Sidestepping the combinatorial explosion: An explanation of *n*-gram frequency effects based on naive discriminative learning

R. Harald Baayen

University of Tübingen, Germany

University of Alberta, Edmonton, Canada

Peter Hendrix

University of Tübingen, Germany

Michael Ramscar

University of Tübingen, Germany

Address all correspondence to:

R. Harald Baayen

Department of Linguistics, Eberhard Karls University

Wilhelmstrasse 19, 72074 Tübingen

E-mail: harald.baayen@gmail.com

Phone +49 (0)7071 29-73117, fax +49 (0)7071 29-5212

Sidestepping the combinatorial explosion: An explanation of *n*-gram frequency effects based on naive discriminative learning

Arnon and Snider (2010) documented frequency effects for compositional 4-grams independently of the frequencies of lower-order *n*-grams. They argue that comprehenders apparently store frequency information about multi-word units. We show that *n*-gram frequency effects can emerge in a parameter-free computational model driven by naive discriminative learning, trained on a sample of 300,000 4-word phrases from the British National Corpus. The discriminative learning model is a full decomposition model, associating orthographic input features straightforwardly with meanings. The model does not make use of separate representations for derived or inflected words, nor for compounds, nor for phrases. Nevertheless, frequency effects are correctly predicted for all these linguistic units. Naive discriminative learning provides the simplest and most economical explanation for frequency effects in language processing, obviating the need to posit counters in the head for, and the existence of, hundreds of millions of *n*-gram representations.

Keywords: naive discriminative learning; Rescorla-Wagner equations; *n*-gram frequency effects; computational modeling

Introduction

In a recent study, Arnon and Snider (2010) reported frequency effects on response latencies to four-word *n*-grams (see also, e.g., Bannard & Matthews, 2008; Tremblay & Baayen, 2010). These effects were interpreted as evidence that multi-word phrases are units of representation, just as frequency effects for regular morphologically complex words have been taken to imply the presence of representations for such words in the mental lexicon (see, e.g., Baayen, Dijkstra, & Schreuder, 1997): " At a minimum, the current findings add multi-word phrases to the units that influence processing in adults (Arnon & Snider, 2010:76).

These word *n*-gram frequency effects are of considerable theoretical interest, as they fit well with theories such as data-oriented parsing (Bod, 2006) and memory-based learning (Daelemans & Bosch, 2005). These theories assume that large numbers of instances of language use are stored, subserving on-line generalization over the exemplars stored in memory. Having access to the information carried by these stored exemplars typically boosts performance considerably. Both theories predict that multiword sequences, or (partial) parse trees for multiword sequences, are cognitively real and stored in memory, and this recent psycholinguistic evidence appears to support these predictions.

However, the idea that *n*-grams are explicitly, and thus, discretely represented in memory is problematic. First, if individuals were to explicitly represent different linguistic experiences as individual types with their own associated token frequency, then given the way language is typically experienced, the number of and size of these types would grow very rapidly, such that processing over these types would involve increasingly time-consuming searches over an ever increasing instance space. [1] It is therefore not surprising that memory-based learning as implemented in the TiMBL software (Daelemans, Zavrel, Sloot, & Bosch, 2007) offers its users compression algorithms such as information gain trees (Daelemans, Bosch, & Weijters, 1997) that speed up the search for pertitent examplars in the instance space, and provide an algorithmic window on how millions of *n*-grams might be stored without the redundancy of having to specify a given word in full for each of the *n*-grams in which it occurs.

Without the compression offered by mechanisms such as information gain trees, the numbers of instances that need to be stored remains prohibitively huge. The British National Corpus (Burnard, 1995) contains some 40,000,000 word trigrams, and the numbers of *n*-grams for higher n increases rapidly. The storage of *n*-gram representations in memory would result in enormous processing costs, and similar huge costs would be incurred in other cognitive domains, such as vision or speech (see Ernestus & Baayen, 2011, for a discussion of this issue in the context of phonetics, for which the problem of storing exemplars of all phonetic variation is exponentially more complex than for orthographic *n*-grams).

