemblages*; Arbib, 1992). Importantly, schemas are not accessible to conscious inspection. Instead, they are sequenced, activated, and inhibited by the supervisory attentional system via goals. If subsequences within a schema are not

[3] The supervisory attentional system proposed by Norman and Shallice (1986) corresponds to the central executive in Baddley's (1986) theory of working memory (Gathercole & Baddeley, 1993:6).

associated with some goal, they cannot be accessed or altered on-line through controlled processes.

Because speaking is an intentional and routine activity, Norman and Shallice's (1986) theory provides an obvious framework within which to model fluent speech production. The emphasis in the theory on the relationship between goals and schemas suggests a way to bridge "language" and "speech" by defining the conveyance of meaning/information as the goal of language and the plan as an abstract representation of articulatory action linked to this goal. The difficulty, of course, is in identifying specific goals and thus the bounded action sequences (schemas) that constitute units in a speech plan. This difficulty is addressed here from a developmental perspective.

In this paper, language acquisition is viewed as a process of schema formation triggered by the identification and repeated implementation of communicative goals. As children's mean length of utterance grows from one word to two words to several words and so on until adult-like mean length of utterance is achieved, so too does the plan that guides fluent speech action. New schemas are formed with the acquisition of new words and multiword constructions. These schemas may be either basic level schemas that directly represent articulatory action sequences and their sensorimotor consequences or higher order schemas that emerge when basic level schemas are fused through practice. I will assume that basic level schemas are established during the first attempts at a goal, but that their internal structure evolves over developmental time and that the emergence of higher order schemas requires extensive speech practice. These assumptions are consistent with the fast mapping of concept to form that is a feature of word learning (e.g., Trueswell, Medina, Hafri, & Gleitman, 2013) as well as with the protracted development of

speech motor skills (e.g., Smith & Zelaznik, 2004) and the observation from the adult motor learning literature that skill acquisition may continue over hundreds and even thousands of trials (see, e.g., Newell & Rosenbloom, 1981).

Although the model proposed here takes development as the starting point for defining the representations that guide fluent speech production, it is not meant to describe speech-language acquisition in any kind of detail. The model sketched below does not offer a nuanced view of children's abilities at any particular age or stage of acquisition, even though it could be implemented to do so. Take, for example, the proposal that basic level schemas are abstracted from action sequences used to convey specific meanings in early language. These schemas are proposed as initially equivalent to the "word templates" of very early child language (Vihman, 2014:176-77), but are assumed to become elaborated and differentiated as a child's sense of meaning becomes more abstract. The process of elaboration and differentiation is described in detail sufficient for model building in the next section, but not with reference to specific moments in the time course of phonological and conceptual development. This is in keeping with the objective of the current paper, which is to understand the complex system that underlies adult speech-language production in terms of goals and schemas.

As already noted, language acquisition provides a way to understand the structure of the plan that guides fluent speech production at any age. It also suggests the primacy of certain kinds of schemas over others. For example, child language begins with context-limited protowords that "do not represent the activity to which they refer but are part of that activity (Vihman, 2014:161)." Although most of these are later replaced by true words, the protoword stage suggests the primacy of the performative and pragmatic aspects of

speech. If we understand primacy hierarchically, then the developmental trajectory suggests a structure where symbolic acts, instead of being cleaved from performative/pragmatic ones, are embedded within them. This conceptualization is consistent with Searle's Speech Act Theory, which is used below to formalize communicative goals, albeit in the general manner consistent with our focus on fluent speech production. According to the theory, adult language is composed of utterances that are holistic events. They always have illocutionary force (*F*), the performative/pragmatic aspect of an utterance, in addition to referential (*r*) and, very often, propositional content (*p*). Moreover, just as in development, *F* is primary since it is possible even in adult language to have utterances that lack referential or propositional content, but are nonetheless pragmatically appropriate: Yay! Boo! Uh oh! (Searle & Vanderveken, 1985:9).

3. A schema-based model of production

The schema-based model of fluent speech-language production proposed here is conceived of within a connectionist architecture. Again, schemas are temporally structured sequences of remembered action and their sensory outcomes. As in Shallice and Cooper's (2000) model of control over extended skilled action sequences, schemas are linked to goals, and can only be accessed and manipulated by the supervisory attentional system via these goals. Goals are formally defined as speech acts. This definition is compatible with a developmental perspective and our focus on the plan that guides speech action. A complete model would concern itself with the internal conceptual structure inherent to the illocutionary force (*F*) and propositional content (*p*) that comprise a speech act, but an excursion into the construction of meaning is well beyond the scope of the present paper.

The link between a goal and one or more schemas is established upon schema formation, which occurs when some action sequence is remembered following the successful achievement of a communicative goal. Language acquisition structures the plan; for example, children communicate successfully in single word utterances prior to producing multiword utterances. Basic level schemas associated with single word utterances are therefore established first in development. When multiword utterances are produced, the plan can be described as a sequence of basic level schemas, as shown in Figure 1. When multiword utterances include subsequences of words that have been frequently practiced together, then the plan can be described as hierarchically structured in that the subsequences will represent higher order schemas: sets of basic level schemas fused together through use.

