Comprehension without segmentation:
A proof of concept with naive discriminative learning

R. Harald Baayen[a]
Cyrus Shaoul[a]
Jon Willits[b]
Michael Ramscar[a]

[a]Eberhard Karls University, Tübingen, Germany
[b]University of Indiana at Bloomington, USA

Abstract

All current theories of auditory comprehension assume that the segmentation of speech into word forms is an essential prerequisite to understanding. We present a computational model that does not seek to learn word forms, but instead decodes the experiences discriminated by the acoustic contrasts in the input. At the heart of this model is a discrimination learning network (Ramscar et al., 2010; Ramscar and Baayen, 2013), trained not on isolated words, but on full utterances. This network constitutes an atemporal long-term memory system. A fixed-width short term memory buffer projects a constantly updated moving window over the incoming speech onto the network's input layer. In response, the memory generates temporal activation functions for each of the output units. Output units (lexical contrasts, or lexomes) with high extended activation reflect a high degree of confidence that the cues that discriminate it from other possible lexomes are present in the external world. Lexomes that are not encoded in the signal give rise to little or no interference. We show that this new discriminative perspective on auditory comprehension is consistent with young infants' sensitivity to the statistical structure of the input. Simulation studies, both with artificial language and with English child directed speech, provide a first computational proof of concept and demonstrate the importance of utterance-wide co-learning.

keywords discriminative learning, auditory comprehension, word segmentation, phonotactics, Rescorla-Wagner equations

1 Introduction

The writing technology with which English and related languages encode speech in the form of structured patterns of ink has had a pervasive influence on the conceptualization of auditory comprehension. When rendering an utterance in written form in alphabetic writing systems, the speech signal has to undergo two processes of discretization: segmentation into a sequence of words, to be divided by spaces, and segmentation of these words into a sequence of letters. For auditory comprehension, it is likewise assumed that listeners have to segment the speech stream into phonemes, and segment the stream of phonemes into words. For example, the ShortList-B model (Norris and McQueen, 2008) characterizes lexical access in auditory comprehension as targeting a path in a word form lattice in which the word forms, represented by strings of phonemes, are properly lined up but without the spaces familiar from writing.

The absence of delimiters in the speech signal raises the question of how children learn where words begin and end, and how listeners partition of the speech signal into the correct sequence of word forms. For example, in Saffran et al. (1996) and many subsequent studies, children learn to segment the speech stream into words with the help of low-probability phonotactic transitions. Based on infants' looking behavior when presented with sequences of simple syllables, they concluded that with only 2 minutes of exposure, 8-month old infants segment the speech stream into words using only the statistical relationships between neighboring phonemes. According to Norris and McQueen, the correct segmentation into words is obtained by making optimal rational decisions on the basis of Bayesian probabilities that are continuously updated as the speech signal unfolds over time.

The present study outlines a completely different, non-decompositional, computational perspective on auditory comprehension. Our approach eschews the structuralist two-tiered perspective on language that is axiomatic for models such as Shortlist-B. According to Martinet (1965), a core design principle of language is its "double articulation". The structuralists and their descendents argue that on a first tier, sounds group together to form words, independently of meaning, and that at a second tier, words — the basic meaning bearing units — group together to form sentences. However, it is well known that this division of labor falls apart on closer inspection. The sign is not arbitrary (Bolinger, 1949), as becomes clear immediately to any student of onomatopea, sound symbolism, ideophones, and phonaesthemes. Moreover, phonaesthemes (e.g., *gl* in words such as *glow, glimmer, glitter, glisten*, and *gleam*, which all relate to light and its perception) show priming effects similar to those for regular morphemes (Bergen, 2004; Pastizzo and Feldman, 2009). More recently, Monaghan et al. (2014) provided further evidence for the non-independence of formal and semantic similarity in the lexicon. Whereas those committed to the double articulation of language will dismiss these findings as just not diagnostic, we accept, as an essential tenet of the scientific method, that evidence that inconveniences standard theory should not be not ignored, but rather that it should be used to develop more adequate theories. Instead of marginalizing these phenomena, we therefore take them as evidence against a two-tiered model of language.

Accordingly, our investigations examine what can be achieved when the relationship between form and meaning is the product of discriminative learning within a *system* of forms and meanings. This contrasts with traditional approaches in which this relationship is indirect, with mediating abstract representations such as phonemes and word forms. Earlier work Baayen et al. (2011) in this vein showed that for reading, a two-layer Rescorla-Wagner network correctly successfully accounts for a wide range of the effects observed in experimental studies of reading. In the present study, we extend their approach to lexical processing in auditory comprehension.

The algorithmic core of our model is a simple network architecture with two layers of localist representations with connection weights that are estimated with the help of the Rescorla-Wagner equations (Rescorla and Wagner, 1972).

1.1 The input layer: triphone cues

The input layer has units (henceforth cues) for n-phones. We have several reasons for opting for cues larger than the phoneme.

First, although the phoneme plays a prominent role in many models, its status as an abstract unit is highly problematic (see, e.g., Port and Leary, 2005). Phonemes are inadequate from the point of view of perception, so that, for instance, voiceless stops are discriminated primarily by contrasts in the formant transitions in adjacent vowels rather than by the contrasts that the phonemes posited for these stops are supposed to represent, while from the point of view of production, it has long been clear that a phonemic representation is inadequate to encode the detail of actual speech signals. In

order to do better justice to the pervasive consequences of co-articulation in the speech signal (see also Browman and Goldstein, 1992; Wickelgren, 1969), our input units span multiple phonemes, and critically, rather than assuming a binary mapping between notional representational units and phonetic "units", these units combine to enable continuous representational values to accrue across multiple input units in a discrimination learning network. In what follows, we make use of triphones as inputs, but other choices are demi-syllables or diphones.

Second, various units that are both larger and smaller than single phonemes or single letters are found in many other models of lexical processing (Taft, 1994; Levelt et al., 1999; Dell, 1986). These are often motivated by theories about the internal structure of the syllable. However, an important property of distributed representations for learning is that because informative patterns among cues develop competitively during learning, it enables us to investigate the development of a child's perception of a continuous phenomenon such as speech without having to overcommit to an ontology of discrete acoustic "units" that do not actually exist in the speech stream.

Third, the adoption of larger input units allows us to avoid the problem of overfitting: For example, if a model is built with letter unigrams and letter digraphs as cues, the digraphs have greater discriminative power than the letters: The *a* and *q* in *qaid* and *quad* do not tell these words apart, whereas the digraphs *qa* and *qu* do so perfectly. Due to cue competition, the learning algorithm will send the weights on the connections from *a* and *q* to *qaid* and *quad* towards zero, while strengthening the corresponding weights for *qa* and *qu*. Similarly, once trigrams are used, both unigrams and digrams become superfluous. In our experience, when working with English, triphones provide excellent discrimination without overfitting: longer n-phones would become too word-specific, causing the model to loose productivity. For other languages with different phonotactics, such as Vietnamese, diphones may be more appropriate than triphones (see Pham and Baayen, 2015, for the case of visual comprehension). In summary, using triphones instead of phonemes offers the important advantages of doing better justice to the continuous acoustic properties of speech, and of providing insight into the actual nature of the discrimination learning that takes place during the development of speech perception, without overfitting.

A nice, and very relevant example of the dangers posed by overfitting for theory is provided by the widespread belief in the existence of abstract perceptual categories inferred from, for instance, the phenomenon of categorical perception. Although there is a body of evidence that is consistent with abstract categories such as phonemes in the literature (see, e.g., Lisker and Abramson, 1964), it is not the case that this evidence can only be explained by assuming that discrete phoneme categories exist. Figure 1 illustrates that categorical discrimination can occur with a two-layer network trained with the Rescorla-Wagner equations. Two output units (henceforth outcomes) are connected to 20 input cues that partition a phonetic continuum x. The two outcomes are characterized by distributions on x with different means (4 and 6) and the same standard deviation (top panel). A constant background cue and some 20 random cues representing further sources of variation were added. As shown in the bottom panel of Figure 1, the model's support for the outcomes (its activation, see below for further details on Rescorla-Wagner networks) shows the typical cross-over pattern characterizing categorical perception. It is important to note here that the outcomes can be any of the words that differ minimally on a phonetic continuum, such as the frication in the segments *f* and *s*. The network can be extended with any number of additional pairs differing on this continuum, and the network will show categorical discrimination for each pair, without the help of abstract phonemic categories.

This network model is also able to explain the disambiguation of ambiguous segments and adaptation, two other phenomena that are put forward as evidence for the existence of "categorical perception". For example, McQueen et al. (2006) trained subjects on ambiguous constructed segments that were midway between *f* and *s*. When subjects are trained on words in which the

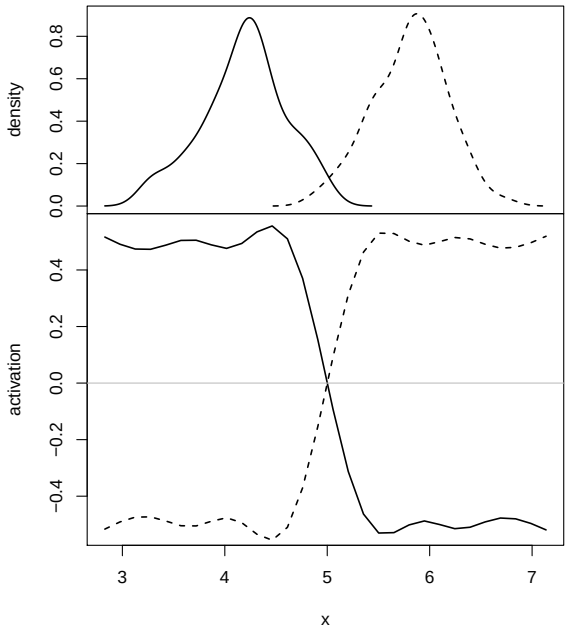

Figure 1: Categorical perception in a discriminative learning model with 20 cues partitioning a phonetic continuum x and two outcomes with distributions (top panel) with different means (4 and 6) on this dimension, and the same standard deviation (0.5). The bottom panel shows activation of the two outcomes as a function of x.

f is replaced by this ambiguous segment, and are simultaneously also trained on words with an unambiguous *s*, the ambiguous segment affords facilitation when used as a prime in a subsequent lexical decision task. Of special interest is that this effect was observed for words that had not been presented during training. The authors take this as evidence for the modulation of abstract prelexical phoneme categories.

However, the effect arises straightforwardly in a discriminative framework. We defined a simple lexicon with two pairs of words contrasting *f* and *s*. Using the Dutch examples of McQueen et al. (2006), we refer to the first pair as *naaldbos* and *witlof*, and the second pair as *doof* and *doos*. We first trained the Rescorla-Wagner network on a 100 tokens of each of these four words, using the distributions illustrated in Figure 1. We then trained the network with the first pair of words (*naaldbos* and *witlof*), using the same (unambiguous) distribution for *s*, but making the *f* ambiguous by narrowing down the distribution for *f* to a small interval around the midpoint of the two distributions (5). Finally, we examined the predictions of the network at this midpoint value for the second pair of words (*doof* and *doos*), that had not been seen during the second training phase. After the second training phase, the probability of *doof* at the midpoint of the phonetic dimension phase was significantly higher than it was after the first training phase. In other words, adaptation took place in the network, exactly in the direction observed by McQueen et al. (2006), even though there were no prelexical phoneme units in the model. These simulation examples show how discrimination learning can provide an alternative perspective on the categoricity of perception that does not require the psychological reification of phonemes. This point is particularly salient given that although empirical studies of categorization offer abundant evidence of the remarkable flexibility of human discrimination learning, this evidence does not provide conclusive support for the idea of discrete categorical representations . Further, while the computational and neuroscience literatures on categorization both provide consistent evidence in support of the kind of processes simulated in

the example above, they also provide a wealth of reasons to doubt the neuropsychological reality of the discrete categorical units that are thought to be driving categorical perception (Ramscar and Port, 2015).