There is, of course, the possibility that the computational mechanisms used by the brain actually allow for massive storage of individual exemplars across all cognitive domains, such that a dedicated neuron or assembly of neurons fires for any given exemplar. Alternatively, if the brain were discovered to be using a form of quantum computing (see, e.g. Bruza, Kitto, Nelson, & McEvoy, 2009), the conceptual problems posed by promiscuous exemplar storage could, perhaps, be substantially reduced (cf. Skousen, 2000). Given what we presently know about the brain, the chances of it being being a quantum computing device seem very small indeed.

Additionally — and importantly — the fact that interactive activation models can employ computational shortcuts like shortlists (Norris, 1994) in order to scale up to realistically sized networks indicates that some exemplars may not be computationally relevant to processing anyway. Given that it is likely that a great many conceivable exemplars will be irrelevant for processing purposes, and given the absence of any implemented (or even implementable) algorithms to indicate how an unbounded set of n-gram exemplars might be processed in a way that is neurocomputationally plausible, the exploration of alternative computational architectures, and alternative conceptualizations of linguistic processing, seem very worthwhile pursuits.

Moreover, the idea that linguistic experience is represented at a lexical level by a (vast) series of holistically stored n-grams raises more than just processing concerns: This conception of how experience is sampled and stored is at odds with our understanding of learning. A large body of evidence, gathered from both human and animal experiments, indicates that learning is a dynamic, discriminative process that represents knowledge in an "associative" system that is constantly updated by experience (Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010). Thus while one might expect children to initially represent more holistic linguistic experiences such as word n-grams (Dabrowska, 2000; Tomasello, 2003), one would also expect that subsequent experience would weaken the associations between the components of these initial representations. This process will tend to increase the degree to which smaller elements — rather than the larger structures from which experience discriminates them — are represented in their own right, thereby decreasing the degree to which word n-grams themselves might be considered to be "stored", see Dabrowska (2000); Borensztajn, Zuidema, and Bod (2009) and Arnon and Ramscar (2012) for a computational simulation of this process that yields successful empirical predictions.

Second, the explicit representation of linguistic experience by means of holistic storage of word n-grams in lexical representations sits ill with what we understand about learning. In both humans and animals, learning is a dynamic, discriminative process in which knowledge is represented as a system in which associations between elements are constantly updated by experience (Ramscar et al., 2010). While one might expect children to initially form explicit representations of more holistic linguistic experiences such as word n-grams (Dabrowska, 2000; Tomasello, 2003), one would also expect subsequent experience to weaken associations between the components of initially holistic representations, increasing the degree to which these sub-elements are represented in their own right, and decreasing dependency on the word n-grams as such (Dabrowska, 2000; Borensztajn et al., 2009; Arnon & Ramscar, 2012).

Finally, it is unclear what the function of holistic representations for n-grams would be. Frequency effects for linguistic units are often interpreted as evidence for the existence of representations for these units. However, for a representation to be cognitively plausible, it should have more functionality than just being the receptor for a frequency 'counter in the head' (see, e.g., Morton, 1968; McClelland & Rumelhart, 1981; Norris & McQueen, 2008, for theories that take a word's unigram frequency and integrate it into a model as a resting activation level or as an a-priori probability). Interestingly, even in exemplar models, an exemplar is not a holistic, atomic, unit but typically a set of feature-value pairs over which some similarity metric is defined that informs prediction from the exemplar space. Therefore, even in exemplar theory, holistic unstructured representations for exemplars are somewhat superfluous, since the crucial information that drives similarity-based prediction is provided by the set of feature-value pairs internal to an exemplar. One perspective on exemplar theory as opposed to rule-based theories (Keuleers, 2008) is that exemplar theory stores (structured) exemplars of experience, and invokes similarity-driven rules at run-time, whereas rule-based theories such as proposed by

Albright and Hayes (2003) compile the rules during learning (with all the evidence available at the learning stage) and then apply the rules (without learning) during run-time.

