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

Common to all current theories of auditory comprehension is the belief that segmentation of the input into word *forms* is a prerequisite for understanding speech. We present a computational model that does not seek to learn word forms, but instead decodes the experiences communicated (represented in the model by lexemes) directly from the n-phones in the input. At the heart of this model is a Rescorla-Wagner network, trained not on isolated words, but on full utterances. This Rescorla-Wagner 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 activation functions for the lexemes over time. Lexemes encoded into the speech signal are decoded by monitoring for temporally extended high activation. Unintended lexemic competitors 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 computational proof of concept and demonstrate the importance of utterance-wide co-learning.

1 Introduction

The 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 presents a completely different computational perspective on auditory comprehension. We reject 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). Instead of marginalizing these phenomena, we take them as evidence against a two-tiered model of language.

We are therefore investigating what can be achieved with an approach in which the relation between form and meaning is the outcome 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. Baayen et al. (2011) showed for reading that a two-layer Rescorla-Wagner network correctly predicts a wide range of effects observed in experimental studies of reading. In the present study, we extend their approach to lexical access in auditory comprehension.

The algorithmic core of our model is a simple network architecture with two layers of localist representations. The input layer has units for n-phones. Although widely used, the problematic status of the phoneme as an abstract unit is well known (see, e.g., Port and Leary, 2005), and is nothing more than a back-projection of the letters familiar from western writing systems onto the speech signal. 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. Possible choices are demi-syllables, diphones, or triphones. In what follows, we make use of triphones.

The output layer contains units that we refer to as lexemes. In spite of the pervasiveness of structural metaphors 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 the words or sentences (Ramscar et al., 2010; Ramscar and Port, 2015). Instead, speech enables senders and receivers to discriminate experiences and goals on the basis of a common code. For example, in a world with just two experiences (being hungry; being satiated) and no noise, a code containing just two discrete signals, 0 and 1, would be sufficient (Ramscar and Baayen, 2013). In reality, the discreteness of elements in a system of lexical items in a code varies. However, what is important to note is that in a discriminative learning model, suppletive forms such as mice/mouse serve to accelerate the rate at which a speakers' representation of a specific form/meaning contrast becomes discriminated from form classes that express similar contrasts, such as rat/rats (Ramscar et al., 2013b). That is, experience will increasingly cause all form meaning and contrasts to become increasingly discrete within a system (Ramscar et al., 2013c), while the degree to which any given form or meaning contrast is discretized at any given point in time will depend on the status of the contrast within the overall system, and a speaker's experience of that system.

Lexemes thus serve to discretize, for the purposes of modeling, the more or less discrete symbols that conventionalize many common distinctions in the linguistic codes that have evolved amongst speaker communities. Lexemes are not form units, nor are they semantic units, but rather they represent the points of contrast that both form and meaning serve to mediate in lexical systems (see also Aronoff, 1994). In the two lexeme system we described above, a listener identifying the speech as "1" *simultaneously* resolves her uncertainty about the form *and* the meaning of a speaker's message. Accordingly, we do not assume that learners are faced with the task of associating words with concepts, but rather, we see language learning as occurring in a context in which 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 lexeme are subject to continuous change. We assume that this holds just as well for the experiences that are associated with any given lexeme (Ramscar et al., 2013a,c), 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 lexemes — and the lexemes within it — changes constantly with experience, both with respect to the objects and events in the world, as with respect to the phonetic cues, which are constantly being updated while speaking and listening.

In the network, each n-phone is connected to every lexeme. Connection strengths (weights) are estimated with the learning equations of Wagner and Rescorla (1972), or with the equilibrium equations for the Rescorla-Wagner equations of Danks (2003). Central to the learning of the weights is the concept of a *learning event*, an event in which a set of n-phones in the auditory signal co-occur with a set of lexemes. The input n-phones (the cues) predict the lexemes (the outcomes). Depending on whether these predictions are correct, the weights from the n-phone cues to the lexeme outcomes are adjusted. Thus, the computational engine of our approach is driven by prediction error. The activation of a lexeme upon presentation of a signal with a set of n-phone cues is obtained by summation of the weights on the connections from the n-phone cues in this set to that lexeme.

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). When a video camera records a boy and a girl walking, 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 boy and then the girl. 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.