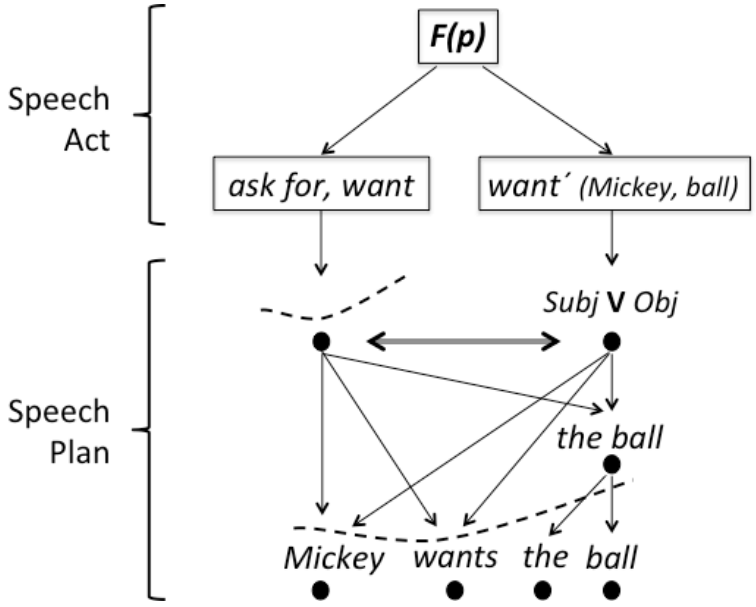

Figure 1. A hierarchical speech plan comprised of basic level and higher-order schemas is activated by the supervisory attentional system via an overall communicative goal (speech act). Note that every schema is associated with some specific, communicatively meaningful goal (not shown) and it is only via these that the supervisory attentional system may truncate or restart speech.

Early word combinations are controlled by the supervisory attentional system via linked goals. More abstract schemas (i.e., constructions) will emerge through time with language acquisition as, say, the class of schemas linked with verbal semantics come to include serial order information through association with those that are linked to verbal arguments. These most abstract schemas may be better imagined as routines through a semantically defined space of goals rather than as ordered sets of basic level schemas. In this way, serial order information at the level of argument structure is more closely tied to meaning than to articulatory action per se, and so is not considered here in any detail. Nonetheless, there is a hypothesized limit on the length of such routines. This limit is attributed to the temporal decay in activation of the overall communicative goal and/or linked schemas (see, e.g., Choe & Redford, 2012). This temporal constraint is presumed to be domain general and due either to the capacity limit of the episodic buffer (Baddeley, 2000) or of attentional focus (Cowan, 2001) or to the window of consciousness that is the psychological present (Grondin, Meilleur-Wells, Lachance, 1999).

In the remainder of this section, I provide a simple but formal treatment of communicative goals in the context of speech-language acquisition, and then focus on schemas that are at the heart of the proposed model in order to justify further and describe more specifically the structure of the speech plan.

3.1. Communicative goals

The typically developing infant acts and reacts in an interpersonal context from the moment of birth on (Trevarthen & Aitken, 2001). Interactions built around, for example, the "I cry, you soothe" interaction are precursors to language, but lack the intentional quality we associate with language insofar as crying is a reflexive behavior. Similarly, very

early vocal-facial imitative behaviors (Meltzoff & Moore, 1977; Kuhl & Meltzoff, 1996) may provide a foundation for speech movement by helping to establish an auditory-motor map, but the intentional quality of such interactions does not go much beyond the "I cry, you soothe" interaction. After all, baby macaques also imitate facial gestures (Gross, 2006). So, although communication is inherent to our social nature, very early communication is not sufficiently intentional to be relevant to the definition of goals that are responsible for schema formation and activation.

In order to define communicative goals and, by extension, the action represented and guided by schemas, we must first identify when the child begins to communicate with intention. I will follow Vihman (2014:32-33) and assume that intentional communication begins just before the onset of first words, between 9 and 12 months of age. At this stage of early vocal development, infants use protowords, deictic and other communicative manual gestures ("showing," "giving"), and may even imitate simple words when primed to do so. Intriguingly, infants' ability to understand an action sequence (e.g., pulling a cloth off a toy) in terms of a final goal (e.g., retrieving a toy) also begins to emerge between 10 and 12 months of age (Sommerville & Woodward, 2005), which is compatible with the analogy adopted here between speech production and skilled non-speech action. Infants who are better able to engage in the action sequence also appear to better understand an actor's intent when engaging in that sequence (*ibid.*), which—assuming the analogy with skilled non-speech action is appropriate—is compatible with the embodied speech-language representations proposed here.

In the 1970s, a number of child language researchers characterized infants' earliest attempts at intentional communication within Speech Act Theory (e.g., Bates, Camaioni, &

Volterra, 1975; Bruner, 1975; Dore, 1975). I will do the same since the theory allows us to formalize communicative goals sufficiently to identify and define associated schemas. Recall that Searle's (1969) formulation of Speech Act Theory divides an utterance into two components. The first is illocutionary force (*F*), which combines the point of the utterance (e.g., question, promise, demand) with pragmatic and performative factors. The other is propositional content (*p*), which is conditioned in part by the illocutionary force of the utterance. Thus, in Searle's terms, speech acts have the structure *F*(*p*). This formulation is convenient for what it suggests about goals and the structure of a plan. In particular, it suggests that *F* and *p* are separable. I hypothesize that this separability implies separate coding in the action system. In the model, *F* is linked to actions at the glottis associated with suprasegmental pitch contours and (initially) with voice quality, and *p* (or whatever content is conditioned by *F*) to action associated with segmental articulation.

F is fundamental to communication. Its primacy is evident in the fact that a speech act can be little more than *F* if the utterance consists of just a single referent, *F(r)*, or may not even require a referent at all, that is, *F( )* (Searle & Vanderveken, 1985:9). From a developmental perspective, which assumes that the fundamentals of human behavior are either innate or early acquired, it is not surprising that *F* is also the primary component of early speech acts; for example, infant's earliest intentional vocalizations have been characterized as *F( )* only (Bates et al., 1975) [4] and early speech as *F(r)* or "holophrastic" (Dore, 1975).

[4] What I describe here as *F( )* only speech acts / goals, Bates et al. (1975) call performatives, which they define as "signals with an illocutionary force" (p. 214).

With regards to *F( )* only speech acts, let us consider Bates and colleagues description of an imperative produced by the 1-year-old child referred to as C:

> C is seated in a corridor in front of the kitchen door. She looks toward her mother and calls with an acute sound "ha." M comes over to her, and C looks toward the kitchen, twisting her body and upper shoulders to do so. M carries her into the kitchen and C points toward the sink. M gives her a glass of water, and C drinks it eagerly (p. 217).