We should note also that the triphones that we use as input cues for our model are by themselves too coarse to capture the full range of low-level phonetic detail (see, e.g., Gaskell and Marslen-Wilson, 1996; Davis et al., 2002; Salverda et al., 2003; Ernestus and Baayen, 2006; Kemps et al., 2005a,b). One solution would be to specify lower-level models discriminating between triphones on the basis of phonetic cues. Another option would be to replace triphones by acoustically motivated cues (see, e.g., Gold and Scassellati, 2006). We leave these issues for further research.

1.2 The output layer: lexome outcomes

The output layer contains units that we refer to as lexomes. Following Milin et al. (2015), we define the lexome as a theoretical construct at the interface of language and a world that is in constant flux with the flow of experience. Lexomes are the lexical dimensions in the system of knowledge that an individual acquires and constantly modifies as the outcome of discriminatively learning from experience within a culture. Because lexomic contrasts serve as communicative counterparts to the specific experiences individuals and cultures discriminate for practical and communicative purposes, they can be evoked in context either by language use or real world experience. Accordingly, the more that a lexome is activated in a given context, the greater the degree of confidence that the cues that culturally discriminate it from other outcomes are present in the external world. Lexomes can be compared, following De Saussure (1966, p. 88), to the pieces in a game of chess.

> ... a state of the set of chessmen corresponds closely to a state of language. The respective value of the pieces depends on their position on the chessboard just as each linguistic term derives its value from its opposition to all the other terms.

The strategic value of a pawn depends on where it is on the board and the positions of the other pieces. Similarly, the value of a lexome such as WALK depends on the other lexomes encoded in the language signal, and their relation to the other lexomes and experiences in a speaker or listeners' current system of knowledge. In an utterance such as *I'll walk home*, the lexome WALK thus correlates with a cultural and behavioral discrimination between between bipedal locomotion and other ways of transportation. In *I'll walk the dog*, the same lexome, together with the lexome DOG, discriminates between the daily exercise regime required to keep a dog healthy, and other activities such as recreational walking.

Expectations about the strategic moves that might unfold in a game of chess change with each turn taken. Similarly, a listener's expectations about the experiential contrast being communicated in a signal change with each next word read or listened to: when *home* follows *I'll walk*, both the structure of the available cues and the outcomes discriminated by the actualized aspects of the message change, which has the effect of altering the interpretation of the signal as compared to when *the dog* follows. A lexome is, in some ways, similar to the lexeme as defined by Aronoff (1994, p. 11) as 'a (potential or actual) member of a major lexical category, having both form and meaning but being neither, and existing outside of any particular syntactic context'. However, from a theoretical perspective, it is important to note that we conceptualize lexomes within a discriminative account of meaning, in which signals serve to reduce a listener's uncertainty about what a message means (a process that relies on having learned a predictive system of lexical and conceptual contrasts, Ramscar et al., 2010). Given the discriminative framing of lexomes, we should note that unlike lexemes, the term lexome can apply to any systematic lexicalized contrast, including grammatical dimensions such as tense, aspect, and number, and combinations thereof.

In this light, it is worth emphasizing that lexomes are not containers of meaning, even though in English and related languages structural metaphors are pervasive that see language as a conveyor belt transporting boxes with meanings from speaker to listener (Reddy, 1979). There are many good reasons to believe that meanings do not reside in words or sentences (Ramscar et al., 2010; Ramscar and Port, 2015). Accordingly, we do not assume that learners are faced with the task of associating word forms with concepts, but rather, we see language learning as a systematic process that occurrs continuously in context, such that language learners simultaneously master both the relevant distinctions in their environments along with the lexical distinctions with which they correlate. To reflect this, in the model we present below, the weights on the n-phone units feeding into a lexome are subject to continuous change. We assume that this holds just as well for the experiences (at least those that we have learned to discriminate in the world) that are associated with any given lexome (Ramscar et al., 2013b,a,d), even though in our simulations we do not address this aspect of the dynamics of learning. In other words, the 'scope' of a system of lexomes — and the lexomes within it — changes constantly with experience, both with respect to the objects and events in the world, and with respect to the phonetic cues, which are constantly being updated while speaking and listening.

1.3 Estimating weights and activations

Each n-phone cue is connected to every lexome. Connection strengths (weights) are estimated with the learning equations of Rescorla and Wagner (1972), which specify the following recurrence relation in discretized time for a weight w_{ij} from cue i to outcome j at time $t+1$,

$$w_{ij}^{t+1} = w_{ij}^{t} + \Delta w_{ij}^{t}. \tag{1}$$

Weights are updated for the learning event at times $t = 1, 2, \ldots$. A learning event comprises a set of unique cues and a set of unique outcomes. (If a cue or outcome occurs more than once, it is included only once.) Let $\textsc{present}(C_i, t)$ denote the presence of a cue C_i in a given learning event E_t taking place at time t, and let $\textsc{present}(O_j, t)$ denote the presence of outcome O_j in E_t. The weight w_{ij}^t from C_i to O_j at time t is updated at $t+1$ by Δw_{ij}^t, defined as

$$\Delta w_{ij}^{t} = \begin{cases} 0 & \text{if } \textsc{absent}(C_i, t), \\ \alpha_i \beta_1 \left(\lambda - \sum_{\textsc{present}(C_k, t)} w_{kj} \right) & \text{if } \textsc{present}(C_i, t) \;\&\; \textsc{present}(O_j, t), \\ \alpha_i \beta_2 \left(0 - \sum_{\textsc{present}(C_k, t)} w_{kj} \right) & \text{if } \textsc{present}(C_i, t) \;\&\; \textsc{absent}(O_j, t). \end{cases} \tag{2}$$

Here, α_i denotes the salience of the cues, and β_j the strength of positive versus negative learning. The parameter λ denotes the maximum amount of learning. In all models reported below, we use the default values $\alpha_i = \beta_j = 0.1$ and $\lambda = 1.0$.

The activation $a_{O_j,t}$ of an outcome (in the models below, a lexome) at time t is given by the sum of the weights on the connections from the cues $\{C_t\}$ in the input signal at t to that outcome:

$$a_{O_j,t} = \sum_{C_i \in \{C_t\}} w_{ij}^{t}. \tag{3}$$

Unlike in standard connectionist models, we do not make use of activation functions such as sigmoid squashing functions or hyperbolic tangents to normalize activations. There are no hidden layers of any kind. As we do not use backpropagation, there is no need to ensure differentiability. At any particular point in the continuing development of the system of lexomes, a high activation of a specific lexome simply reflects a high degree of confidence that the cues that discriminate it from other lexomes are present in the external world (compare the fifth dogma of Barlow, 1972).

A Rescorla-Wagner network can be trained in two ways. Both training regimes have in common that a set of learning events has to be defined. For instance, an utterance in a film subtitle corpus can be taken as a learning event, with trigraphs or triphones as cues and lexomes as outcomes. When there is an intrinsic order to the learning events, equation 2 can be used, updating weights for each successive event. When there is no such intrinsic order, an alternative is to use the equilibrium equations for the Rescorla-Wagner equations developed by Danks (2003). In the present study, we use the equilibrium equations only for simulated data without order to the learning events. For such data, they show the endstate of learning, after infinite learning experience. For real data, weights are best estimated learning event by learning event.

1.4 Temporal dynamics

The temporal dynamics of auditory comprehension, as revealed by, e.g., gating tasks and the visual world paradigm (Grosjean, 1980; Salverda et al., 2003) arise in the model as a consequence of the temporal unfolding of the speech signal over time. Although the Rescorla-Wagner network itself captures minimal temporal sequences by means of the triphone cues, no other mechanisms are built in to represent time. Thus, there is no recurrent hidden layer as in the PDP model of Gaskell and Marslen-Wilson (1997a), nor are networks duplicated for successive time steps as in the TRACE model. For practical and explanatory purposes, the network itself (as evolved after a given number of learning events) can be conceptualized as a-temporal. However, when a speech signal is presented incrementally to the network, the cues in the input change as time unfolds. This, in turn, leads to changes in the extent to which different lexomes are supported by the input. We return in more detail to this aspect of our model below.

1.5 Discrimination instead of hierarchical decomposition

Fundamental to our approach is the argument that it is counterproductive to seek to segment the speech signal into a hierarchy of increasingly smaller bits of signal. The deconstruction of the signal into hierarchies of form units is fundamentally at odds with the central insights of information theory (Shannon, 1948, 1956). To see this, consider Figure 2. Four experiences of the world, one of a fountain, one of a fountain pen, one of an orange, and one of a glass of orange juice, are coded with two binary digits. Table 1 lists the amount of information in each picture, estimated by the file size of the jpg pictures. Only two bits are required to discriminate between these four experiences (see Table 1). Notably, the experiences (i.e., the pictures in Figure 2) are much more complex than the simple code that can discriminate between them.

Importantly, the four two-bit signals can be randomly assigned without the communication code incurring a loss in effectiveness. Even more importantly, it is not necessary to decompose the code into 'meaning-ful' parts. In Figure 2, one could seek to interpret the first zero as a linguistic sign for fountain, that then re-occurs in fountain pen. Likewise, an initial one could be taken to signify oranges. However, objects referred to as fountain pens are not compositional functions of objects referred to as fountains. The function of the initial zero is to discriminate between the experiences on the top row and those on the bottom row of Figure 2. If we reverse the order of the digits, an initial zero eliminates fountain pens and orange juice from the set of experiences potentially encoded in the signal, and the second digit then eliminates remaining uncertainty with respect to fountain and orange. In short, the successive digits in the binary code serve to zoom in on the experience encoded in the signal by successively eliminating the other experiences that the code supports.

In current standard theories of lexical processing, by contrast, the signals (digits/constituents) encoding fountain pen (or orange juice) are first split into their constituents, which supposedly

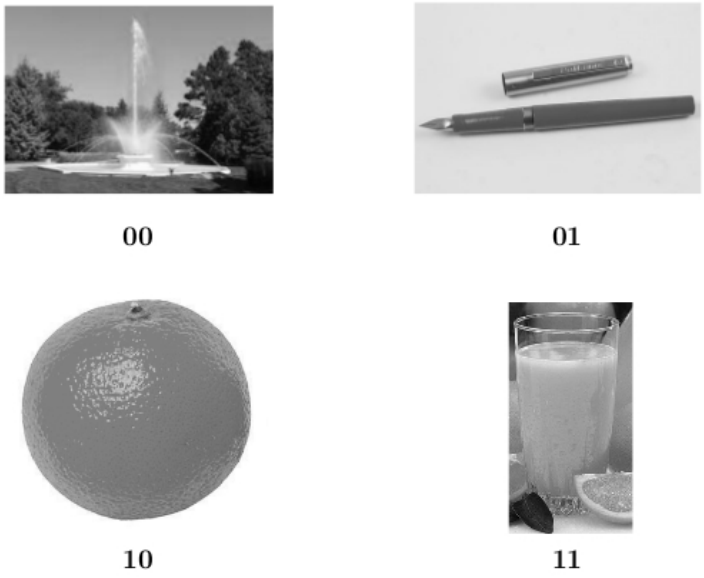

Figure 2: Binary coding for four experiences (images from Wikipedia).

signal	size	experience	compression size
00	2 bits	fountain	9.3 kB
01	2 bits	fountain pen	11.7 kB
10	2 bits	orange	11.8 kB
11	2 bits	orange juice	14.6 kB

Table 1: Binary coding for four experiences with different information loads, estimated by the file compression sizes.

activate their constituent meanings (fountains and pens, or oranges and juice). However, since fountains do not contribute in an obvious way to the semantics of fountain pens, one then has to assume that the constituents provide pointers to the intended meaning of fountain pen.