The goal of the present study is to show how word n-gram effects can arise in the adult system as a consequence of *the process by which a system of mappings from orthography to decompositional semantics is learned*, such that excellent classification performance can be obtained without storing exemplars and without explicit rule induction.

The theoretical approach to learning adopted here is explicitly discriminative, implemented in the Naive Discriminative Reader model (henceforth NDR) proposed by Baayen, Milin, Filipović Durdević, Hendrix, and Marelli (2011). After describing the NDR model, we will show how an effect of word n-gram frequency on response latencies can be predicted straightforwardly, without having to posit representations for word n-grams. We will complement this evidence for our argument with a non-linguistic predictor based on the arithmetic properties of English words. This Boolean predictor, which evaluates to TRUE if the sum of the positions of a word's letters in the English alphabet is a prime number and to FALSE if it is not, is significant both for the observed and for the simulated latencies. For this predictor, an explanation in terms of dedicated representations would be absurd. We will suggest that an explanation of word n-gram frequency effects in terms of representations is barely more plausible.

Discrimination learning and the Naive Discriminative Reader

Consider what we know about children's language learning: While children are born knowing little about language, they are simultaneously sensitive to very fine grained phonetic discriminations. As their knowledge of language increases, their sensitivity to phonetic information declines (Werker & Tees, 1984), a seemingly baffling phenomenon that is incompatible with the idea that language learners utilize a fixed set of representations (in the earliest stages of language development, at least). This finding is, however, neither incommensurable with, nor even baffling, from the discriminative perspective that has come to dominate models of learning in the study of animal learning, and in human neuroscience. In a discriminative learning model, "knowledge" is initially represented by an undifferentiated complex of cue potentials that are shaped by experience as the model uses prediction-error to find the set of weights between cues and a set of outcomes (events, etc.) that best predicts the environment (see, e.g., see Elman, 1990; Kamide, Altmann, & Haywood, 2003; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Elman, 2009; Arnon & Ramscar, 2012). Because discriminative learning is at heart a process of learning to ignore uninformative cues, and because linguistic events can be both outcomes (given earlier events) and cues to later events, the process can naturally capture the way sensitivity to uninformative phonetic information declines as the number of informative phonetic discriminations increases. Indeed, it can serve to illustrate how losses in phonetic sensitivity to cues go hand in hand with increasing sensitivity to the discriminations appropriate to a given system (Ramscar, Suh, & Dye, 2011).

The same discriminative approach can be applied productively at other levels of linguistic abstraction, beyond phonetics. The computational model we use here for simulating the results of (Arnon & Snider, 2010) works in exactly the same way. Starting with inputs whose weights are not differentiated, the model learns to discriminate between better and worse cues to events as they unfold. This Naive Discriminative Reader model, which is described in detail in Baayen et al. (2011), is a simple two-layer network that takes letter unigrams and bigrams as input cues, and uses them to predict lexical and grammatical "meanings" (e.g., 'hand', 'plural'). Baayen et al. (2011) experimented with higher order letter n-grams as input cues, but found these to only lead to a marginal increase in goodness of fit.

In the NDR model, each outcome is linked to all cues, and a subnet consisting of a given outcome and the set of cues is therefore formally a perceptron. The weights for each perceptron subnet are set by means of the Rescorla-Wagner discriminative learning equations (Wagner & Rescorla, 1972). Given the presence ($\textsc{present}(X, t)$) or absence ($\textsc{absent}(X, t)$) of a cue or outcome X at time t, the Rescorla-Wagner equations provide the association strength V_i^{t+1} between outcome O and cue C_i at time $t+1$:

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

where the change in association strength is defined as:

$$\Delta V_i^t = \begin{cases} 0 & \text{if } \textsc{absent}(C_i, t) \\ \alpha_i \beta_1 \left(\lambda - \sum_{\textsc{present}(C_j,\, t)} V_j \right) & \text{if } \textsc{present}(C_j, t) \ \&\ \textsc{present}(O, t) \\ \alpha_i \beta_2 \left(0 - \sum_{\textsc{present}(C_j,\, t)} V_j \right) & \text{if } \textsc{present}(C_j, t) \ \&\ \textsc{absent}(O, t) \end{cases} \quad . \tag{2}$$

The NDR model uses the standard settings for the parameters, which are $\lambda = 1$, all α's equal, and $\beta_1 = \beta_2$. As can be seen in 2, the association of a cue with an outcome is strengthened when the cue and outcome appear together and weakened when the cue occurs while the outcome is absent.