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 seems so puzzling when considered as a word-at-a-time process is actually a straightforward product of this system.

Of course, training in literacy adds further layers of complexity, with knowledge of words' or-

thographic 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.

Our rejection of word form segmentation as part of auditory comprehension presents a unique perspective on the results obtained by Saffran et al. (1996) and the claim 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. In what follows, we discuss the phenomenon of low-probability phonotactic transitions (n-phone troughs), and 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 speechs 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 lexemes. 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.

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. We then explain, using discriminative learning, why infants respond to transitional troughs.

2.1 The non-optimality of segmentation

To illustrate the disadvantages of segmentation, we consider a simple artificial language. Words in this language consist of one or two syllables. Each syllable has a CCVC structure. 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 lexicon 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 1: Phrase forms, lexemes, and segmentation for simulation 1.

A total of 500 three-word phrases was generated by randomly selecting three words from the lexicon, in proportion to their frequency. Table 1 lists examples of the phrases, their constituent lexemes (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 decode the lexemes from the signal, i.e., from the unsegment phrases, without any further information such as a lexicon 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 1 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

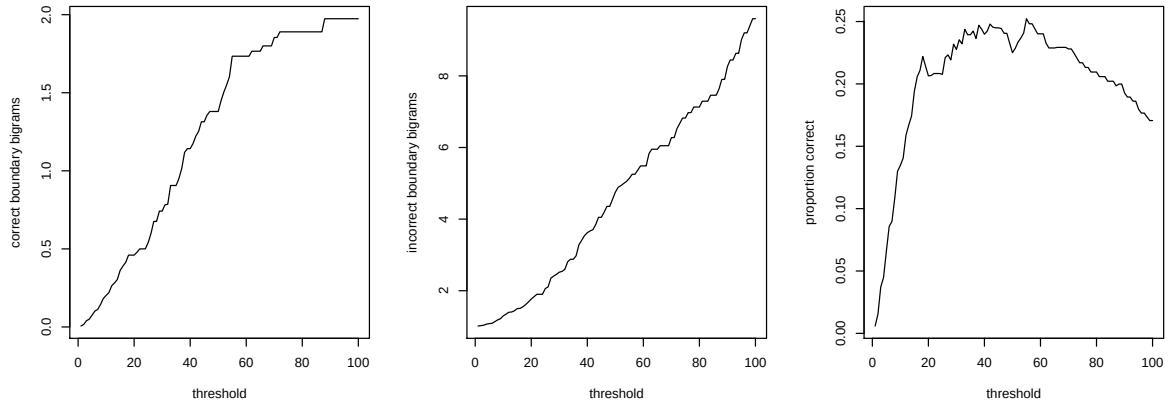

Figure 1: Accuracy of detection of word boundaries as a function of frequency threshold.

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 lexicon 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 lexemes, this high accuracy extends to the identification of the lexemes.

What can be accomplished by capitalizing low transitional diphone probabilities as segmentation cues? The crucial question here is what low is. Figure 1 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 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 the highest activations for each of the three words across all 500 phrases. Given the principles of error-driven learning, principles which have been shown to predict not only animal learning (Rescorla, 1988) but also human learning (Ramscar and Yarlett, 2007; Ramscar et al., 2010, 2011, 2013a,b, 2014), subword cues can discriminate perfectly between the lexemes that are encoded in the signal, and those that are not.

Let's now consider a simulated language with more variable phonotactics. Table 2 provides examples of phrases generated using a lexicon 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 words across all 500 phrases.

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 2: Phrase forms, lexemes, and segmentations for simulation 2.

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 lexemes encoded in the signal, and for 99.4% of the phrases, the three correct lexemes are among the top four most highly activated lexemes.

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. Furthermore, 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%.

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 decode the lexemes 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). The answer, from a discriminative perspective, is straightforward: The transitions with lower probability are predicted less well, hence greater updates of the weights are required. In other words, the infants are simply more surprised and undergo a stronger learning experience.