With regard to *F(r)* speech acts or holophrastic speech, let us note that this simply captures the intuition of every parent, caregiver, or older sibling that when a child, let's call him Max, is uttering one of his only words, for example, *ball*, in a pleading tone with rising intonation, then he is requesting the ball. When Max utters *ball* in a harsh tone with falling intonation, he is keeping the ball.

In sum, I use Speech Act Theory to motivate an understanding of overall communicative goals because the theory is consistent with the prosody-first view of speech-language production and the developmental perspective adopted here. In particular, the primacy of *F* in Speech Act Theory and in early intentional communication captures the intuition that the performative and pragmatic aspects of language are fundamental to speaking. Since these aspects are most frequently conveyed by intonation, and since pitch manipulation is perhaps the most practiced aspect of vocalization prior to the onset of first words (see, e.g., Buder, Chorna, Oller, & Robinson, 2008), the primacy of *F* suggests also the primacy of prosody in speech planning.

3.2. Basic level schemas

Basic level schemas are fundamental units of speech production in that they provide all information necessary for speech motor control. They are distributed representations that include (a) information about the relative timing of effector movements; (b) the proprioceptive (kinesthetic) and exteroreceptive (acoustic) consequences of these movements; and (c) a link between this correlational structure and the communicative goal. When goal achievement relies on the use of conventionalized forms (i.e., language), then the schema is linked, via the goal, to a perceptual representation built up from ambient language input. Perceptual representations are critical for the evolution of schema representations through developmental time so that, for example, what is first uttered as [bɑː] for "ball" may eventually become [bɑl]. Perceptual representations are also critical to the initial formation of schemas. Whereas form-meaning correspondences between intonation contours / voice quality and the illocutionary force of an utterance might be discoverable through trial and error, attempts to communicate something about what a word denotes will fail unless what is produced approximates some shared perceptual form of the word.

Figure 2 illustrates the schemas associated with an *F(r)* goal, which is realized as the utterance "Mickey" with rising intonation. Note that sequencing information is intrinsic to the representation, consistent with the generalized motor programs of schema theory (Schmidt, 1975; Arbib, 1992; Shea & Wolf, 2005) and the gestural score representations of Articulatory Phonology (Browman & Goldstein, 1986; 1992). Note also that the figure shows only memory of exteroreceptive information even though memory for proprioceptive information is also assumed. The relative timing information that specifies

articulatory action is read off by the motor system and controlled via efferent copy, which, along with the reafferent input, can be compared during learning by the supervisory attentional system to a separate perceptual representation built up from the input.

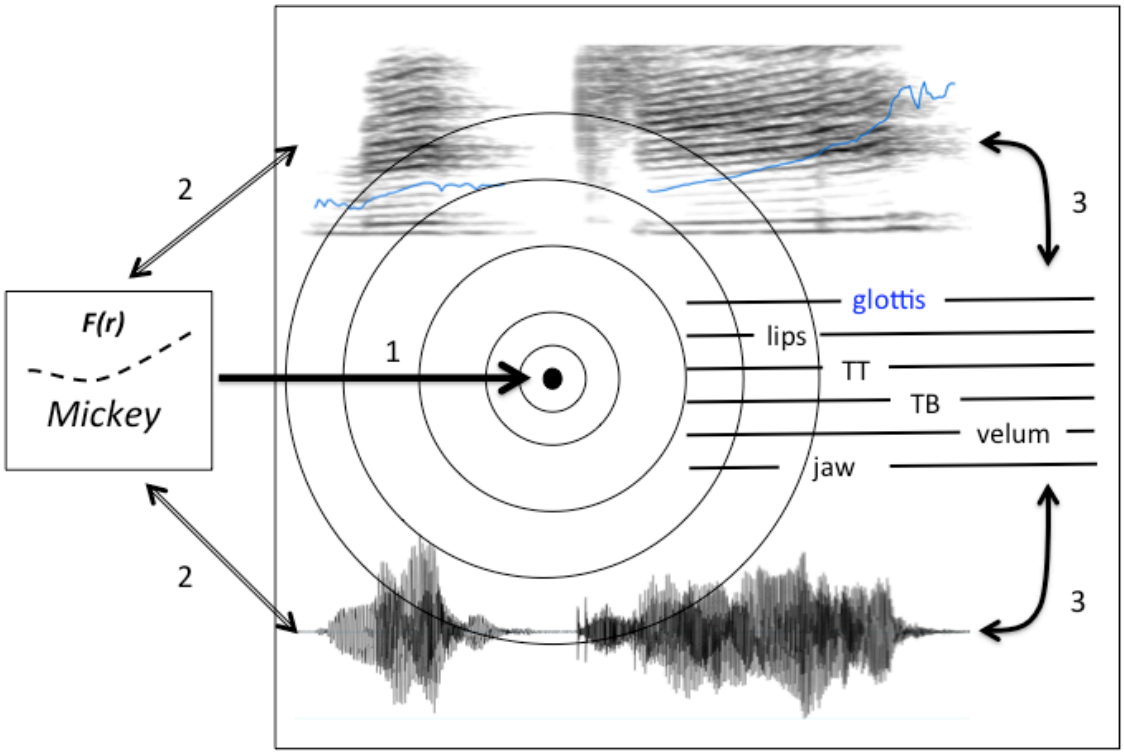

Figure 2. The architecture of a plan is illustrated for the overall goal *F(r)*, which is shown as "Mickey" with rising intonation to code the performative/pragmatic aspect of the goal (*F*). The basic level schemas linked to *F* and *r* represent memory for action that is coded in abstract form. These memories include relative timing information, notated here with reference to vocal tract gestures; memory for action-dependent proprioceptive information (not shown), and for exteroreceptive (acoustic) information (shown). The plan is activated by some goal ("1"). Execution is guided by the information present in the schemas. It is also monitored on-line by the supervisory attentional system ("2") with reference to a targeted perceptual representation (not shown) that is mediated by the goal. If the executed action departs from the intended goal, the memory for action is disrupted so that future attempts at the goal are slightly modified ("3").