The Rescorla-Wagner equations describe the learning process as it unfolds over time. Here, we are interested in the result of this learning process: the adult language processing system. The NDR model gauges the properties of this system through the equilibrium equations for the Rescorla-Wagner algorithm (Danks, 2003). These equilibrium equations define the connection strength (V_{ik}) of cue (C_i) to outcome (O_k) as:

$$\Pr(O_k|C_i) - \sum_{j=0}^{n} \Pr(C_j|C_i) V_{jk} = 0 \tag{3}$$

where $\Pr(C_j|C_i)$ is the conditional probability of cue C_j given cue C_i, $\Pr(O_k|C_i)$ is the conditional probability of outcome O_k given cue C_i and $n+1$ is the number of different cues. Because each perceptron subnet is modeled independently of all other perceptron subnets, Baayen et al. (2011) refer to the NDR model as an instantiation of *naive* discrimination learning.

Given a specific word or a phrase as input, only a small subset of all the unigram and bigram cues will become active. If the set of active input cues is defined as $\mathcal{C}$ and the set of active oucomes as $\mathcal{O}$, the activation of the meanings associated with the input cues is given by:

$$a_i = \sum_{k \in \mathcal{O}} \sum_{j \in \mathcal{C}} V_{jk}. \tag{4}$$

where j ranges over the active cues, k ranges over the active outcomes and V_{jk} is the equilibrium association strength for cue C_j and outcome O_k.

We note here that the accumulation of the activation a_i will almost always take place over time, as word n-grams will generally not be read with a single fixation. Instead, as the eye moves through the n-gram, the activation for the n-gram accumulates. We also note that the eye-movement record may also provide positional information that is not encoded at the level of letter unigrams and bigrams, allowing n-grams such as *take the down escalator* and *take down the escalator* to be distinguished. However, the details of how bottom-up information, including positional information, is integrated as the eye scans a sequence of words, an issue that we are currently exploring, are irrelevant for understanding how word n-gram frequency effects arise in naive discrimination learning.

The activation a_i gives the total posterior evidence for a given meaning outcome given the input cues. The more posterior evidence for this outcome, the faster a stimulus is processed. Modeled response latencies in the NDR model are therefore inversely proportional to a_i:

$$\text{simulated RT}_i \propto \log(1/a_i) \tag{5}$$

where the log-transform is applied to remove a rightward skew from the distribution of activations for ease of statistical analysis.

The NDR activations provide an estimate of how well a given meaning has been learned from the orthography. The NDR, as currently implemented, does not incorporate an algorithm for response conflict, which typically arises for many words when other meanings receive similar or even higher activations upon presentation of the visual stimulus. Interestingly, the neuroscience literature offers a range of findings in a variety of tasks in support of the idea that a circuit involving anterior cingulate (ACC) and pre-frontal cortex (PFC) serves to monitor conflicts that might arise between previously learned responses prior to final response selection (see, e.g., Montague, Hyman, & Cohen, 2004; Yeung, Nystrom, Aronson, & Cohen, 2006; Ramscar, Dye, Gustafson, & Klein, in press). In other words, response activation and response selection involve (at least) two functionally separable processes. As a consequence, decision latencies are likely to be determined by two functionally distinct factors: on the one hand, the degree to which meanings (including non-targeted meanings) are activated by the input (currently captured by the NDL), and the degree to which the co-activation of multiple candidate responses invokes subsequent cognitive control processes, which are not modelled explicitly at present.