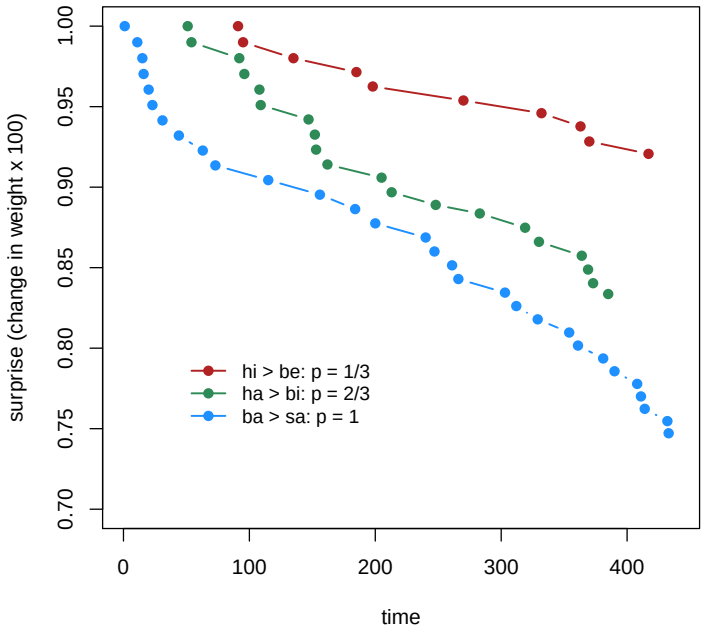

Figure 2: Surprise, measured as change in weight ($\times 100$) for a Rescorla-Wagner network ($\lambda = 1, \alpha = 0.1, \beta = 0.1$) with the current CV syllable as cue and the next CV as outcome, for three CV syllables with transitional probabilities of 1/3, 2/3, and 1, across 440 learning events.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

To clarify 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.

Figure 2 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. 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

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 3: Short-term moving window for the initial part of sentence 4 of simulation 1. The # represents the absence of signal, i.e., silence.

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-lexeme decoding

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

For predicting the timecourse of lexeme activation, we take a moving window of the incoming speech signal and use it as the input to a pre-trained Rescorla-Wagner network. 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 1, 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*. The active cues are connected, with individual weights, to all lexeme outcomes. The activation of a given lexeme is defined by the sum of the weights on the connections from the active cues to that lexeme. As the length of the window is fixed and independent of the lengths of the words in the signal, lexemes will tend to be activated when the window moves into the area where their word form is located, and they will tend to de-activate when the window passes out of their word form area. Figure 3 illustrates this pattern for the fourth sentence in Table 1.