I hypothesize that speech-language schemas are formed with every act of intentional communication. Schema formation follows from the implicit remembering of how to achieve a communicative goal. The link between schemas and goals allows for monitoring and an assessment of goal fulfillment ("2" in Figure 2). If the outcome of this assessment is negative, some amount of disruption to the remembered speech actions

follow ("3" in Figure 2). The disruption allows for a different instantiation of the goal on subsequent attempts at communication ("1" in Figure 2).

Different kinds of basic level schemas are formed over developmental time with the different types of speech acts attempted. For example, an *F( )* speech act can be achieved in a variety of ways in adult language, but the way that is most relevant to the formation of basic level schemas is through the manipulation of pitch and voice quality. Consider Bates and colleagues' (1975) report of C's realization of different *F( )* utterances, described earlier with reference to an imperative speech act. Child C, a baby girl, is reported as uttering the same "ha" sound of the imperative sequence in a proto-declarative sequence, which also involved pointing. The difference between the imperative and proto-declarative sequence is in how "ha" was uttered. In the imperative, C produced this protoword "as an acute sound." In the proto-declarative, she produced it as "a breathy sound."

For infants like C different *F( )* goals are achieved if they get what they want as a consequence of their behavior. Schema formation depends on segmenting the action sequence that precipitates goal fulfillment, and then attaching the memory of the sequence to a memory of the goal. In the scenarios that Bates and colleagues describe, the candidate action sequences include glottal changes, supraglottal movement patterns, and whole body as well as arm movements. All of these patterns could be initially represented as the means to an end, provided an equal weighting of glottal, vocal tract, and other body effectors. But because the child will eventually demonstrate an understanding that a declarative or imperative intent is effectively communicated by intonation alone, I assume that this is discovered through repeated trials where pitch changes are exercised independently of supraglottal articulatory and other body movements. The overarching point, though, is that

goal fulfillment (response outcome) drives action segmentation and fixes in memory some portion of the immediately preceding action pattern, including the sensory consequences of that pattern. This is the moment of initial schema formation and it is linked to some specific goal.

If initial representations are memories of a preceding action sequence, then schemas may initially code fairly complex action patterns. These patterns may be distributed across all the articulators involved in conveying meaning, including the non-speech articulators such as the head and hands. It may be only in the random perturbation of a remembered action sequence that the child discovers an optimal pattern for achieving said goal, thereby, separating speech from non-speech representations, and suprasegmental representations from segmental ones. For example, a pointing gesture may cease to accompany a declarative sequence if such a sequence is initiated when the child's arms are otherwise occupied (e.g., they are wrapped around a big object that the child wants to hold on to even while she seeks to direct the interlocutors attention to another object). The converse can also be imagined: the child may find that pointing alone is entirely adequate to achieve specific *F( )* goal when the voice is otherwise occupied (e.g., if the child is drinking out of a sippy-cup, but wants to direct the interlocutors attention elsewhere). If vocal cues, "ha," or pointing alone is found to be sufficient to achieve the goal, then these can each be represented separately. These schemas, attached as they are to the same goal, can then be implemented independently to achieve this goal, combined with each other for emphasis, or combined with other schemas to achieve some other goal.

3.3. Fundamental units of production

Basic level schemas provide the action specifications, relative timing and sensorimotor information necessary for speech motor control. As such they are the fundamental units of production. Since the discovery procedure outlined above will lead to differentiated schemas that can be recombined to achieve new goals, a question arises as to the nature and temporal extent of these fundamental units in an utterance. As units of recombination, the answer must be that basic level schemas represent all and only the unique action sequences associated with the successful achievement of some specific goal, their boundedness varying according to whether the sequences are linked to the illocutionary force or content aspects of an overall goal. Given these two aspects, two types of basic level schemas are proposed: those that are abstract representations of particular intonation contours, and those that are abstract representations of words. Voice quality, while important to early representations formed with the achievement of *F( )* goals, is assumed to be differentiated from intonation during development and, lacking temporal structure at the level of an utterance, is abstracted as a global parameter setting rather than as a schema.

Basic level schemas that are abstract representations of particular intonation contours include relative timing information needed for the appropriate sequencing and spacing of low and high tones across an utterance. The temporal extent of these schemas is always the same as the temporal extent of the plan itself, which is determined by the content of the speech act (see Figure 1). In adult speech, this content will frequently be a proposition (*p*). I assume that *p* activates an associated construction, which has some maximal length due to a basic temporal constraint on planning speech. As suggested

earlier, this domain general constraint reflects the decay in activation of either the overall communicative goal during the window of awareness associated with the psychological present (Grondin, et al., 1999) or decay in activation of the units that have been readied for execution in the episodic buffer (Baddeley, 2000) or within the focus of attention (Cowan, 2001). Both the psychological present and the capacity limits of working memory are approximately 2 seconds in duration.

The hypothesis of a temporal constraint on the size of a plan is consistent with Chafe's (1987:22) idea that a speaker "verbalizes one piece of temporarily active information after another…. (which) is expressed (as)… an 'intonation unit'... new intonation units typically begin about two seconds apart." But the hypothesis is actually motivated by a surprising recent finding from my laboratory based on a longitudinal study on school-aged children's speech. We found that regardless of the child's age at the start of the study, their average intonational phrase duration remained a constant 1 second across a 3 year period (Redford, Foroughifar, & Dilley, 2014). Ongoing work with adults of different ages suggests that the average duration of an intonational phrase may be longer under some task conditions and shorter under others, but the prediction is that these should not exceed 2 seconds.