Intuitively, it seems entirely rational to think that of course a fire engine is a truck carrying equipment for putting out fires, but speakers of Vietnamese feel equally rational about the juxtaposition of fire and engine denoting the vehicle that produces the power that pulls a train. Languages are economical in their use of lexical forms, and re-use words in other words. Since reusing shorter sequences in longer sequences cannot be avoided (1 reoccurs in two positions in 11), it makes sense that as languages evolve, this re-use has some partial, albeit from a cross-linguistic perspective, clearly idiosyncratic, motivation. But, as folk etymologies illustrate, the bewilderment that we often experience at why complex words are what they are, can lead to explanations that have little to do with the actual historical origins of a given onomasiological convention (e.g., 'fountain' in 'fountain pen' originally denoted the reservoir in which ink is stored). However, folk etymologies as well as our rationalizations of the supposed logic of language have little to say about how the language code actually works. Instead, they are informative about our cultural pre-conceptions about the supposed logical nature of our language.

When a video camera records a fountain pen and its cap (as in the upper right panel of Figure 2), and communicates the recording to a display screen through an electrical wire, it is not the case

that the electrical signal in the wire first decompositionally transmits the pen and then its cap. The electrical signal encodes, to the outside observer, encrypts, the visual scene using an error-corrected optimized code that transmitter and receiver share, and which allows the display screen to discriminate the steps that result in the reproduction of the recording. It is this code, the set of algorithms that make it possible for speakers to use linguistic signals to discriminate the various experiences they wish to communicate about, that we believe is central to a proper understanding language and language processing.

With its rejection of any segmentation operations on the signal, our approach distinguishes itself from other computational models of lexical access in auditory comprehension. For instance, both the TRACE model (McClelland and Elman, 1986), and Shortlist-B are supplied with a lexicon with pre-segmented word forms and their frequencies. Both models are designed to recover word forms and their order from a stream of phonemes obtained by concatenation of word forms. Neither model offers insights as to how these word forms are learned. A similar problem arises with the recurrent PDP model of Gaskell and Marslen-Wilson (1997b), which represents the speech signal as a sequence of phonetic feature bundles, which is paired, bundle by bundle, with matching bundles of semantic features. How this model comes to know about this pairing of phonetic feature bundles and semantic feature bundles is left unexplained. Furthermore, it is unclear what exactly the semantic features actually represent (Ramscar and Port, 2015), or how the coupling of these constructs to word forms is supposed to take place in learning. In fact, given that the effectiveness of features as diagnostics for categories is subject to continuous modulation (see, e.g., Love et al., 2004; Marsolek, 2008; Ramscar and Port, 2015), the assumption of a time-invariant, Platonic, semantic vector prototype is highly implausible.

In our model, word forms are never learned. Instead, learners acquire and learn to use a lexical *system*. As we shall see, not only is it not necessary to learn words forms, it is even counterproductive to do so. Much of the "heavy lifting" that can make language acquisition seem so puzzling when considered as a word-at-a-time process is actually a straightforward product of this system. Importantly, it does not make sense within our approach to hope for form representations being implicitly coded in the connection weights. The forms themselves have no theoretical relevance whatsoever in the model. There is no need to re-represent the signal 'internally': It is the (rich) experiences of the world that we have learned to discriminate between (such as the pictures in Figure 2) that the comprehension system is decoding from the signal.

Of course, training in literacy adds further layers of complexity, with knowledge of words' orthographic forms generating expectations about corresponding phonological forms. These added complexities are beyond the current scope of our model, which addresses the learning of auditory comprehension before the onset of literacy.

Of the substantial literature on segmentation in auditory processing (see, e.g., McQueen et al., 1994, 1995; Vroomen and De Gelder, 1995; Johnson and Jusczyk, 2001), the study by Saffran et al. (1996) has been particularly influential. These authors obtained evidence congruent with the possibility that young infants are using transition probabilities between phonemes (or other sound units) to segment the speech stream into words. Like Saffran et al., we agree that their results demonstrate impressive learning capabilities of young infants, and suggest that experience-dependent (i.e., learning) processes have been underappreciated in many theories of language acquisition. However, we argue that taking a "discriminative" stance — rather than a "decompositional" stance as is commonly assumed by most research — may offer a better characterization of the language acquisition problem.

This paper is not intended to be a comprehensive presentation of our approach to lexical processing. However, whereas previous work in this area addressed visual comprehension (Baayen et al., 2011, 2013; Milin et al., 2015), the present study outlines a very simple but highly effective

computational architecture for auditory comprehension that is inspired by the discriminative stance.

In what follows, we first discuss the phenomenon of low-probability phonotactic transitions (n-phone troughs) by means of a series of simulation studies using artificial grammars. We then clarify how the evidence from infant looking behavior that appears to support segmentation can be understood from the perspective of discrimination learning. We then illustrate, using the English child-directed speech in the CHILDES database (MacWhinney, 2000), how comprehension can proceed perfectly well without segmentation.

2 Segmentation and discrimination

Within-word phoneme transition probabilities tend to be higher than between-word phoneme transition probabilities. Low transitional probabilities have been put forward, together with prosodic and co-articulatory information, as cues for segmenting the speech stream into words (Christiansen et al., 1998; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Saffran et al., 1996), and for segmenting words into their constituent morphemes (Seidenberg, 1987; Hay, 2002, 2003).

From a discriminative perspective, low-transitional probabilities are not 'separators', but 'binders': They are excellent cues for discriminating between lexomes. Consider the word sequence *klejpɔt*, *clay pot* , i.e., a pot made of clay. Of the triphones for this word pair, *kle, lej, ejp, jpɔ, pɔt*, the first two are unique to *clay*, the last is unique to *pot*, and the third and fourth are unique to the phrase. Since *clay* and *pot* are much more frequent than *clay pot*, the cues *kle, lej* will develop strong weights for *clay* and weak or even negative weights to *clay pot*. Similarly, the cue *pɔt* will predict *pot*, but will provide only weak evidence for *clay pot*. By contrast, the low-frequency cues *ejp, jpɔ* will be learned to support *clay pot*. They constitute the only evidence in the signal that supports the specific meaning 'pot made of clay'.

Decompositional theories first segment *klejpɔt* into *klej* and *pɔt*. At this point, these theories have to deal with the problem that the meaning of *clay pot* is not a-priori predictable from the meanings of its parts — a *clay pot* could also mean a pot for storing clay. As a consequence, decompositional theories are forced to view *clay* and *pot* as pointers in a hash table to 'a pot made of clay'. By first taking the signal apart and then putting it together again, processing becomes much more complex than it need be: the boundary n-phones *ejp, jpɔ* provide exactly the critical information for targeting the appropriate interpretation. Since many words have highly context-dependent meanings (compare *eat your porridge* with *eat your hat*), segmentation into words systematically ignores valuable information in the signal, and gives rise to exacerbated problems of disambiguation at 'post-lexical' stages of processing.

How does this approach deal with novel words such as *polka-dot dingo*? Let us assume that the constituents *dingo* (an Australian wild dog) and *polka-dot* (being covered with colorful circles or spots, as in *polka-dot dress*, *polka-dot plant*, and *polka-dot man*) have been encountered. In this case, *polka-dot* is associated with three lexomes, one discriminating between dresses patterned with large colored dots and other dresses, one discriminating between plants with spotted leaves and other plants, and one discriminating between a criminal in batman comics wearing an outfit with large colored circles and other men. The simultaneous availability of these three lexomes, and the lexome for *dingo*, is what our current implementation produces. Returning to de Saussure's chess metaphor: Our model identifies several potential pieces, but it necessarily remains silent about how these pieces contribute to the game, as this depends on the other pieces and their positions in the game. In other words, context and the prior experience of an interpreter have to be taken into consideration.

Depending on this prior experience, modification of *dingo* by *polka-dot* might reduce uncertainty

to dingos characterized either by spots or by colorful circles. In most cases, because an interpreter's prior learning will result in a given context implicitly reducing the likelihood of alternative interpretations such that one interpretation becomes most salient, context will suffice to bias the interpreter to one interpretation or the other (see Ramscar et al., 2013d, for an example of the powerful, and surprisingly uniform effects of this kind in adult paired associate learning). However where sufficient ambiguity still remains, we assume that higher-order reasoning processes (see, e.g., Ramscar et al., 2013b) will guide an interpreter in the direction of either a dingo with mud spots, or a dingo in a comic with brightly colored dots.[1]

In what follows, we first present a series of simulation studies illustrating why segmentation is not necessary and non-optimal. We also clarify why it is impossible to bootstrap word boundaries from transition troughs only. We then explain, using discriminative learning, why infants respond behaviorally to transitional troughs.

2.1 The non-optimality of segmentation

To illustrate the disadvantages of segmentation, we consider a simple artificial language. The design of this language is inspired, to some extent, by Vietnamese (see, e.g., Pham and Baayen, 2015). As in Vietnamese, words in this language consist of one or two syllables. Each syllable has a highly constrained CV structure, limited here to the pattern CCVC. The first consonant was selected randomly from the set $\{p, t, k, b, d, g\}$, the second consonant was selected from the set of fricatives $\{f, s, x, v, z, G\}$. The vowel was one of the 5 cardinal vowels $\{a, e, i, o, u\}$, and the final consonant was selected randomly from the set $\{p, t, k, b, d, g, f, s, x, v, z, G, r, l, h\}$. A total of 100 monosyllabic words was generated, and assigned frequencies sampled from a lognormal(4,2) distribution. Next, a total of 900 two-syllable words was constructed by concatenation of two syllables sampled from the monosyllabic words, with a probability proportional to their frequency. The sampling frequencies of these 900 two-syllable words were combined with frequencies sampled from a lognormal(4,2) distribution. This resulted in a list with 100 monosyllabic and 900 bisyllabic words. Word frequencies and syllable family sizes approximately followed Zipf's rank-frequency power law.

Forms	Lexemes	Parse
pfehdvazdGatpsugtGap	100, 837, 924	pfeh+dvazdGat+psugtGap
tGupgvalgsukdvazkzuptsok	340, 745, 493	tGupgval+gsukdvaz+kzuptsok
dvoskzuppzehtfiGbxuxksub	773, 982, 533	dvoskzup+pzehtfiG+bxuxksub
pvopdsobgsukdsazpzizksub	892, 189, 898	pvopdsob+gsukdsaz+pzizksub
dviGdvazpzehtfiGbfahpvop	998, 982, 801	dviGdvaz+pzehtfiG+bfahpvop
pzizgvaldviGksubbsusdzel	694, 677, 312	pzizgval+dviGksub+bsusdzel

Table 2: Phrase forms, lexomes, and segmentation for simulation 1.

A total of 500 three-word phrases was generated by randomly selecting three words from the list of words, in proportion to their frequency. These 500 phrases are all that we make available for learning. The list of words itself, from which the words in the phrases were sampled, was withheld.

[1] Note that we associate the modifier *polka-dot* with three different lexomes. When a particular lexome is experienced especially frequently (e.g., in a series such as *polka-dot dress, polka-dot shirt, polka-dot pants, ...*), it will acquire stronger associations during learning, and hence will dominate understanding. This is how our approach explains the experimental results that CARIN theory (Gagné and Shoben, 1997; Gagné, 2001) accounts for by means of abstract decontextualized concepts (such as "polka dot") and an associated probability distribution over a set of abstract conceptual relations.

Table 2 lists examples of the phrases, their constituent lexomes (indexed by integers), and the segmentation of the phrases into word forms. Of the 88 constituents in the complex words, 18 are bound stems that occur in at least one other word (compare English *mit* in *transmit, commit, emit, submit*) and 7 are cranberry morphs that are attested in only a single complex word (compare *cran* in English *cranberry*). The phrases were assigned a uniform frequency distribution. The task for a computational model is to discriminate the lexomes based on the information in the signal, i.e., from the unsegmented phrases, without any further information such as a the original list of word forms.