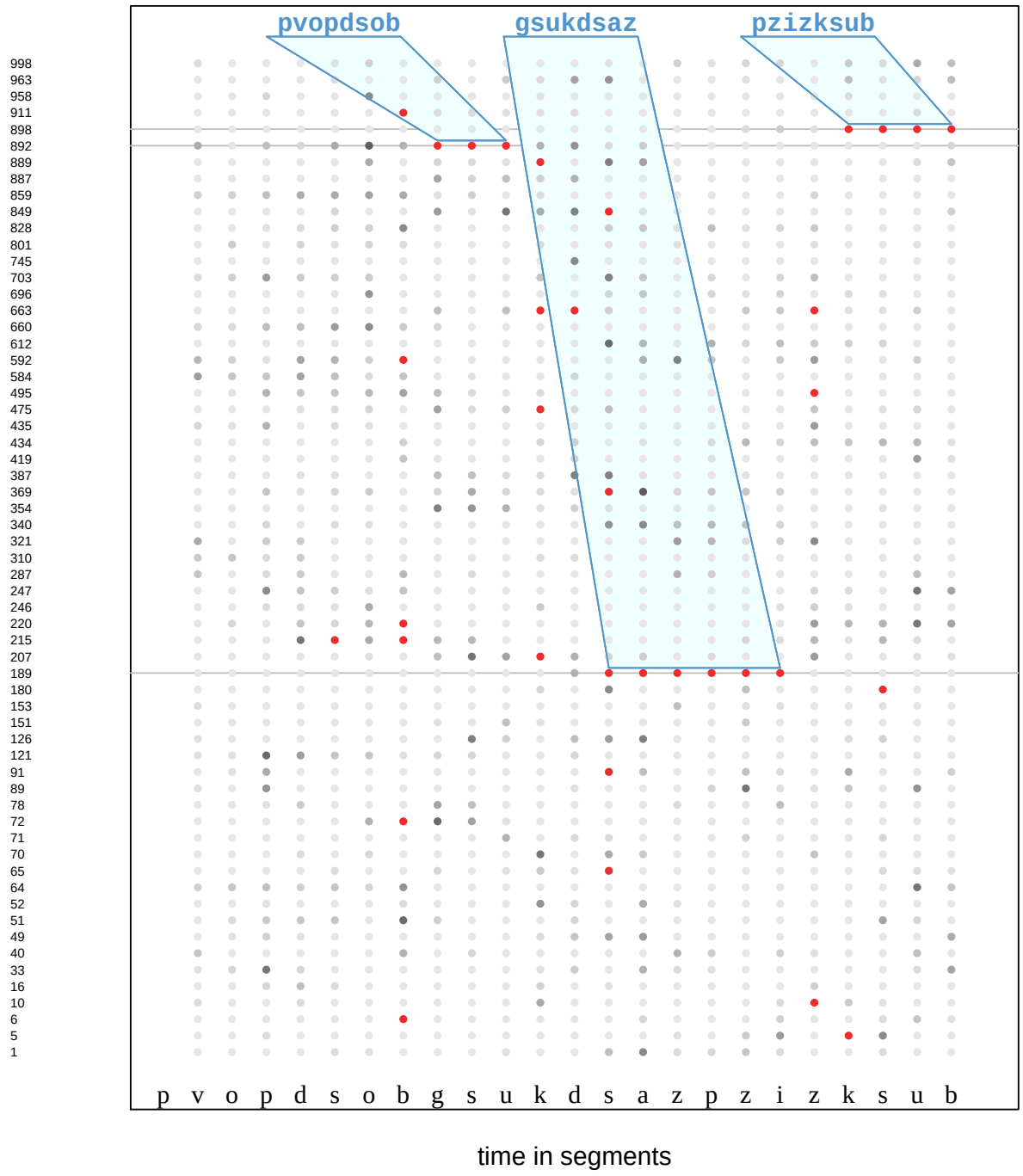

Figure 3: Activation as a function of time for the fourth simulated phrase in Table 1. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.5 are highlighted in red. Polygons highlight time intervals where the signal provides continuous support for words' lexemes.

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

The timecourse of lexeme activation illustrated in Figure 3 is one in which the lexemes 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 word form. 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 word, *pvopdsob*. The first point in time where this word is highly activated is when the sliding window has moved over to the first segment of the *second* word, the *g* of *gsukdsaz*. This is because the boundary triphones, *obg* and *bgs*, are the most powerful discriminators for lexeme 892. What we see here is a phenomenon that we call *co-learning*. Lexemes are not learned in isolation, but in context. Lexemes 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 4 for the first phrase in Table 1. Here, we see that strong support for the first two words arises only when the moving window has reached the third word. An important aspect of discriminative decoding of the signal, illustrated in both Figure 4 and Figure 3, is that strong activations for lexeme competitors are ephemeral. Competitor activations above threshold are typically restricted to one time unit. The only exception in Figure 4 is for lexeme 207, which in this simulated language is the high-frequency word *tGaptGuz*, the first syllable of which is identical to the second syllable of the third word *psugtGap*. But even for this strong competitor, the temporal extension of strong activation is more restricted than that of the lexemes that are actually encoded in the signal.

Results thus far are based on small samples of simple artificial languages. Doesn't this come with the risk of overfitting the data? We don't think so. We have trained Rescorla-Wagner networks on corpora of up to 9 billion words, and obtained excellent predictions for visual lexical decision latencies. In what follows, we focus on a much smaller 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. The words were used as outcomes. For future work, we plan to pre-process the words so that inflectional variants such as *play*, *plays* and *playing* will be represented by a common lexeme for the experience of playing, as well as by additional grammatical lexemes for number, tense, and aspect.