Just like basic level schemas associated with specific intonation contours, basic level schemas associated with word forms are coextensive with the plan, but only in early development where utterances are well described as *F(r)* speech acts. Thus, the hypothesis is that the basic level schemas that guide segmental articulation are the size of a word. This hypothesis assumes that two word utterances in child language are either always combinations of words that have been previously uttered in a single word context or that at

least one of the words in the utterance was produced previously, such that only the novel action sequence will trigger the formation of a new basic level schema should the communicative goal be achieved.

The hypothesis that word-sized units are fundamental units of production is not new. It is also consistent with work that has investigated speech production from a motor learning perspective (Sternberg, Monsell, Knoll, & Wright, 1978; Sternberg, Wright, Knoll, & Monsell, 1980; Klapp, 2003). For example, in simple reaction time experiments, Sternberg et al. (1978) found no latency effects due to word length in number of syllables, but expected increases with list length. They interpreted these findings to suggest the "stress group" as the relevant response unit (i.e., basic level schema), which in their study was equivalent to a word of any length with primary stress on one of the syllables.

The hypothesis that intonational phrases are also fundamental units of production is more novel, but consistent with the prosody-first view of speech production embraced by a number of researchers (Byrd & Saltzman, 2003; Krivokapić, 2007; Turk & Shattuck-Hufnagel, 2014; Shattuck-Hufnagel, 2015). The hypothesis is also compatible with the widely held view that the speech plan is hierarchically structured, as first suggested by Lashley (1951).

A hierarchically structured speech plan is consistent not only with the temporal patterns inherent in speech and associated with language rhythm, but also with the motor learning literature. For example, Sternberg and colleagues (1988) argued that the response units (words) in their experiment were embedded in a larger motor program, equivalent to the present conception of a speech plan. Consistent with the hypothesis of a temporal constraint on speech plans, they found that these larger programs had a maximum length

(see, e.g., their Figure 8). Sternberg et al. were not certain what constrained the length of their larger motor programs, but they argued that it was *not* working memory span. Their argument was based both on the number of items produced and on utterance duration:

> The breakpoint for monosyllables substantially exceeded the same subjects' memory span (7.4±0.6 words) for the same material, further indicating that the capacity under study is distinct from short-term memory. Moreover, the difference between capacity estimates in terms of number of words does not translate into equal measures of capacity in terms of spoken duration, as might be expected if programs and short-term memory capacities were the same… Utterance duration at their respective breakpoints is substantially and significantly shorter for monosyllable than for trisyllable utterances (p. 191).

Sternberg et al.'s observation that the length of an utterance and working memory span are uncorrelated supports the present argument against an active process of phonological / phonetic encoding. The structure that Sternberg et al. describe is also consistent with the two types of basic level schema defined here, assuming that the utterances in the Sternberg et al. study were defined by a coherent intonation contour. Although they did not report average durations for the utterances in question, the present hypothesis of a domain general temporal constraint on planning predicts that these were not more than 2 seconds in duration.

3.4. Higher-order schemas

Basic level schemas can be combined to produce wholly novel sequences that achieve new communicative goals. When the goal has the structure *F(p)*, schemas are serially ordered by whatever construction has been activated. Let us call the formulation of

communicative goals language planning and the order defined by constructions the grammar. Further discussion of either are beyond the scope of the current paper.

Although abstract constructions represent well worn routes through a semantic space, the execution of linked basic level schemas represents a new action pattern if the specific content of the construction is new. That is, new arrangements of basic level schemas do not conform to the present definition of schemas as temporally structured sequences of remembered action and their sensory outcomes. However, a sequence of basic level schemas can become a schema in its own right with sufficient practice. This is because extensive practice has consequences for meaning, here associated with goals, as well as for the temporal patterning of an action sequence, here associated with schemas.

Specifically, practice with a sequence leads to meaning that can be considered more than the sum of its parts (Goldberg, 2006). Consider, for example, the collocation *pay attention*, which is a common way of saying "be attentive to." If the speaker wants to convey this meaning they might activate the goal associated with the collocation rather than the distinct goals associated with the concepts *pay* and *attention*. Note, though, that even if such a higher-order goal is established, the basic meanings of the individual words is still accessible. Goals associated with the basic meanings allow the supervisory attentional system to segment *pay* from *attention* during speaking.

Extensive practice with a sequence also has consequences for the temporal patterning of action (Bybee, 2001). This observation is well established for lexical items where, for example, rate normalized word duration is known to vary in homophones as a function of word frequency (e.g., *thyme* versus *time*, Gahl, 2008). The effect of frequency can even lead to substantial differences in the relative timing patterns of near minimal

pairs (e.g., *memory* versus *mammary*, Bybee, 2001). In the current model, frequency effects on timing at the word level would be represented in the basic level schemas. Higher-order schemas are the representations that capture timing patterns that emerge from repeated practice of a sequence of basic level schemas.

The effect of practice on the temporal patterning of word sequences depends on the semantics of the associated goals. For collocations like *pay attention*, where both words have equal semantic weight, the representation of relative timing in the higher-order schema may be identical to that which is described by the sequence of basic level schemas. However, when the co-occurring words differ in semantic weight, higher-order schemas may come to represent the temporal patterning that results from weight-dependent differences in parameter setting during fluent speech production. I am thinking in particular of the reduction of function words relative to adjacent content words (e.g., *the ball*), which results in the percept of prosodic words. The specific hypothesis is that speakers engage in meaning-based modulation of speech, which is expressed in the parameter settings used when a schema is activated. For highly practiced sequences of basic level schemas, on-line modulation gives way over time to the fixed temporal patterns we associate with prosodic words. These temporal patterns will be represented in the relative timing structure of a higher-order schema so long as this higher-order unit is itself associated with a specific new or extended meaning. Otherwise, it is only the structure of the basic level schema itself that is changed with the repeated selection of the same parameter settings during production.