First consider what might be done using a segmentation-driven approach. For this particular simulated language, phonotactic constraints on words provide very strong cues for syllable boundaries: A boundary follows the initial C in any CCC sequence. However, syllables have to be grouped into words. The problem that has to be addressed is that many of the phrases can be segmented in multiple ways (median: 3). For instance, the third phrase in Table 2 has five different segmentations:

dvos kzuppzeh tfiG bxuxksub
dvos kzup pzehtfiG bxux ksub
dvos kzup pzehtfiG bxuxksub
dvoskzup pzehtfiG bxuxksub
dvoskzup pzehtfiG bxux ksub

As a first step, one could select that parse for which the product of the sample probabilities of its constituent is maximal. The resulting proportion of correctly selected segmentations is 0.322. Accuracy can be improved to 0.978 by calculating the probabilities of word forms on the basis of their occurrences across all possible segmentations, and then selecting that parse for which the product of these constituent probabilities is greatest. (The resulting accuracy is identical to the accuracy obtained when the population probabilities of the constituents in the original list of words are used.)

Thus, given a simulated language with highly restricted phonotactics, and a correct guess about the syllable structure, probabilistic reasoning makes it possible to get the word forms right almost all of the time. Given a one-to-one mapping of word forms to lexomes, this high accuracy extends to the identification of the lexomes.

What can be accomplished by capitalizing on low transitional diphone probabilities as segmentation cues? The crucial question here is what low is. Figure 3 illustrates that as the threshold for a 'low' frequency diphone is increased, the number of correctly detected boundaries increases (left panel) to its maximum (2), as expected. At the same time (center panel), the number of spurious boundaries increases as well, and more rapidly to a higher number. The proportion of correctly identified boundaries is highest for thresholds around 60 occurrences, and then deteriorates. The highest proportion of correct syllable boundaries is 0.25. Unfortunately, there is not a single instance across the 500 phrases for which both boundaries are identified correctly. The problem is that languages typically come with many low-frequency segment transitions that are not boundary transitions. For any given frequency threshold, boundary transitions with a frequency exceeding the threshold will not be available for segmentation, resulting in actual word boundaries being missed. Conversely, non-boundary transitions below the threshold will give rise to spurious word boundaries. Bootstrapping from phonotactics only simply does not work.

Very different results are obtained with discriminative learning. Using the NDL package (Shaoul et al., 2013) in R version 3.0.2 (R Core Team, 2014), a Rescorla-Wagner network, with weights estimated by the equilibrium equations, was trained on the 500 phrases. This network predicts, when presented with all the cues of a given phrase (below, we will consider the temporal dynamics), the highest activations for each of the three words across all 500 phrases. Clearly, subword cues can

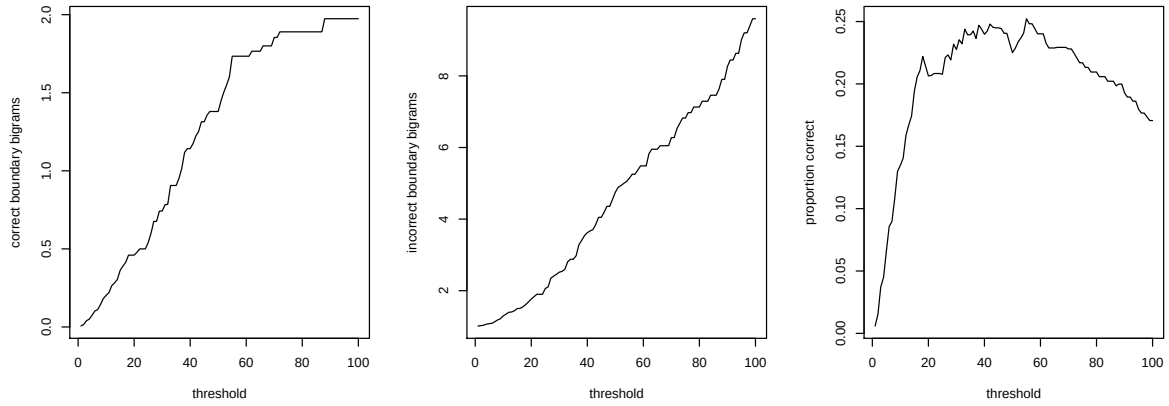

Figure 3: Accuracy of detection of word boundaries as a function of frequency threshold.

discriminate perfectly between the lexomes that are encoded in the signal, and those that are not. This illustrates that the scope of error-driven learning is not restricted to animal learning (Rescorla, 1988) but extends to the challenges encountered in human learning (Ramscar and Yarlett, 2007; Ramscar et al., 2010, 2011, 2013a,c, 2014). .

Forms	Lexemes	Parse
fubaerouboggGoGvaha	176, 175, 37	fubaero+uboggGoG+vaha
fubagGoradaotuadaGebe	505, 922, 665	fubagGor+adaotu+adaGebe
isorkoxoosogGoGodas	74, 827, 891	isor+koxooso+gGoGodas
kxoGgokurivukiisahkiG	785, 754, 825	kxoGgok+urivuki+isahkiG
gGokaxaGgGoksufi	77, 933, 83	gGok+axaGgGok+sufi
ivefubavahasufi	187, 37, 83	ivefuba+vaha+sufi

Table 3: Phrase forms, lexomes, and segmentations for simulation 2.

Let's now consider a simulated language with more variable phonotactics. Table 3 provides examples of phrases generated using a word list in which simple words can have not only CCVC structure, but also CVC, CVCV, VCVC, or VCV structures. Again, a Rescorla-Wagner network assigned the highest activations to the correct lexomes across all 500 phrases.

Does discriminative learning scale up? Using the same varied phonotactics, we increased the number of simple words to 2700, the total number of words to 30,000, and the number of phrases to 10,000. For 94.5% of the phrases, the model correctly predicts the highest activations for the lexomes encoded in the signal, and for 99.4% of the phrases, the three correct lexomes are among the top four most highly activated lexomes.

By contrast, the percentage of correctly identified boundaries on the basis of low-probability transitions, for the optimal threshold, is a mere 0.4%. As before, none of the phrases is correctly segmented. We anticipate that more sophisticated segmentation induction techniques such as adaptor grammars (see, e.g., Synnaeve et al., 2014) will yield much better performance.

Adaptor grammars make assumptions about the grammar generating the phrases. We therefore also considered a simulated data set where all information useful to adaptor grammars is removed. For this final set of phrases, words have no phonotactic structure whatsoever. Instead of assigning a lognormal distribution to word frequencies, word frequencies follow a uniform distribution. Fur-

thermore, a random half of the phrases have four words instead of three, obtained by splitting one two-syllable word into two one-syllable words. Under the assumption that an adaptor grammar gets all the syllable boundaries right, 92.2% of the segmentations can be reconstructed. The accuracy of our Rescorla-Wagner network is at 100%.[2]

This final simulation illustrates that phonotactic restrictions are not necessary for making sense of the signal. Phonotactic restrictions arise due to constraints on the coordination of our articulators in speech production. Similarly, a Zipfian power law is not necessary for discriminative learning to be effective. Word frequency distributions follow, albeit typically only approximately (see, e.g., Baayen, 2001), a power law because the events, states, objects and properties in the world tend to follow power laws (see, e.g. Good, 1953; MacArthur, 1957). Since discriminative learning as formalized by Rescorla and Wagner benefits from diversity in the signal, the comprehension-external forces shaping and condensing the lexicon actually render discrimination in comprehension more difficult: Words become more similar than they would have been otherwise, and phrases become more ambiguous.

2.2 Low-probability phonotactics and infant looking behavior

We have seen that Rescorla-Wagner networks are able to discriminate the lexomes from the signal with very high accuracy, whereas theories assuming that segmentation into words is the gateway to understanding perform less well. Bootstrapping word forms from troughs in transitional probabilities was shown to be especially problematic. This raises the question of why young infants are paying attention to low-probability phonotactic transitions (Saffran et al., 1996). Several models have been put forward that demonstrate that transitional probabilities can inform the discovery of word boundaries (see, e.g., Cairns et al., 1997, for an implementation using a recurrent network). In what follows, we show that the data of Saffran et al. are equally consistent with a scenario in which infants are not seeking to discover word boundaries at all. In this scenario, their looking behavior reflects the unexpectedness of syllables and the concomitant stronger learning experience requiring greater adjustments of weights in the network.

To illustrate this point, we constructed simulated data that approximates the experimental design of Saffran et al. (1996). A total of 440 CV syllable tokens (representing 15 syllable types) was presented one after the other to a Rescorla-Wagner network (*ba sa hi bo si ho bi se he bu ...*). Some syllables were always followed by exactly the same next syllable (e.g., *ba* was always followed by *sa*). Some syllables were followed by one syllable in two thirds of the cases, and by another in one third of the cases (e.g., *ha* was followed by *bi* two thirds of the time, and by *bo* one third of the time). Finally, some syllables were followed by any of three syllables with equal probability (e.g., *hi* by *be, bo, bu*). The task of the network was to predict the next syllable (the outcome) given the current syllable (the cue).

The rationale for this set-up of the simulation is that infants participating in experiments such as described by Saffran et al. (1996) are listening to a sequence of meaningless syllables. We assume that the minima in spectral energy in the speech signal demarcate boundaries on the individual speech events. In other words, we assume that the infants are sensitive to syllable identity. In the absence of any meaningful communication taking place in the course of the experiment, the implicit learning system predicts upcoming syllables. At each subsequent syllable, we adjust weights according to the Rescorla-Wagner equations.

[2] Milin et al. (2015) report a computational modeling study in which a Rescorla-Wagner network was trained on 4.8 million utterances (subtitles accompanying movie scenes) from an English subtitle corpus. With only a year's worth of reading experience (some 22 million word tokens, 84,000 types) the model correctly predicted the highest activations for all lexomes in an utterance for 65% of the utterances.

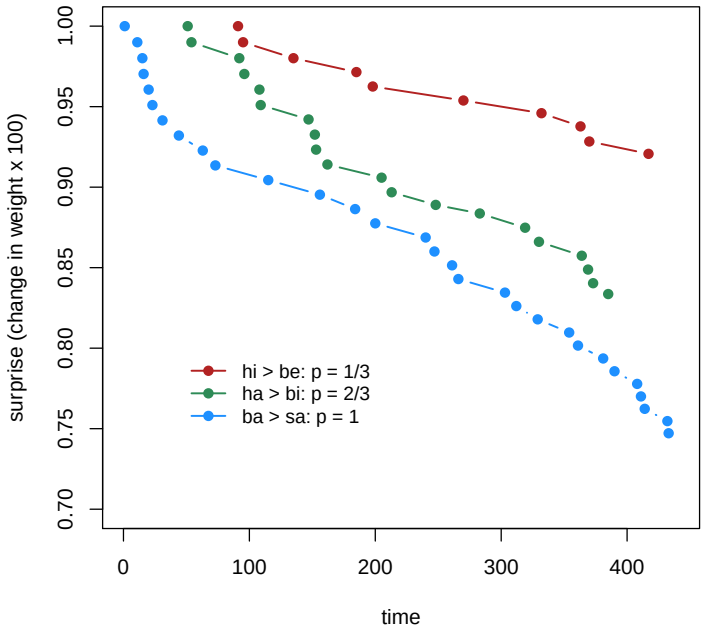

Figure 4: Surprise, measured as change in weight ($\times 100$) for a Rescorla-Wagner network ($\lambda = 1, \alpha = 0.1, \beta = 0.1$) with the current CV syllable as cue and the next CV as outcome, for three CV syllables with transitional probabilities of 1/3, 2/3, and 1, across 440 learning events. Time on the horizontal axis is in learning event units.

Figure 4 summarizes the changes in the weights. These reflect the model's surprise about its prediction error, as it develops over the course of the experiment. The model results mirror the data from Saffran et al. closely. For the syllable transitions with probability 1, the weight adjustments decrease most quickly. For the most uncertain transitions, the adjustments in the weights decrease slowly. The transitions with medium uncertainty pattern in between. Since the surprise at having made a wrong prediction is greatest for the low-probability transitions, it is no wonder that infants look at these more. There is strong evidence that the type of implicit learning involved here is mediated by dopaminergic cells in specific areas of the human brain (Schultz, 1998). How exactly changes in the firing rate of these dopaminergic cells give rise to infants' head-turning behavior we do not know. But at the functional level, the Rescorla-Wagner equations offer a simple and straightforward explanation for the observed head-turning behavior.