Figure 5 presents the activation dynamics when a seven segment wide sliding window is moved

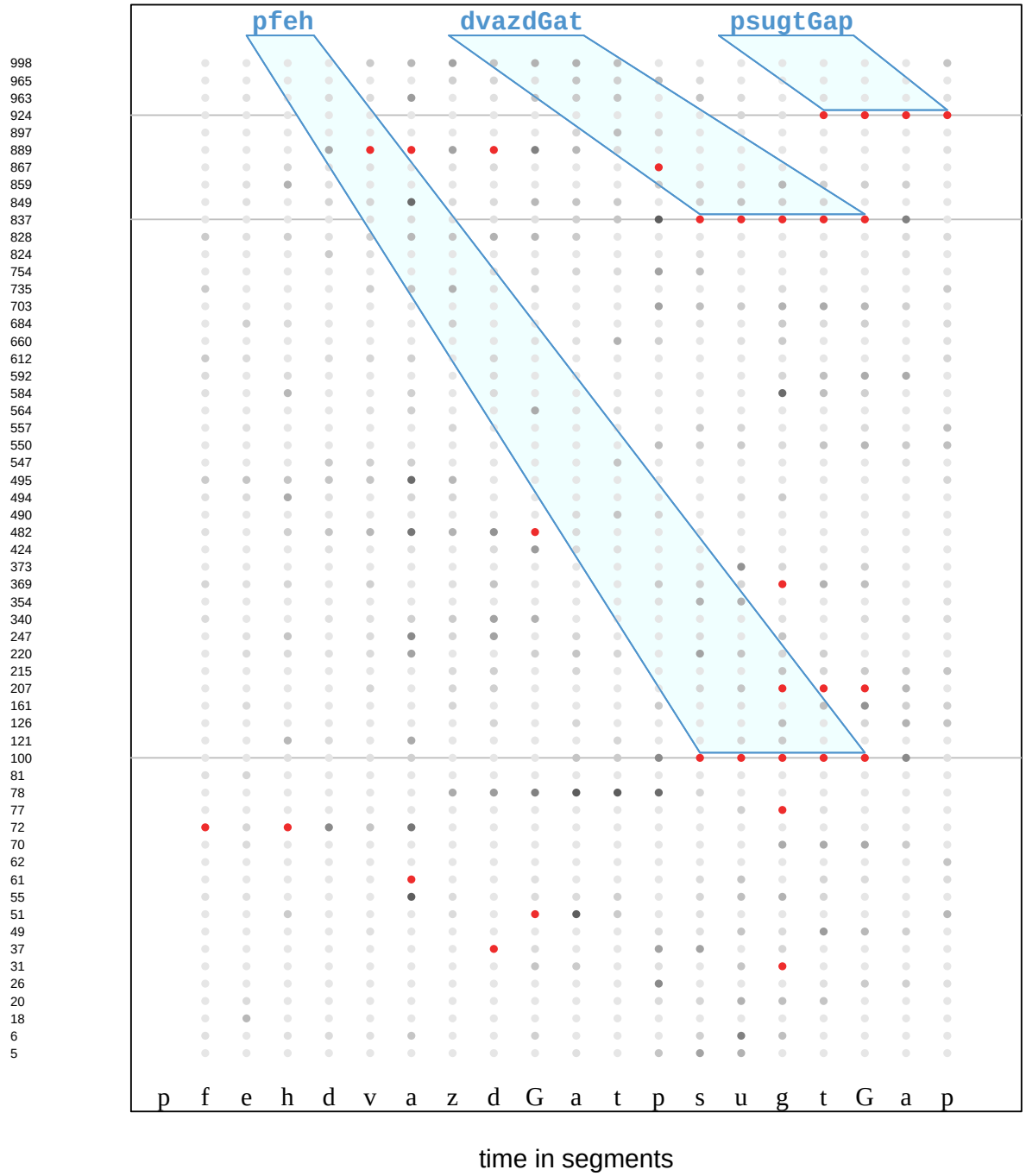

Figure 4: Activation as a function of time for the first phrase in Table 1. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.5 are highlighted in red. Polygons highlight time intervals where the signal provides continuous support for words' lexemes.

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 illustration of the productivity of a Rescorla-Wagner memory. As for the preceding figures, the left axis lists any lexeme that at a given point in time belonged to the set of 6 words with the highest activations. Target lexemes are highlighted.

Several aspects of this example are noteworthy. First, lexemes become highly activated in roughly the same order as the words are arranged in the utterance. The only lexeme that fails to be sufficiently activated to appear in the signal-to-lexeme 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 lexemes, 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 word 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 lexemes, with *piggy* sharing the category-denoting content lexeme *pig* with its base, but in addition having a separate lexeme 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 in the form of 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 ephemeral activations.