The evidence from child language suggests that higher-order schemas may only emerge with extensive speech practice, and so that the speech plan may initially consist

only of basic level schemas. To wit, children produce the vowel duration patterns associated with strong-weak and weak-strong lexical stress by age 2 years (Pollock, Brammer, & Hageman, 1993; Kehoe, Stoel-Gammon, & Buder, 1995; Schwartz, Petinou, Goffman, Lazowski, & Cartusciello, 1996), but stress-timing at the phrase level does not emerge in English speaking children's speech until sometime between age 5 and 8 years of age (Grabe, Post, & Watson, 1999; Bunta & Ingram, 2007; Payne, Post, Astruc, Prieto, & Vanrell, 2012). Work in my laboratory suggests that the slow acquisition of phrase-level rhythm in English can be attributed to children's production of function words (Sirsa & Redford, 2011), consistent with an observation first made by Allen and Hawkins (1978). Compared to adults, 5-year-old children produce function words with relatively longer vowels, and the vowels produced are less influenced by the word-initial consonant of a subsequent content word (Redford, 2014).

3.5. Schema-internal patterns and the emergence of sub-lexical goals

Basic level schemas are under closed loop control, which simply means that their execution is monitored and can be interrupted. The fact that disfluencies will often take the form of word truncation is consistent with closed loop control. It also suggests attention to schema-internal patterns. This is particularly true when truncation is triggered by a speech error, that is, by an error most often described as the transposition or substitution of segment-sized units in production[5]. But if we assume that monitoring represents a

[5] This observation begs the question of how speech errors might emerge in the present model, but work in phonetics also calls into question the notion that speech errors are due to the displacement of segments (e.g., Frisch & Wright, 2002; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007). And Dell and colleagues (Dell, Juliano, & Govindjee, 1993) have found that they can replicate many speech error patterns in a connectionist model without directly encoding segmental structure. Thus, one way to approach speech errors in the present model might be to explore how these could emerge from competing activation of

comparison of efferent copy against a stored perceptual form that is linked to the intended goal (stage 2 in Figure 1), then segment-sized units are not necessary. This point is briefly elaborated below.

Auditory feedback perturbation experiments have shown that participants will adjust their production of individual sounds, for example, a vowel, in response to altered exteroreceptive feedback, for example, a change in formant frequency (see Houde, 2015, for a review). The change in production is attributed to a motor control strategy that takes an acoustic representation of the vowel as its goal. Representations such as these might suggest that something like an acoustic segment or coordinative structure is the fundamental unit of production, but the discrepancy between this idea and the hypothesis put forward here may be mainly due to a difference in focus. Whereas studies in speech motor control tend to focus on segments, our concern here is with speaking. Even so, the claim in the speech motor control literature is not usually that goals are segment-sized (though, see, Löfqvist 1990), but rather that articulation references an internal representation of the proprioceptive and exteroreceptive spaces established during development (e.g., Guenther, 1995; Schwartz, Basirat, Ménard, & Sato, 2012). In the model imagined here, the mapping between movement and its sensory representation is schema-internal. Goals are only communicative. And, since it is possible to imagine monitoring based on matching the efferent copy to a distinct perceptual form without segmentation, it is not clear to me that individuated coordinative structures or sensory goals would in fact be necessary for speaking under normal circumstances.

---

networked representations due to the many-to-one association between schemas and goals.

Still, many short acoustic intervals carry a functional load in language. For example, whole words can often be segmented into multiple units, each with its own distinct meaning. Although children first acquire morphologically complex words as whole entities (MacWhinney, 1985), they quickly learn the meaning of bound morphemes that are highly productive, like the *–er* nominalizer in English. Once identified, children will use these morphemes appropriately in their own language, albeit inconsistently at first (e.g., Clark & Hetch, 1982). In the present framework, the productive use of bound morphemes suggests that their forms are represented as schemas. When children learn that words may be comprised of meaningful bits, the relevant perceptuo-motor intervals are linked to individuated meanings, and then activated via these meanings during the production of novel forms.

4. Conclusion: implications and open questions

A schema-based model of speech-language production has been offered as a way to unify "speech" and "language" representations so that the automaticity of speaking is accommodated, but its intentional and controlled aspects are also acknowledged. The speaker is argued to control speech action indirectly via goals that are meaning-based. The action itself unfolds with reference to a hierarchically structured plan comprised of basic level and higher-order schemas, which are temporally structured memories of articulatory action and their proprioceptive (somatosensory) and exteroreceptive (acoustic) consequences. The scope of the plan is determined by an overall communicative goal, the speech act, which is comprised of two aspects: illocutionary force and propositional content. Illocutionary force is primarily conveyed in spoken language through intonation. Since illocutionary force is more fundamental to the speech act than propositional content,

as evidenced by *F( )* and *F(r)* productions in adult speech and by the developmental trajectory from protowords to holophrastic communication to word combinations, schemas that code for intonation contours are assumed to be basic and coextensive with the plan itself. In this way, the current model shares an assumption with other prosody-first models of speech-language production (Byrd & Saltzman, 2003; Krivokapić, 2007; Turk & Shattuck-Hufnagel, 2014; Shattuck-Hufnagel, 2015); namely, that the speech plan references a domain that is at least the size of an intonational phrase[6].

The propositional content of a speech act is conveyed through the serial ordering of words. Word order depends on the grammar, more specifically, on the construction that is most highly activated by the propositional content. Together, content and construction spread activation to linked schemas. When the specific words activated by the intended propositional content have not previously co-occurred in the order required by the construction, segmental articulation is guided by word-sized schemas. In this way, word-sized units are seen as fundamental to production, a view motivated by the developmental perspective adopted here, but one that is also consistent with the adult literature as well. When any sub-sequence of words within the construction is highly practiced, higher-order schemas are activated and segmental articulation follows from the specific relative timing patterns encoded therein. Higher-order schemas thus add an additional layer of hierarchical structure to the plan beyond the intonational phrase and word levels.