3 The time-course of signal-lexome decoding

Thus far, we have evaluated the performance of the Rescorla-Wagner networks by inspection of the activations of the lexomes in simple phrases. Across simulations, the networks successfully discriminated between the pertinent lexomes encoded in the signal and other lexomes by assigning the former the highest activations. In this section, we consider in more detail the timecourse of lexome activation.

For predicting the timecourse of lexome activation, we take a moving window of the incoming

t	window	cue_1	cue_2	cue_3	cue_4	cue_5	cue_6	cue_7	cue_8
1	pv	#pv	pv#						
2	pvo	#pv	pvo	vo#					
3	pvop	#pv	pvo	vop	op#				
4	pvopd	#pv	pvo	vop	opd	pd#			
5	pvopds	#pv	pvo	vop	opd	pds	ds#		
6	pvopdso	#pv	pvo	vop	opd	pds	dso	so#	
7	pvopdsob	#pv	pvo	vop	opd	pds	dso	sob	ob#
8	vopdsobg	#vo	vop	opd	pds	dso	sob	obg	bg#
9	opdsobgs	#op	opd	pds	dso	sob	obg	bgs	gs#
10	pdsobgsu	#pd	pds	dso	sob	obg	bgs	gsu	su#
11	dsobgsuk	#ds	dso	sob	obg	bgs	gsu	suk	uk#
12	sobgsukd	#so	sob	obg	bgs	gsu	suk	ukd	kd#

Table 4: Short-term moving window of width 8 for the initial part of sentence 4 of simulation 1. The # represents the absence of signal.

speech signal in discretized time and use it as the input to a pre-trained Rescorla-Wagner network. The unit of time is the segment, i.e., it takes three time steps for the moving window to slide over a triphone. Triphones become active only when they are fully supported by the information in the moving window. The moving window that we use spans 8 segments. This implementation of time involves obvious simplifications, as actual acoustic durations of triphones can vary substantially both between and within triphones. (An alternative implementation, that however requires access to the audio, is to define a window in milliseconds, to move this window across the acoustic signal by a fixed increment, and to collect those triphones that are supported by the acoustic signal in the window.)

The network serves as a memory that is itself a-temporal, but that, due to the sequential nature of the cues (n-phones), implicitly captures rich temporal information. The moving window, illustrated for the fourth simulated sentence in Table 4, represents the part of the incoming signal that can be held in a short-term memory buffer. As with other domains of temporal cognition, whether it be navigation through space, listening to music, or remembering a story or a film, complete paths of non-trivial length are impossible to hold in mind at once. Typically, we have to replay these paths step by step, where any given small segment that we can hold in mind at time t in the sequence becomes the stepping stone to the next small segment at time $t + 1$.

The moving window defines the set of n-phone cues that are available at a given point in time, henceforth the *active cues*. A cue is active when it fully matches a (three segment long) substring in the current window. The active cues are connected, with individual weights, to all lexome outcomes. The activation of a given lexome is defined by the sum of the weights on the connections from the active cues to that lexome. As the length of the window is fixed and independent of the lengths of the words in the signal, lexomes will tend to be activated when the window moves into the area where their triphone cues are located, and they will tend to de-activate when the window passes out of their cue region. Figure 5 illustrates this pattern for the fourth sentence in Table 2.

Time is displayed on the horizontal axis, with segments as units of time. Along the vertical axis, a subset of the lexomes is shown. This subset contains any lexome that, at any point in time, belongs to the six highest activated outcomes at that point in time. The activations of these lexomes are represented by discs coded with grayscales, with darker shades of gray representing higher activations. For ease of visual inspection, activations exceeding a threshold of 0.5 are presented in red. Horizontal gray lines highlight the lexomes encoded in the signal. The polygons highlight the

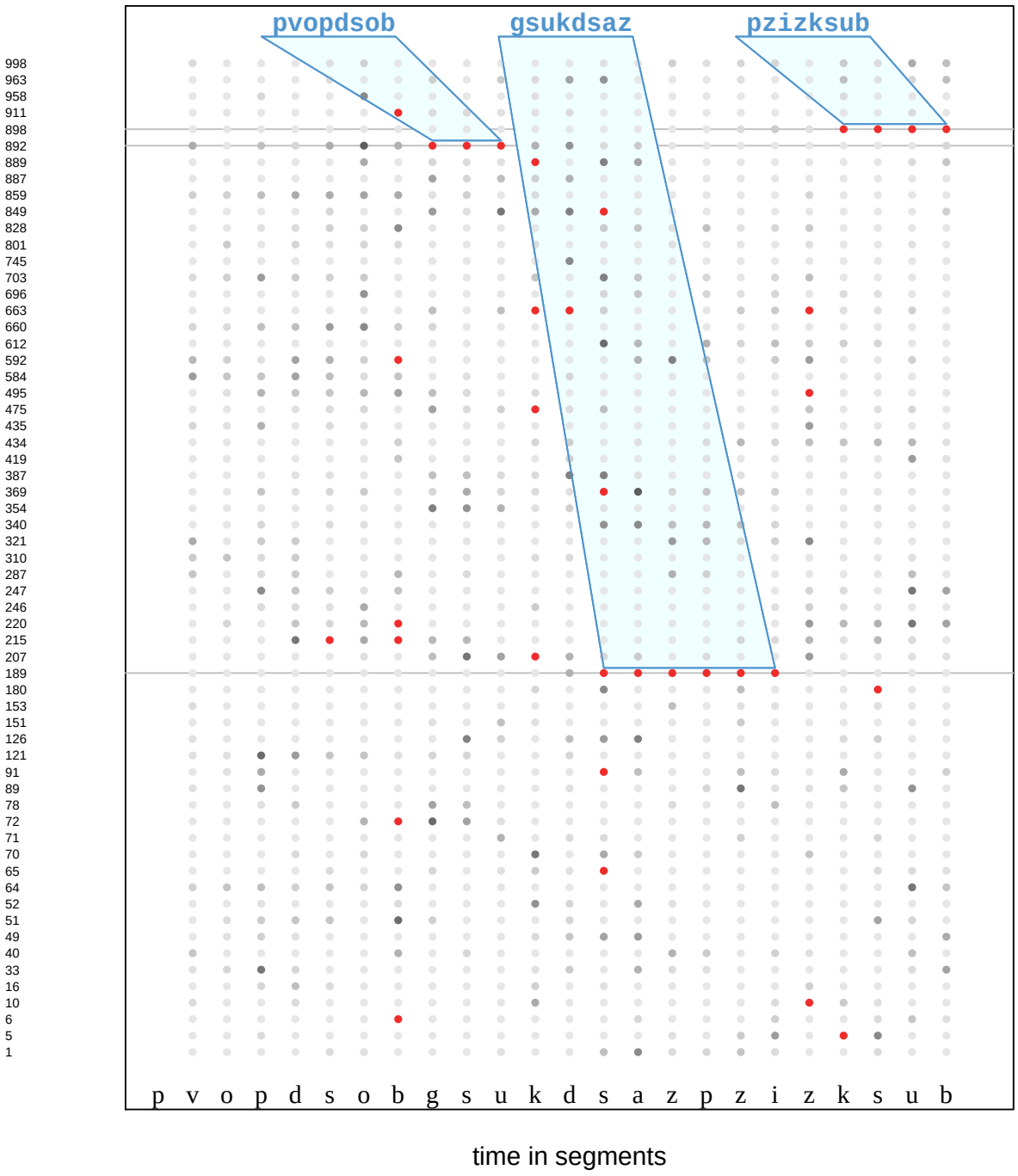

Figure 5: Activation as a function of time for the fourth simulated phrase in Table 2. Numbers in the left margin indicate lexomes. Darker shades of gray indicate greater activation of these lexomes. Activations exceeding a threshold set at 0.5 are highlighted in red. Polygons highlight time intervals where the signal provides continuous support for words' lexomes.

intervals of time at which the triphone cues in the signal activate their lexomes above the threshold.

The timecourse of lexome activation illustrated in Figure 5 is one in which the lexomes of the three words are activated in order. Word polygons have a rightwards orientation, consistent with the accumulation of evidence as the sliding window covers more of the words' cues. We note here that word forms (as displayed above the polygons) do not have any theoretical status in our model. They are shown only to facilitate interpretation.

The details of the timecourse of activation as predicted by our model can be quite subtle. Consider, for instance, the polygon for the first lexome, 892 (encoded as *pvopdsob*). The first point in time where this lexome is highly activated is when the sliding window has moved over to the first segment of the second lexome, the *g* of *gsukdsaz*. This is because the boundary triphones, *obg* and *bgs*, are the most powerful discriminators for lexome 892. What we see here is a phenomenon that we call *co-learning*. Lexomes are not learned in isolation, but in the context of utterances with other triphone cues and lexome outcomes. Lexomes are thus part of a system, a system that is much richer and informative than expected given segmentation-driven theories. Co-learning on the basis of co-occurrences of subword units like n-phones across different word forms lies at the heart of the parsimoneous explanation of frequency effects for word n-grams given by Baayen et al. (2013).

The effects of co-learning can be much more salient, as illustrated in Figure 6 for the first phrase in Table 2. Here, we see that strong support for the first two lexomes arises only when the moving window has reached the triphone cues of the third. This example illustrates a long-distance dependency (see, e.g., Schreuder, 1990, for examples from morphology), with uncertainty about the first two lexomes only being resolved once the cues of the third lexome become available. To understand why the first lexome is activated without its triphones being supported by the signal, it is crucial to keep in mind that the model is trained not on single words, but on utterances. In the course of training, the boundary triphone *tps* (and subsequent triphones in the initial stretch of *psugtGap*) happened to acquire strong weights to lexomes 100 (*pfeh*) and 837 (*dvazdGat*). Given the distributional properties of this artificial language, these triphones are much more powerful as discriminators for these lexomes than the triphones in *pfeh* and *dvazdGat* themselves.

An important aspect of discriminative decoding of the signal, illustrated in both Figure 6 and Figure 5, is that strong activations for lexome competitors are short-lived. Competitor activations above threshold are typically restricted to one time unit. The only exception in Figure 6 is for lexome 207 (encoded as *tGaptGuz*), which in this simulated language has a high frequency of occurrence and has a first syllable that is identical to the second syllable of the third lexome (encoded as *psugtGap*). But even for this strong competitor, the temporal extension of strong activation is more restricted than that of the lexomes that are actually encoded in the signal.

Results thus far are based on small samples of simple artificial languages, which raises the question of whether this approach scales up to real language. Our experience with modeling visual lexical decision latencies suggests it does. We have trained Rescorla-Wagner networks on corpora of up to 9 billion words, and obtained excellent predictions for response latencies.