Fourth, words that appear more than once in an utterance, such as the definite article in the present example, straightforwardly activate their lexeme at disjunct time intervals. Finally, the model predicts that lexemes 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 lexemes and content lexemes 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 6, 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 lexemes 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 lexemes are overfitted, and bereft of the moderating benefits that accrue thanks to cue competition in contextual learning.

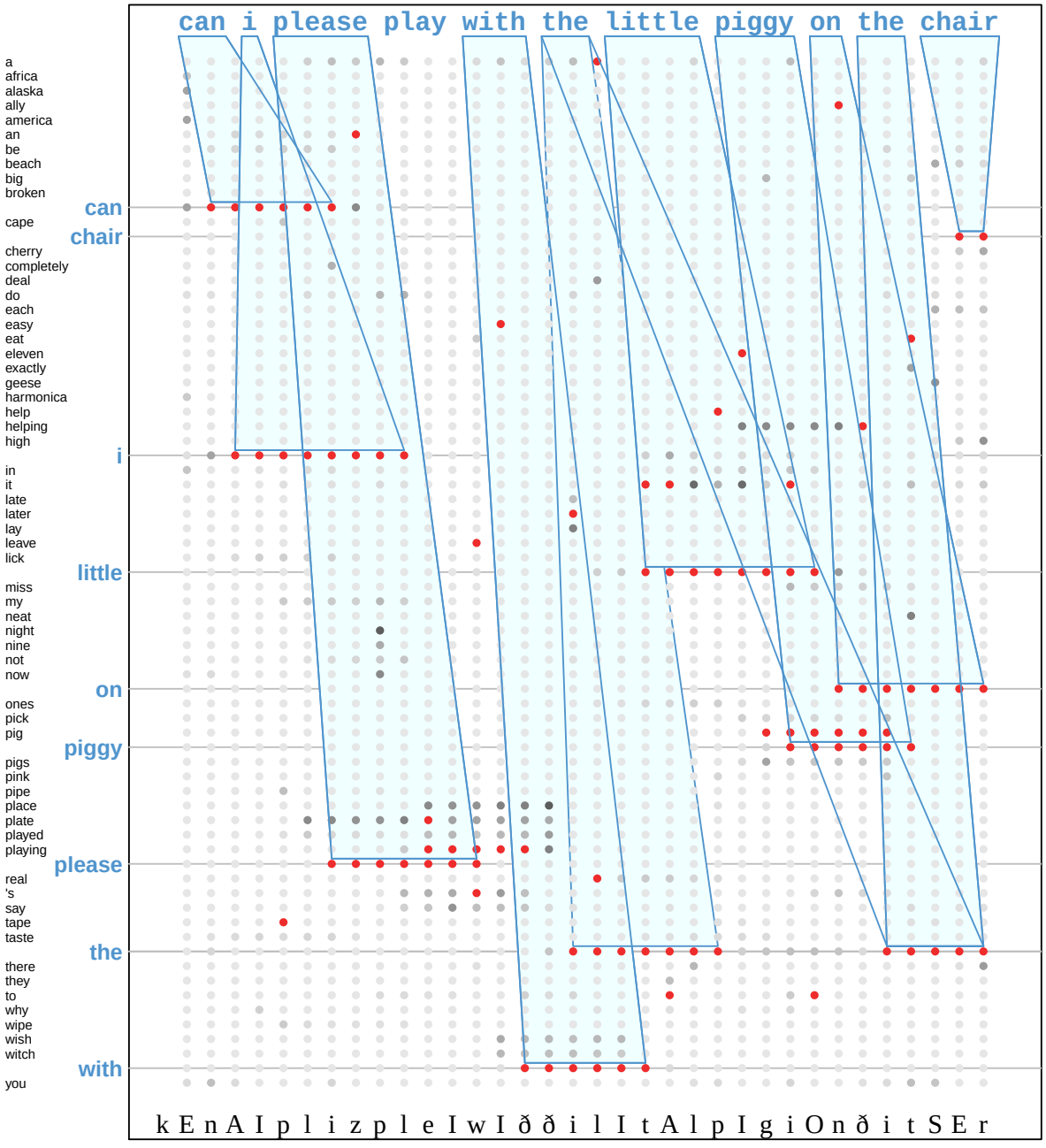

Figure 5: Activation as a function of time for a Rescorla-Wagner memory trained on full utterances in CHILDES. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.2 are highlighted in red.