Model building is an exercise in hypothesis generation within a particular framework and from a particular perspective. The framework adopted here is usage-based

[6] Byrd & Saltzman (2003) do not talk about a plan per se, but instead of a pi-gesture that has consequences for timing at the intonational phrase-level and so can be imagined as co-extensive with some plan.

and the perspective is developmental; the assumptions are that language form follows from communicative function, and complex structure emerges over time. Many hypotheses were advanced. Chief among these is the hypothesis that speaking is guided by schema representations. This hypothesis is similar to the hypothesis from Articulatory Phonology that speaking is guided by a gestural score. But the schemas proposed here are different from gestural scores in several ways: they incorporate sensory information, are linked via goals to separate perceptually-based representations, and are an explicit product of language acquisition.

The emphasis on development in the present model leads to novel hypotheses and testable predictions. For example, the assumption that schema formation begins with intentional communication led to the hypothesis that schemas initially represent whole body and manual gestures with speech information. This hypothesis is relevant to work on co-speech gesture. The informal model sketched here thus could be elaborated to provide a specific and reasonably low-level account of the link between gesture and language in speaking (McNeill, 2008), and especially an account of why non-speech movement facilitates lexical access (e.g., Rauscher, Krauss, & Chen, 1996; Ravizza, 2003).

With regards to testable predictions, the hypothesis that articulatory action is governed by schemas predicts, for example, that relative timing patterns associated with, say, word production are represented separately from the details of articulatory coordination necessary to achieve the sequential sensory targets that define word forms. Put another way, the prediction is that relative and absolute timing are dissociable. Some preliminary evidence in support of this prediction comes from a study we recently conducted comparing child and adult speakers' ability to produce context-specific vowel

duration differences (i.e., relative timing) to their ability to repeatedly produce the same vowel with the same duration in the same context (i.e., absolute timing; Redford & Oh, 2015). The groups differed only in measures of absolute timing, consistent with the predicted dissociation between relative and absolute timing. The direction of the difference suggests that it is on-line motor control that is immature in the school-aged children we study, and not their lexical representations that provide the plan for sequential action.

The model proposed here also raises many questions for further consideration and exploration. These include, but are not limited to, questions concerning the nature of the patterns remembered, in particular, how holistic they are initially, and what is the evidence for their differentiation over developmental time; questions about the recombination of schemas, and how free or constrained the process might be; questions about how the intonational and segmental aspects of speech are aligned in the plan; questions about the nature of a temporal constraint on speech planning, and whether there is in fact one at all. These open questions represent challenges for the present model, but also a kind of success. Model building allows us to search the same landscape again, but from a slightly different perspective. This provides a new understanding of well-known patterns. It also leads us to ask new questions, which can result in the identification of important new patterns. Finally, model building encourages us to critically evaluate competing hypotheses within the context of a research program that has as its goal to reject some ideas and preserve others as our collective understanding of the phenomenon in question increases.

Acknowledgments

This work was supported by Award Number R01HD061458 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD). The content is solely my responsibility and does not necessarily reflect the views of NICHD. I am grateful to my colleagues and to the graduate students in the Linguistics Department at the University of Oregon for providing an intellectual environment that encourages me to think more explicitly about the relationship between language and speech. I am also grateful to Sergei Bogdanov for discussions of mechanism that helped me to better formulate the proposed model, and to Marilyn Vihman and an anonymous reviewer for their trenchant comments and criticisms, which were critical to improving the presentation of ideas.

References

Allen, George D. & Sarah Hawkins. 1978. The development of phonological rhythm. In Alan Bell & Joan Bybee Hooper (eds.), *Syllables and Segments*, 173–185. New York: North-Holland Publishing.

Arbib, M. A. (1992). Schema theory. *The Encyclopedia of Artificial Intelligence*, 2, 1427-1443.

Baddeley, A. (1986). *Working memory*. Oxford: Clarendon.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4, 417-423.

Bates, E., Camaioni, L., & Volterra, V. (1975). The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly of Behavior and Development*, 21, 205-226.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219-252.

Browman, C. P., & Goldstein, L. M. (1992). Articulatory phonology: An overview. *Phonetica*, 49, 155-180.

Bruner, J. S. (1975). The ontogenesis of speech acts. *Journal of Child Language*, 2, 1-19.

Buder, E. H., Chorna, L. B., Oller, D. K., & Robinson, R. B. (2008). Vibratory regime classification of infant phonation. *Journal of Voice*, 22, 553-564.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31, 149-180.

Chafe, W. (1987). Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984.* (Vol. 11, pp. 21-51). Amsterdam: John Benjamins.

Choe, W. K. & Redford, M. A. (2012). The distribution of speech errors in multi-word prosodic units. *Laboratory Phonology*, 3, 5–26.

Clark, E. V., & Hecht, B. F. (1982). Learning to coin agent and instrument nouns. *Cognition*, 12, 1-24.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17, 297-338.

Cooper, W., & Paccia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Cowan, N. (2000). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87-185

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech, Language, and Hearing Research*, 11, 707-721.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93, 283-321.

Dell, G., Juliano, C. & Govindjee, A. (1993) Structure and content in language production: a theory of frame constraints in phonological speech errors. *Cognitive Science*, 17, 149–195.

Dore, J. (1975). Holophrases, speech acts and language universals. *Journal of Child Language*, 2, 21-40.

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

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, 30, 139-162.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, 41, 29-47.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language,* 84, 474-496.

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language Production, Volume 1* (pp. 177-220). London: Academic Press.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. New York: Taylor & Francis.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, 103, 386-412.

Grabe, E., Post, B., & Watson, I. (1999). The acquisition of rhythmic patterns in English and French. In *Proceedings from the 17th International Congress of Phonetic Sciences*, (ICPhS-99, San Francisco), 1201–1204.