In what follows, we focus on a much smaller but still non-trivial data set, consisting of the child-directed speech in the English section of the CHILDES database (MacWhinney, 2000), comprising 6,653,023 word tokens representing 34,082 word types, instantiated across 1,674,811 utterances. We ordered utterances chronologically by the age of the children addressed. A Rescorla-Wagner memory was constructed by applying the Rescorla-Wagner equations, rather than the Danks equilibria equations, with as learning events the 1,674,811 utterances, using a development version of the NDL package (Shaoul et al., 2014). Each utterance was converted into a segment stream of IPA symbols, using the CMU dictionary (Weide, 1998). For a given learning event, the cues were the set of unique triphones in the segment stream. The total number of unique triphone cues across all utterances was 23,229. In this exploratory study, the word strings were used as outcomes. For future work,

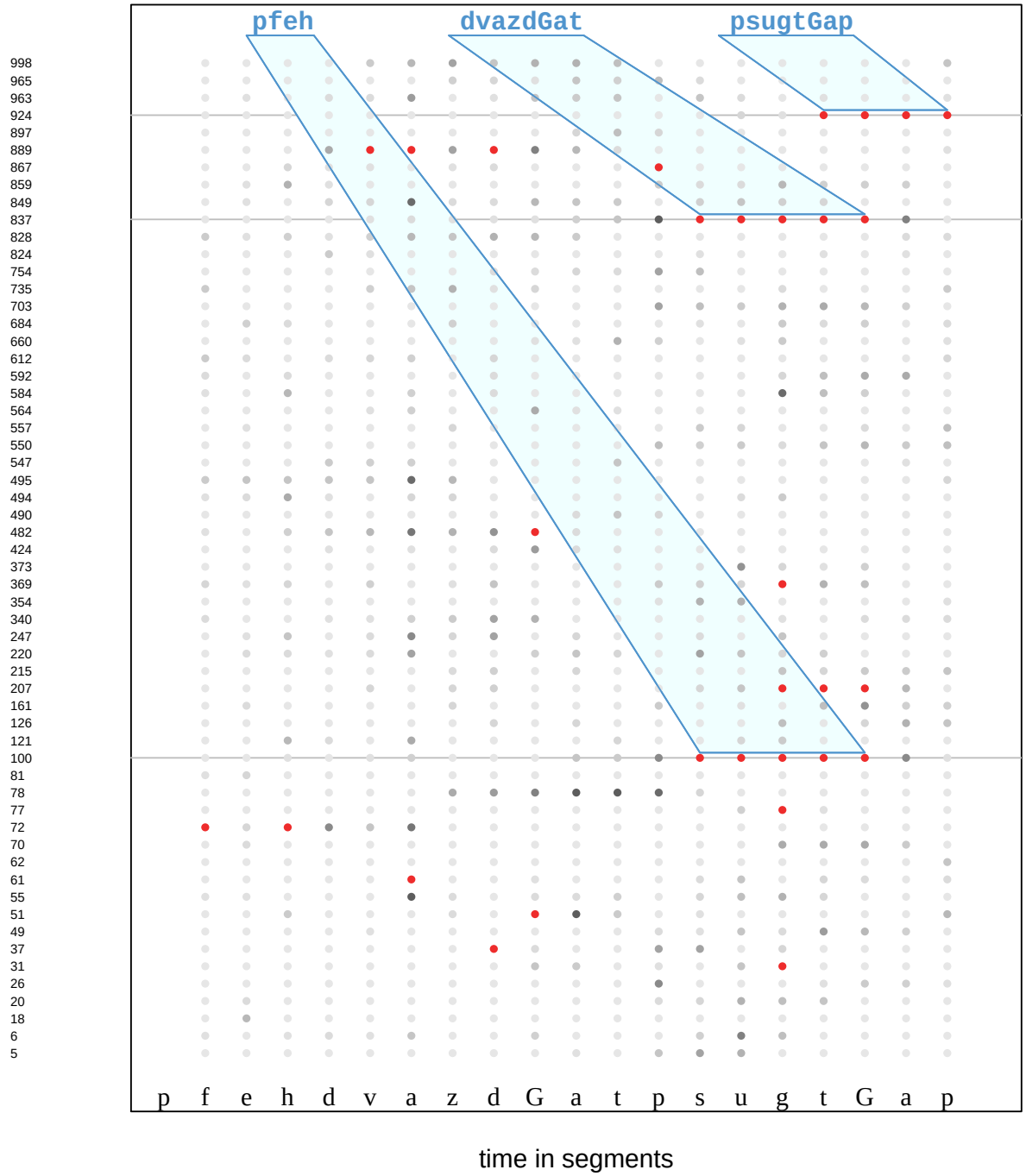

Figure 6: Activation as a function of time for the first phrase in Table 2. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.5 are highlighted in red. Polygons highlight time intervals where the signal provides continuous support for words' lexomes.

we plan to pre-process the word strings so that inflectional variants such as *play*, *plays* and *playing* will be represented by a common lexome for the experience of playing, as well as by additional grammatical lexomes for number, tense, and aspect.

Figure 7 presents the activation dynamics when a seven segment wide sliding window is moved over the sentence *can I please play with the little piggy on the chair*, a made-up sentence that does not occur in the corpus and serves as an illustration of the productivity of a Rescorla-Wagner memory. As in the preceding figures, the left axis lists any lexome that at a given point in time belonged to the set of 6 lexomes with the highest activations. Target lexomes are highlighted.

Several aspects of this example are noteworthy. First, lexomes become highly activated in roughly the same order as their triphone cues are arranged in the utterance. The only lexome that fails to be sufficiently activated to appear in the signal-to-lexome decoding time map is *play*. However, *played* and especially *playing* are highly activated. We anticipate that in a future implementation of the model in which inflected words are linked to both content and grammatical lexomes, this problem will not arise.

Second, *pig* and *piggy* are strongly activated, with strong activation for *pig* emerging one timestep earlier than for *piggy*, and with strong activation continuing one timestep longer for the diminutive. That a base and its derivative show co-temporal activation is not surprising, and both can be argued to contribute to the semantic percept of the diminutive. A more sensitive coding of the lexomes, with *piggy* sharing the category-denoting content lexome *pig* with its base, but in addition having a separate lexome for, e.g., affectiveness, will of course change the activation dynamics of the two words. A more important shortcoming of our present implementation is that acoustically, the independent word *pig* and the base *pig* in *piggy* have different acoustic characteristics (Hawkins, 2003; Salverda et al., 2003; Kemps et al., 2005a,b). As a consequence, there is discriminative information in the speech signal that is lost in our current implementation of cues operationalized as triphones.

Third, *it* is an embedded word in *little*. Even though of a very high frequency, it is not as well supported as *little*, with only two adjacent timesteps with strong activation. All other lexical competitors, such as *eat* in *the chair* (which our text-to-phone system converted to ðitʃɛr), have only short-lived high activations.

Fourth, words that appear more than once in an utterance, such as the definite article in the present example, straightforwardly activate their lexome at disjunct time intervals. Finally, the model predicts that lexomes can be strongly co-activated for overlapping time intervals. The present example illustrates this for *can I* and for *on the chair*. Since languages may express abstract features such as number, person, case, etc. by means of suprasegmentals such as stress, segmental duration, glottalization, tone, and nasalization (Hyman and Leben, 2000), we know that grammatical lexomes and content lexomes can be activated simultaneously. (The same point can be made on the basis of English irregular verbs such as *run* and *ran*, where present versus past tense is activated cotemporally with the lexical meaning.)

It is important, when building a Rescorla-Wagner memory, to use full utterances as training events, and not isolated words. This point is illustrated by Figure 8, which presents the same sentence played to a network exposed to single-word learning events. Many things now go wrong. The function words *I, the* and *on* have strong activations only at single timesteps, making it impossible to distinguish them on the basis of temporal span from spurious intruding lexomes such as *a, an, to* and *it*. Furthermore, many other words now receive extensive temporal support, such as *night, think, feeling, helping, wipe* and *each*. By withholding contextual information in the utterance, the weights from a word's cues to its lexomes are overfitted, and bereft of the moderating benefits that accrue thanks to cue competition in contextual learning.

We note here that the threshold we use here is nothing else but a tool for evaluating the perfor-

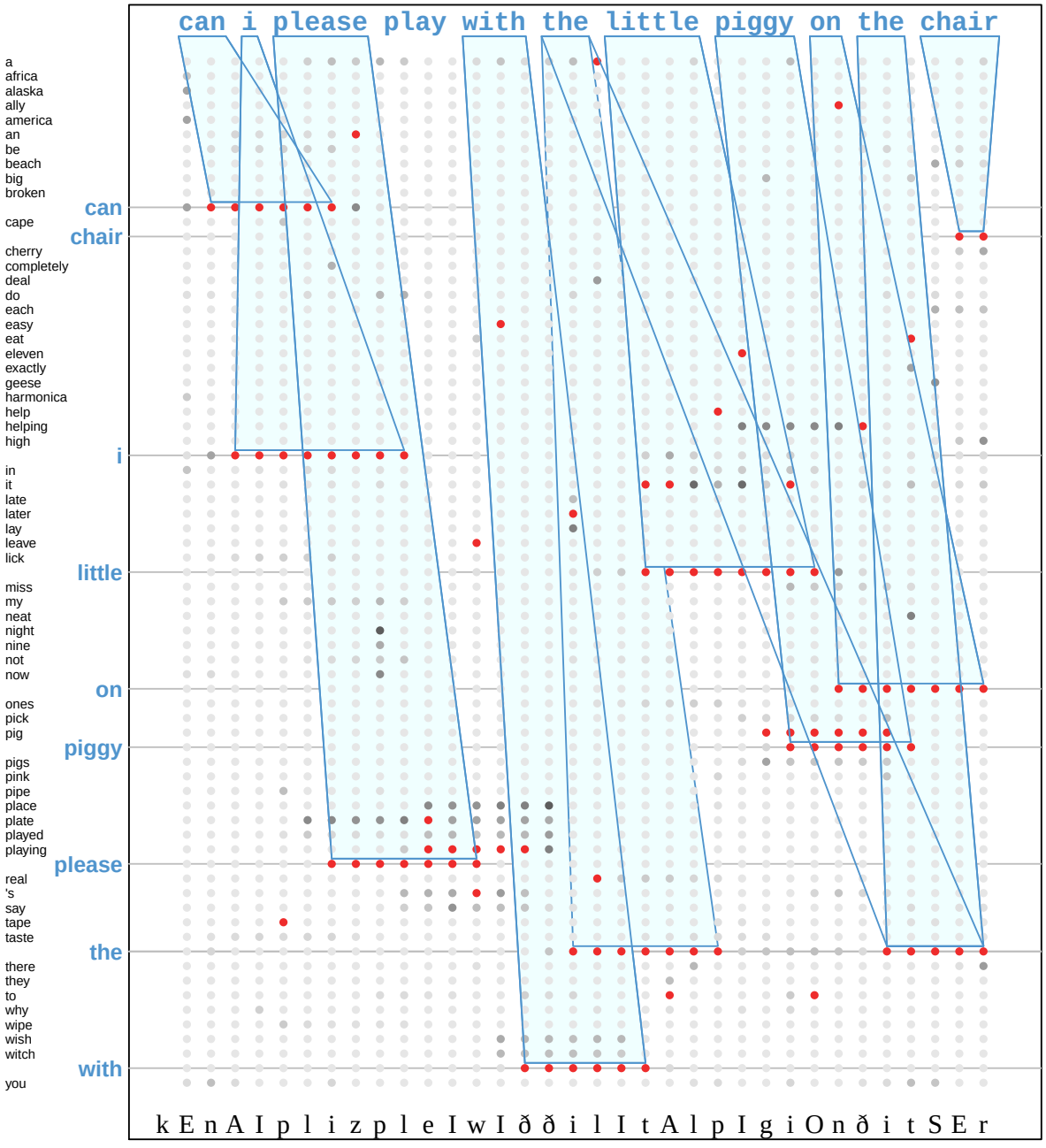

Figure 7: Activation as a function of time for a Rescorla-Wagner memory trained on full utterances in CHILDES. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.2 are highlighted in red.