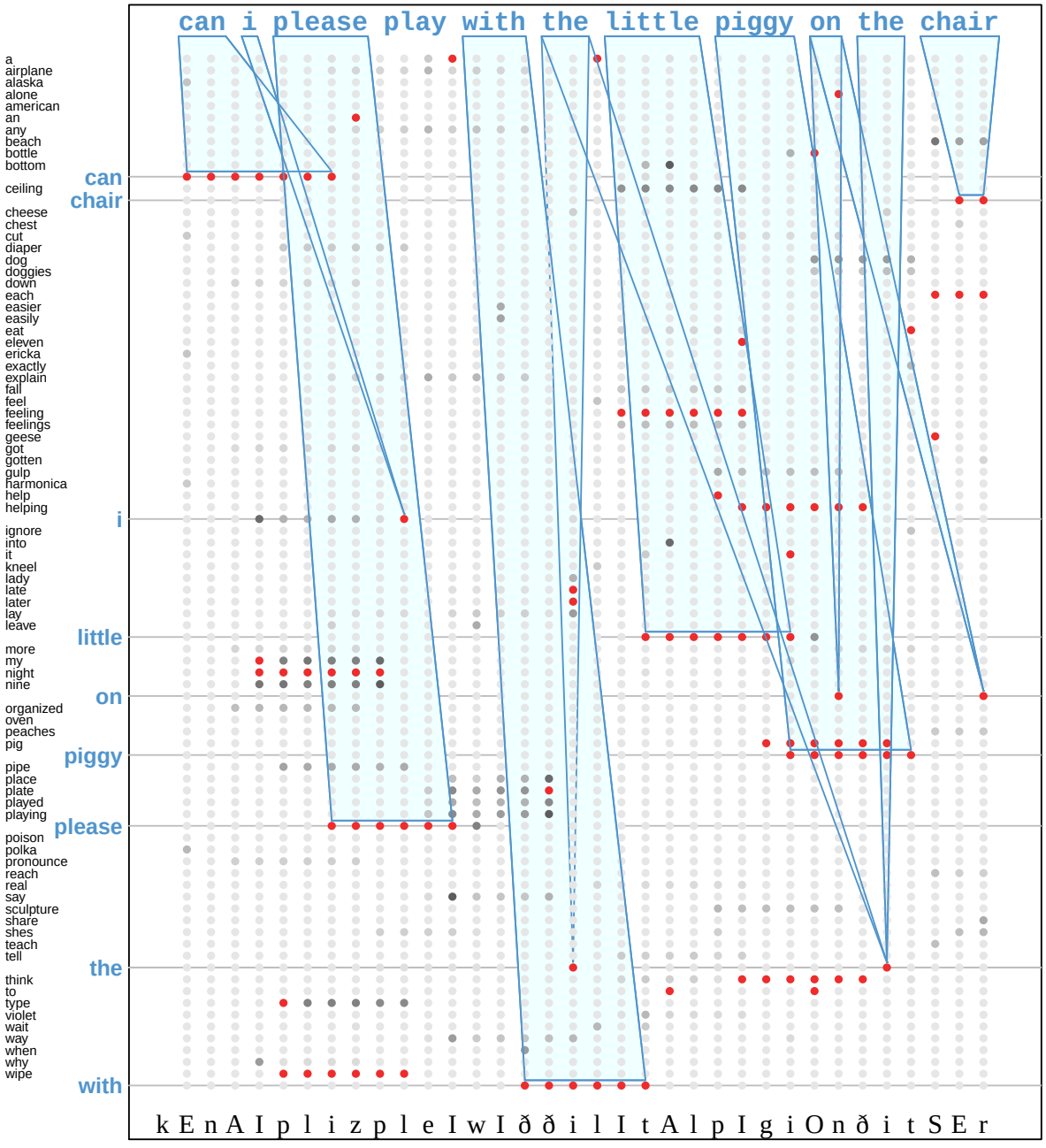

Figure 6: Activation as a function of time for a Rescorla-Wagner memory trained on isolated words in CHILDES. Darker shades of gray indicate greater activation. Activations exceeding a threshold set at 0.2 are highlighted in red.