Grondin, S., Meilleur-Wells, G., & Lachance, R. (1999). When to start explicit counting in a time-intervals discrimination task: A critical point in the timing process of humans. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 993-1004.

Gross, L. (2006). Evolution of neonatal imitation. *PLoS Biology*, 4, e311.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102, 594.

Houde, J. (2015). Auditory feedback. In M.A. Redford (Ed.), *The Handbook of Speech Production* (pp. 267-297). Boston, MA: Wiley.

Huber, J. E. (2008). Effects of utterance length and vocal loudness on speech breathing in older adults. *Respiratory Physiology & Neurobiology*, 164, 323-330.

Kehoe, M., Stoel-Gammon, C., & Buder, E. H. (1995). Acoustic correlates of stress in young children's speech. *Journal of Speech, Language, and Hearing Research*, 38, 338-350.

Klapp, S. T. (2003). Reaction time analysis of two types of motor preparation for speech articulation: Action as a sequence of chunks. *Journal of Motor Behavior*, 35, 135-150.

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, 35, 162-179.

Kuhl, P. K., & Meltzoff, A.N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, 100, 2425-2438.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112–131). New York: Wiley.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Löfqvist, A. (1990). Speech as audible gestures. In W. J. Hardcastle & A. Marchal (Ed.), *Speech production and speech modelling* (pp. 289-322). Dordrecht: Kluwer Academic Publishers.

MacWhinney, B. (1985). Hungarian language acquisition as an exemplification of a general model of grammatical development. In D. I. Slobin (eds.), *The Crosslinguistic Study of Language Acquisition, 2* (pp. 1069-1155). Hillsdale, NJ: Erlbaum.

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, 25, 187-205.

McNeill, D. (2008). *Gesture and thought*. Chicago, IL: University of Chicago Press.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198, 75-78.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (Ed.), *Cognitive Skills and their Acquisition* (pp.), Hillsdale, NJ: Erlbaum.

Norman, D. A., & Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In R. Davidson, G. Schwartz, & D. Shapiro (eds.), *Consciousness and Self-Regulation: Advances in Research and in Theory, Volume 4* (pp. 1-18). New York: Plenum Press.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39, 151-168.

Payne, E., Post, B., Astruc, L., Prieto, P., & Vanrell, M. M. (2012). Measuring child rhythm. *Language and Speech*, 55, 203-229.

Pollock, K. E., Brammer, D.M., & Hageman, C.F. (1993). An acoustic analysis of young children's productions of word stress. *Journal of Phonetics*, 21, 183–203.

Prieto, P., Shih, C., & Nibert, H. (1996). Pitch downtrend in Spanish. *Journal of Phonetics*, 24, 445-473.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7, 226-231.

Ravizza, S. (2003). Movement and lexical access: Do noniconic gestures aid in retrieval? *Psychonomic Bulletin & Review*, 10, 610-615.

Recasens, D. (1989). Long range coarticulation effects for tongue dorsum contact in VCVCV sequences. *Speech Communication*, 8, 293-307.

Redford, M.A. (2014, April). Rhythm from reduction: The emergence of prosodic words in children's speech. *Workshop on Later Stages in Speech and Communication Development*, London, United Kingdom.

Redford, M.A., Foroughifar, Z., & Dilley, L. (2014, July). Constraints on prosodic phrasing in children's speech. *14th Conference on Laboratory Phonology*, Tokyo, Japan.

Redford, M.A., & Oh, G. (2015, August). Fixed temporal patterns in children's speech despite variable vowel durations. In Stuart-Smith, J., Scobbie, J., Turk, A. (eds.), *Proceedings from the 18th International Congress of Phonetic Sciences, (ICPhS-15 Glasgow)*, 5 pages.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82, 225.

Schwartz, J. L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25, 336-354.

Schwartz, R. G., Petinou, K., Goffman, L., Lazowski, G., & Cartusciello, C. (1996). Young children's production of syllable stress: An acoustic analysis. *Journal of the Acoustical Society of America*, 99, 3192-3200.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge, UK: Cambridge University Press.

Searle, J. R., & Vanderveken. (1985). *Foundations of illocutionary logic*. Cambridge, UK: Cambridge University Press.

Sirsa, H., & Redford, M.A. (2011). Towards understanding the protracted acquisition of English rhythm. In Lee, W.-S., Zee, E. (eds.), *Proceedings from the 17th International Congress of Phonetic Sciences, (ICPhS-11 Hong Kong)*, pp. 1862-1865.

Shattuck-Hufnagel, S. (2015). Prosodic frames in speech production. In M.A. Redford (Ed.), *The Handbook of Speech Production*, (pp. 419-444). Boston, MA: Wiley.

Shea, C. H., & Wulf, G. (2005). Schema theory: A critical appraisal and reevaluation. *Journal of Motor Behavior*, 37, 85-102.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84, 127.

Smith, A., & Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, 45, 22-33.

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition*, 95, 1-30.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information Processing in Motor Control and Learning* (pp. 117-152). New York: Academic Press.

Sternberg, S., Knoll, R. L., Monsell, S., & Wright, C. E. (1980). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica*, 45, 175-197.

Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry*, 42, 3-48.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66, 126-156.

Turk, A., & Shattuck-Hufnagel, S. (2014). Timing in talking: What is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 20130395.

Vihman, M. M. (2014). *Phonological development: The first two years*. Malden, MA: Wiley Blackwell.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18, 3-35.

Wheeldon, L. (2000). Generating prosodic structure. In L.R. Wheeldon (Ed.), *Aspects of Language Production* (pp. 249-274). Philadelphia, PA: Taylor & Francis.

Winkworth, A. L., Davis, P. J., Ellis, E., & Adams, R. D. (1994). Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors. *Journal of Speech, Language, and Hearing Research*, 37, 535-556.