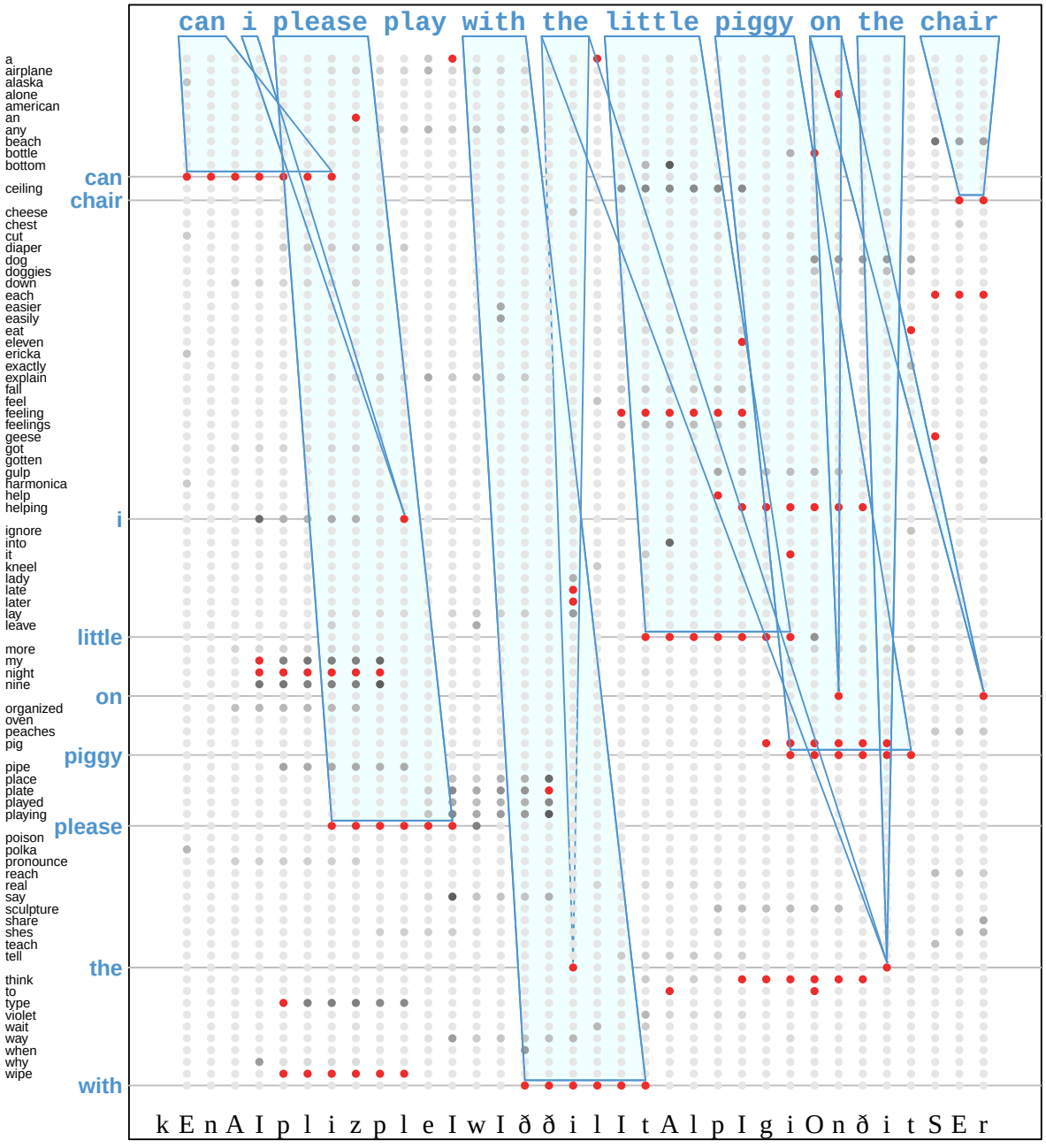

Figure 8: Activation as a function of time for a Rescorla-Wagner memory trained on isolated words in CHILDES. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.2 are highlighted in red.

mance of the network given current evaluation metrics in natural language processing. It allows us to assess the performance of the network with respect to the question whether the correct lexomes are 'identified' by inspecting extent of activation and duration of extended activation. However, the conception that underlies our model does not fit well with the idea of identifying lexomes as winners in a competitive selection process, because lexomes are seen as part of a system. It is the system of lexomes and the amount of support that all lexomes receive within that system that determines meaning, just as the potential and beauty of a chess problem is not defined by the collection of pieces on the board, or by the piece that has just been moved, but rather by the configuration of all the pieces on the board. When listening to the English word *hamster*, one other piece in the game that is especially important is the lexome HAM (Salverda et al., 2003), whereas the German word HAMSTER is positioned against the lexomes for HAMBURG and HAMMER, as the German word for *ham*, *Schinken*, is phonologically completely different. Given that availability of non-targeted lexomes is detectable in semantic categorization tasks (Bowers et al., 2005), our working hypothesis is that the meaning of a signal at time t is defined by the degrees of activations at t of all lexomes in the multidimensional space of contrasts that have been discriminated through experience with the world and language. Therefore, for comparison with other modeling technologies, the threshold serves only to guage which lexomes dominate, as well as the intervals in time when they are dominant.

Above, when evaluating the performance of our approach for artificial languages, we considered how effective a Rescorla-Wagner network is as a memory for utterances it had been exposed to. However, one desideratum for a lexical memory is that it be productive, in the sense that it works well also for unseen utterances. Shaoul et al. (2015) therefore calculated precision, recall, and the F score for a discriminative model, very similar to the one sketched above, for the 6.6 million words of the CHILDES data, using cross-validation. Precision was between 0.75 and 0.80, recall was around 0.95. F scores varied between 0.80 and 0.85. It seems reasonable to suppose that further improvements can be anticipated when lexome-to-lexome predictivity is integrated into future models.

4 Discussion

Standard approaches to language assume grammar to be a form of calculus, a formal system comprising an alphabet of elementary symbols such as stems and morphemes, stored in a mental lexicon, that is combined with a set of rules defining the well-formed symbol sequences of a language. In the context of these standard approaches, it makes sense to consider algorithms that segment the signal into its constituent symbols. Thus, models such as Shortlist-B set out to partition the speech stream into a sequence of word forms that jointly completely cover the speech stream without overlap. By combining Bayesian updating with a path-based search through a word lattice, input such as ðəkætəlɒgɪməlaɪbrɪ is segmented into the sequence of word forms *the catalogue in a library*, succesfully discarding alternative sequences such as *the cat a log in a library*.

The theory we have outlined in this study explicitly rejects the conceptualization of language as a formal calculus. Taking inspiration from Shannon's theory of information, our focus shifts from the internal constituency of the signal to the code encrypting and decrypting the experiences conveyed by the signal. We understand the encoding and decoding processes as fundamentally discriminative in nature, and have found the functional characterization of discriminative learning provided by the Rescorla-Wagner equations to provide an excellent basis for computational implementation.

By means of several simulation studies, we showed that a Rescorla-Wagner network can discriminate the lexomes encoded in a signal without segmenting the signal into word forms, outperforming

segmentation-based models. It is also capable of replicating behaviors seen in categorical perception. We also showed that a Rescorla-Wagner network, exposed to learning events which comprise all sublexical cues and all lexome outcomes present in full utterances, is an effective atemporal long-term memory system. In combination with a short-term memory buffer that projects a moving window over the incoming speech signal onto the network's input layer, the memory generates activation functions for the lexomes over time. Lexomes encoded in the speech signal are identified by monitoring for temporally extended high activation. Potential lexomic competitors give rise to little or no interference, as long as the Rescorla-Wagner memory is trained on whole utterences and not on isolated words. This architecture, which is much simpler than those of the TRACE and Shortlist-B models, offers as additional advantage, thanks to co-learning, the possibility of capturing long-distance dependencies and resolving ambiguities that in decompositional models can only be addressed post-lexically.

Experiments with young infants have been taken as evidence for segmentation of the speech stream into words. Saffran et al. (1996) concluded that apparently with only two minutes of exposure, 8-month old infants were able to find the word boundaries in an artificial language. This evidence is consistent with a Rescorla-Wagner network predicting next syllables. Furthermore, as has been pointed out by numerous people including Saffran and colleagues (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003), bootstrapping word boundaries from transition probability troughs is computationally infeasible, due to many word boundaries having high-transition probabilities. By contrast, Rescorla-Wagner networks designed to predict the lexomes encoded in the signal instead of the boundaries between word forms, do so with a very high accuracy.

Our approach differs from current approaches in machine learning (see, e.g., Brent and Cartwright, 1996; Brent, 1999; Johnson, 2008; Goldwater et al., 2009; Robinet et al., 2011; Synnaeve et al., 2014) in two important ways. First, these studies set the goal of finding the word form boundaries in utterances, whereas we aim at identifying the lexomes in these utterances. It is worth noting that by doing so, we avoid the many problems that arise in the context of "reduced" words that are ubiquitous in actual speech (Johnson, 2004). Ernestus et al. (2002) showed that many such reduced forms are not understood in isolation, and critically depend on sufficient context for comprehension to be succesful. Furthermore, Kemps et al. (2004) observed that the form listeners assume they have heard can be quite different from the form that was actually in the speech signal. Whereas these findings raise all kinds of questions for segmentation-driven approaches, they dovetail well with our approach.

Interestingly, approaches using machine learning techniques report enhanced results when predictivity from word and context is taken into consideration (e.g., Goldwater et al., 2009; Synnaeve et al., 2014). This brings us to the second way in which our approach differs. Machine learning studies on speech segmentation address the segmentation problem using unsupervised learning. If one accepts the axiom of the double articulation of language and if one believes that it is profitable to study form by itself (see, e.g., the phonological component of both traditional descriptive grammars and of generative linguistic theory), then it makes some sense to assume that in language acquisition word forms have to be acquired first and that this acquisition process must be unsupervised. Unfortunately, to our knowledge, evidence that the word forms of a language can be learned effectively (or at least, effective enough to be implementable for second language acquisition programmes) just by attending to a stream of speech has not been forthcoming. The abovementioned studies documenting the importance of lexical and contextual predictivity for word segmentation are consistent with our position that a form-only approach is too restrictive. Importantly, rejection of the axiom of the double articulation of language creates the novel opportunity to zoom in straightforwardly on how form relates to meaning, and, crucially, to switch from unsupervised learning to supervised learning. The extensive sharing of attention that is part of human enculturation (Tomasello, 2009)

in general and language learning in particular is consistent with effective learning to be contingent on meaningful communication.

From the discriminative perspective, phrasing the acquisition problem in terms of unsupervised learning makes the problem much harder than it actually is. As a consequence, strong assumptions have to be made. For instance, the model developed by Goldwater et al. (2009) requires the entire data to be available in memory for the learning algorithm to iterate over, an assumption that the authors themselves are not entirely comfortable with, and which they hope to sidestep with the help of algorithms using particle filters (Doucet et al., 2001). The MBDP-1 model of Brent (1999) and models using adaptor grammars (Johnson, 2008; Synnaeve et al., 2014) make use of advanced probability models that define which of the many possible segmentations for a given utterance is optimal. When learning is supervised, no such assumptions about a-priori available probability models or grammars are necessary. All that is required is the straightforward local updating of weights using the Rescorla-Wagner equations.

We conclude with a comparison of our discriminative learning approach with the Shortlist-B model of Norris and McQueen (2008). These authors argue that word recognition closely approximates optimal Bayesian decision making. However, in Shortlist-B, Bayesian decision making is restricted to a static probability space that does not take the sequence of learning events into account. As such, it must fail to properly predict the phenomenon of blocking (see, e.g. Kamin, 1969; Rescorla and Holland, 1982).

For example, a dog is first trained to expect food when a bell rings, and subsequently trained to expect food when a bell is rung together with a flashing light. However, the dog does not learn to expect food when the light is flashed without ringing the bell. The flashing light is blocked as a predictive cue for food. Bayesian decision making that has access only to the accumulated counts of events fails to predict the dog's expectations. For instance, consider a training sequence in which food is presented half of the time (always with a bell ringing), a light is flashed a quarter of the time together with the bell (all in the second part of the training sequence), and in which the probability of light given food is 0.5 (of all trials with food, half had a light flashing). Then, according to Bayes rule,

$$\Pr(\text{food}|\text{light}) = \frac{\Pr(\text{light}|\text{food})\Pr(\text{food})}{\Pr(\text{light})} = \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{4}} = 1.$$

For a model that does not take learning and cue competition over time into account, this is the only rational prediction, and the prediction of anyone unfamiliar with the empirical findings. Importantly, blocking is not just a curiosity from the animal learning literature: The difficulties of acquiring a second language in the presence of a first language bear eloquent witness to the pervasive effects of blocking in language processing (see Ellis, 2006a,b; Arnon and Ramscar, 2012).