4 Discussion

Many researchers, although working within very different theoretical frameworks, agree that prediction plays a central role in language processing (see, e.g., Hawkins and Blakeslee, 2004; Norris and McQueen, 2008; Frank and Bod, 2011). Central to the present study has been the question of prediction in lexical access during auditory comprehension. The answer that we have offered builds on a discriminative theory of language that is incompatible with standard decompositional and abstractionist theories (see also Baayen and Ramscar, 2015).

Standard approaches to language submit grammar to be a 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 suchs 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 ðəkætəlɒgɪnəlaɪbrɪ 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.

We have shown that a Rescorla-Wagner network, exposed to learning events which comprise all sublexical cues and all lexeme outcomes present in full utterances, can be used as an atemporal long-term memory system in combination with a short-term memory that projects a moving window over the incoming speech signal onto the network's input layer. In response, the memory generates activation functions for the lexemes over time. Lexemes encoded into the speech signal can be decoded by monitoring for temporally extended high activation. Unintended lexemic 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. We note here that the architecture of our model is much simpler than that of the TRACE and Shortlist-B models.

Decompositional theories, founded on the structuralist conception of the dual articulation of language, deprive themselves of the rich sublexical co-occurrence structure that is crucial for rapid and accurate discrimination between encoded and non-encoded lexemes. As a consequence, problems of discrimination and disambiguation are relegated to processing stages following the segmentation of the speech stream into word forms, to the sentence level, whereas they could already have been solved, at least in part, before then. By way of example, consider near homophones (Gahl, 2008) such as *thyme* and *time*. In our approach, these words are associated with different lexemes, and the n-phones contributed by the other words in the utterances in which these words occur contribute to discriminating between them. By contrast, the Shortlist-B model segments out the word form *tʌɪm*, and defers the disambiguation between the concrete and abstract interpretation to subsequent processes.

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. However, their evidence is entirely consistent with the predictions of a Rescorla-Wagner network predicting next syllables. Furthermore, as has been pointed out by numerous people including Saffran and col-

leagues (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 lexemes encoded in the signal instead of the boundaries between word forms, do so with a very high accuracy.

We conclude this study 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. The problem we see here is that Bayesian decision making, in the way they restrict it to a static probability space that does not take the sequence of learning events into account, must fail to properly predict the phenomenon of blocking (see, e.g. Kamin, 1969; Rescorla and Holland, 1982).

When a dog 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, this dog does not expect food when the light is flashed without ringing the bell. 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 (see Ellis, 2006a,b, for detailed discussion).

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). They all have in common that they take into account that evidence accumulates over a learning process, and that in this process, cue evidence has to be weighted for the presence of other cues. Interestingly, as pointed out by Trimmer et al. (2012), 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 is more robust to different configurations of parameters. 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. (2011) restricted itself to learning events with only three words. The other side of the same coin is that we have been massively underestimating the importance of co-learning.

Author note.

This research was supported by an Alexander von Humboldt research award to the first author.

References

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.

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.

Baayen, R. H. and Ramscar, M. (2015). Abstraction, storage and naive discriminative learning. In Dabrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 99–120. De Gruyter Mouton, Berlin.

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.

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

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.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.

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

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.

Frank, S. L. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3):474–496.

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

Hawkins, J. and Blakeslee, S. (2004). *On intelligence*. Henry Holt and Company, New York.

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

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.

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.

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. (2005a). Prosodic cues for morphological complexity: The case of Dutch noun plurals. *Memory and Cognition*, 33:430–446.

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.

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. Part*, 1.

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.

Norris, D. and McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–395.

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.

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*, page 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.

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

Ramscar, M., Dye, M., Popick, H. M., and O'Donnell-McCarthy, F. (2011). The Right Words or Les Mots Justes? Why Changing the Way We Speak to Children Can Help Them Learn Numbers Faster. *PLoS ONE*.

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

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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.

Wagner, A. and Rescorla, 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*, pages 64–99. Appleton-Century-Crofts, New York.

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.