The Rescorla-Wagner equations correctly predict blocking. Several proposals are available for explaining blocking using insights from Bayesian modeling (see Holyoak and Cheng, 2011, for a review). Typically, the probability of light is discounted by arguing that the light is not being attended to. These accounts are more complex than the explanation provided by the Rescorla-Wagner model, and parsimony should prefer the latter explanation. Furthermore, it has been argued that a learning rule that is 'optimal' in the Bayesian sense may be favored less by natural selection in biological systems than the Rescorla-Wagner learning rule, because the latter, as a greedy algorithm, is more robust to different configurations of parameters (Trimmer et al., 2012). Importantly, whatever the mathematical characterization of the biologically optimal way of dealing with prediction error may ultimately turn out to be, it is clear that learning and cue competition must be part of any experience-driven model of language processing.

Once this discriminative aspect of learning is taken seriously, questions must be answered about

the appropriate grain size of learning events and the particulars of the cues and outcomes in these learning events. The grain size of learning events emerged from the present study as much wider than we had originally anticipated — the model of Baayen et al. (e.g., 2011) restricted itself to learning events with only three words. The other side of the same coin is that studies of language acquisition and processing appear to have massively underestimated the importance, and indeed, the ubiquity of co-learning at every level.

Author note.

This research was supported by an Alexander von Humboldt research award to the first author. We are indebted to Petar Milin, James Blevins, and Lotte Meteyard for many discussions of the issues raised in this paper, and to Dennis Norris and two other, anonymous, reviewers for excellent critical discussion.

References

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20: 384–422.

Arnon, I. and Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122(3):292–305. doi:10.1016/j.cognition.2011.10.009.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. The MIT Press, Cambridge, Mass.

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baayen, R. H., Hendrix, P., and Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech*, 56:329–347. doi:10.1177/0023830913484896.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–482. doi:10.1037/a0023851.

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80:290–311.

Bolinger, D. (1949). The sign is not arbitrary. *Boletín del Instituto Caro y Cuervo*, 5:52–62.

Bowers, J., Davis, C., and Hanley, D. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, 52:131–143. doi:10.1016/j.jml.2004.09.003.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105. doi:10.1023/A:1007541817488.

Brent, M. R. and Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61(1):93–125. doi:10.1016/S0010-0277(96)00719-6.

Browman, C. and Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, 49:155–180.

Cairns, P., Shillcock, R., Chater, N., and Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33(2):111–153. doi:10.1006/cogp.1997.0649.

Christiansen, M. H., Allen, J., and Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2-3):221–268. doi:doi:10.1080/016909698386528.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47(2):109–121. doi:10.1016/S0022-2496(02)00016-0.

Davis, M. H., Marslen-Wilson, W. D., and Gaskell, M. G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 28:218–244. doi:10.1037//0096-1523.28.1.218.

De Saussure, F. (1966). *Course in General Linguistics*. McGraw, New York.

Dell, G. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283–321. doi:10.1037/0033-295X.93.3.283.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer, New York. doi:10.1007/978-1-4757-3437-9.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1):1–24. doi:10.1093/applin/ami038.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in l2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, 27(2):164–194. doi:10.1093/applin/aml01.

Ernestus, M. and Baayen, H. (2006). The functionality of incomplete neutralization in Dutch. The case of past-tense formation. *Laboratory Phonology*, 8:27–49.

Ernestus, M., Baayen, R. H., and Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, 81:162–173. doi:10.1006/brln.2001.2514.

Gagné, C. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27:236–254. doi:10.1037/0278-7393.27.1.236.

Gagné, C. and Shoben, E. J. (1997). The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:71–87. doi:10.1037/0278-7393.23.1.71.

Gaskell, M. and Marslen-Wilson, W. (1996). Phonological Variation and Inference in Lexical Access. *Journal of Experimental Psychology: Human Perception and Performance*, 22(1):144–158.

Gaskell, M. G. and Marslen-Wilson, W. (1997a). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12:613–656.

Gaskell, M. G. and Marslen-Wilson, W. (1997b). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*, 12:613–656. doi:10.1080/016909697386646.

Gold, K. and Scassellati, B. (2006). Audio speech segmentation without language-specific knowledge. In *Cognitive science society*, 1370–1375.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54. doi:10.1016/j.cognition.2009.03.008.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28:267–283.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31:373–405. doi:10.1016/j.wocn.2003.09.006.

Hay, J. B. (2002). From speech perception to morphology: Affix-ordering revisited. *Language*, 78:527–555.

Hay, J. B. (2003). *Causes and Consequences of Word Structure*. Routledge, New York and London.

Holyoak, K. J. and Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, 62:135–163. doi:10.1146/annurev.psych.121208.131634.

Hyman, L. M. and Leben, W. R. (2000). Suprasegmental processes. In Booij, G. E., Lehmann, C., and Mugdan, J., editors, *Morphologie: ein internationales Handbuch zur Flexion und Wortbildung. Vol. 1.*, pages 587–594. Walter de Gruyter.

Johnson, E. K. and Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567. doi:10.1006/jmla.2000.2755.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*, pages 29–54, Tokyo, Japan. The National International Institute for Japanese Language.

Johnson, M. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of ACL*, 398–406.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A. and Church, R. M., editors, *Punishment and Aversive Behavior*, pages 276–296. Appleton-Century-Crofts, New York.

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, R. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, 19:117–127. doi:10.1016/S0093-934X(03)00425-5.

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, R. (2005a). Prosodic cues for morphological complexity: The case of Dutch noun plurals. *Memory and Cognition*, 33:430–446. doi:10.3758/BF03193061.

Kemps, R., Wurm, L. H., Ernestus, M., Schreuder, R., and Baayen, R. (2005b). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, 20:43–73. doi:10.1080/01690960444000223.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–38.

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, 111(2):309. doi:10.1037/0033-295X.111.2.309.

MacArthur, R. H. (1957). On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, 43(3):293.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database.* Mahwah, NJ, Lawrence Erlbaum Associates.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Science*, 12(5):176–181. doi:10.1016/j.tics.2008.02.005.

Martinet, A. (1965). *La Linguistique Synchronique: Études et Recherches*. Presses Universitaires de France, Paris.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18:1–86. doi:10.1016/0010-0285(86)90015-0.

McQueen, J., Cutler, A., Briscoe, T., and Norris, D. (1995). Models of continuous speech recognition and the contents of the vocabulary. *Language and cognitive processes*, 10:309–331. doi:10.1080/01690969508407098.

McQueen, J. M., Cutler, A., and Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30(6):1113–1126. doi:10.1207/s15516709cog0000_79.

McQueen, J. M., Norris, D., and Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:621–638. doi:10.1037/0278-7393.20.3.621.

Milin, P., Ramscar, M., Cho, K., Baayen, R. H., and Feldman, L. B. (2015). Cornering segmentation: the perspective from discrimination learning. *Manuscript submitted for publication, University of Tübingen*.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., and Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B.*, 369(1651):20130299. doi:10.1098/rstb.2013.0299.

Norris, D. and McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–395. doi:10.1037/0033-295X.115.2.357.

Pastizzo, M. J. and Feldman, L. B. (2009). Multiple dimensions of relatedness among words: Conjoint effects of form and meaning in word recognition. *The Mental Lexicon*, 4(1):1. doi:10.1075/ml.4.1.01pas.

Pham, H. and Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition, and Neuroscience*, in press.

Port, R. F. and Leary, A. P. (2005). Against formal phonology. *Language*, 81:927–964.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M. and Baayen, R. H. (2013). Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in Language Sciences*. doi:10.3389/fpsyg.2013.00233.

Ramscar, M., Dye, M., Gustafson, J., and Klein, J. (2013a). Dual routes to cognitive flexibility: Learning and response conflict resolution in the dimensional change card sort task. *Child Development*, 84:1308–1323. doi:10.1111/cdev.12044.

Ramscar, M., Dye, M., and Klein, J. (2013b). Children value informativity over logic in word learning. *Psychological science*, 24(6):1017–1023. doi:10.1177/0956797612460691.

Ramscar, M., Dye, M., and McCauley, S. M. (2013c). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4):760–793. doi:10.1353/lan.2013.0068.

Ramscar, M., Dye, M., Popick, H. M., and O'Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PloS one*, 6(7). doi:10.1371/journal.pone.0022501.

Ramscar, M., Hendrix, P., Love, B., and Baayen, R. (2013d). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8:450–481. doi:10.1075/ml.8.3.08ram.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, R. (2014). Nonlinear dynamics of lifelong learning: the myth of cognitive decline. *Topics in Cognitive Science*, 6:5–42. doi:10.1111/tops.12078.

Ramscar, M. and Port, R. (2015). Categorization (without categories). In Dabrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 75–99. De Gruyter, Berlin.

Ramscar, M. and Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31(6):927–960. doi:10.1080/03640210701703576.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957. doi:10.1111/j.1551-6709.2009.01092.x.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. *Metaphor and Thought*, 2:164–201.

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist*, 43(3):151–160. doi:10.1037/0003-066X.43.3.151.

Rescorla, R. A. and Holland, P. C. (1982). Behavioral studies of associative learning in animals. *Annual Review of Psychology*, 33(1):265–308. doi:10.1146/annurev.ps.33.020182.001405.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, pages 64–99. Appleton Century Crofts, New York.

Robinet, V., Lemaire, B., and Gordon, M. B. (2011). Mdlchunker: A mdl-based cognitive model of inductive learning. *Cognitive science*, 35(7):1352–1389. doi:10.1111/j.1551-6709.2011.01188.x.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, 274:1926–1928. doi:10.1126/science.274.5294.1926.

Salverda, A., Dahan, D., and McQueen, J. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90:51–89. doi:10.1016/S0010-0277(03)00139-2.

Schreuder, R. (1990). Lexical processing of verbs with separable particles. *Yearbook of Morphology*, 3:65–79.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27.

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In Coltheart, M., editor, *Attention and Performance XII*, pages 245–264. Lawrence Erlbaum Associates, Hove.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shannon, C. E. (1956). The bandwagon. *IRE Transactions on Information Theory*, 2(1):3.

Shaoul, C., Arppe, A., Hendrix, P., Milin, P., and Baayen, R. H. (2013). *NDL: Naive Discriminative Learning*. R package version 0.2.14, available at `http://CRAN.R-project.org/package=ndl`.

Shaoul, C., Schilling, N., Bitschnau, S., Arppe, A., Hendrix, P., and Baayen, R. H. (2014). *NDL2: Naive Discriminative Learning*. R package version 1.901, development version available upon request.

Shaoul, C., Willits, J., Ramscar, M., and Baayen, R. H. (2015). Simulating multiple aspects of childhood language acquisition using naive discrimination learning. *under revision*.

Synnaeve, G., Dautriche, I., Börschinger, B., Johnson, M., and Dupoux, E. (2014). Unsupervised word segmentation in context. In *Proceedings of* COLING 2014*, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2326–2334, Dublin.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes*, 9:271–294. doi:10.1080/01690969408402120.

Thiessen, E. D. and Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, 39(4):706–716. doi:10.1037/0012-1649.39.4.706.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard University Press.

Trimmer, P. C., McNamara, J. M., Houston, A. I., and Marshall, J. A. R. (2012). Does natural selection favour the Rescorla-Wagner rule? *Journal of Theoretical Biology*, 302:39–52. doi:10.1016/j.jtbi.2012.02.014.

Vroomen, J. and De Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 21:98–108. doi:10.1037/0096-1523.21.1.98.

Weide, J. W. (1998). *The Carnegie Mellon Pronouncing Dictionary v. 0.6. Electronic Document.* School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76:1–15.