In other work, we have addressed, at an abstract level, parts of this conflict resolution. Baayen et al. (2011) take the 20 most activated competitors into account as competitors for the target meaning, and Mulder, Dijkstra, Schreuder, and Baayen (2013) show that within the NDR approach, the summed activation of non-targeted meanings above a certain threshold can be taken into account, in line with the multiple readout framework of Grainger and Jacobs (1996). The experience of these studies is that the target's activation is the main determinant of response latencies, with additional mechanisms for conflict resolution providing minor improvements in goodness of fit. Since a great deal remains unknown about the exact functional architecture of the pre-frontal cortex at present, any model of these processes will currently have to be specified at abstract levels, especially as exact implementations of a checking mechanism could take a number of possible — and currently equally plausible — forms. In the present study, we therefore restrict ourselves to modeling the response latency to a word n-gram as the logarithm of the reciprocal of its associated activation.

The estimation of a perceptron subnet in the NDR is parameter free: It is completely determined by the distributional properties of its input space. Formally, the model is closely related to string kernels in machine learning (Buch, 2011), see also Jäkel, Schölkopf, and Wichmann (2009). The NDR can also be viewed as a statistical classifier that is optimal in the least-squares sense, and that is grounded in well-established principles of animal and human learning. Baayen (2011a), see also Janda et al. (2012), shows that naive discriminative learning, used as a classifier for data on the dative alternation, performs as well as the generalized linear mixed model, a support vector machine, memory-based learning, and random forests. This result is of particular importance as it illustrates that excellent classification accuracy can be obtained without having to posit the existence of exemplars, and without having to perform explicit rule-induction.

The simulated latencies predicted by the NDR reflect a wide variety of distributional effects, including not only whole word frequency effects, but also morphological family size effects (Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004), inflectional entropy effects (Baayen, Feldman, & Schreuder, 2006), constituent frequency effects (Baayen, Kuperman, & Bertram, 2010), and paradigmatic entropy effects (Milin, Filipović Durdević, & Moscoso del Prado Martín, 2009). This is accomplished without any representations for complex words whatsoever — the model is a full semantic decomposition model.

The weights of the NDR as reported by Baayen et al. (2011) were calculated on the basis of the co-occurrence frequencies extracted from 1,496,103 different three-word sequences extracted from the British National Corpus. Three-word sequences rather than single words were chosen in order to create a more realistic learning environment. Interestingly, it turned out that this made it possible for the NDR too also correctly model syntactic relative entropy effects present in single-word lexical decision latencies. Phrasal frequency effects were also predicted, but not tested. The present study addresses the modeling of phrasal frequency effects by taking as point of departure the materials of Experiment 1 of Arnon and Snider (2010). Does the NDR model correctly predict the observed phrasal frequency effect, even though it does not have any representations for word n-grams?

Simulating the n-gram frequency effect

Bannard and Matthews (2008) documented a whole-phrase frequency effect in a language acquisition study. They asked two and three year old children to repeat high frequency phrases such as “a drink of milk” as well as low frequency phrases, such as “a drink of tea”. Phrase pairs were matched for substring frequency. Nonetheless, performance was better and reaction times were shorter for high frequency phrases than for low frequency phrases.

Arnon and Snider (2010) extended these findings to adult language processing. In a phrasal decision task, they asked participants to judge whether or not a four-word expression was possible in English. Items were constructed in pairs, with high and low frequency members of an item pair differing only on the final word and being matched for the frequency of the final word, bigram and trigram. Any effects of phrase frequency could therefore not be due to substring frequency. They found shorter reaction times to high frequency phrases, such as “don’t have to worry” than to low frequency phrases, such as “don’t have to wait”. They interpreted these findings as evidence for full-phrase representations. Here, we present a simulation of the effect documented by Arnon and Snider (2010) with a full-decomposition model at the level of semantics, in which no full-phrase representations exist.

For each of the 47 different final words of the 56 four-word phrases of Experiment 1 of Arnon and Snider (2010), we retrieved all occurrences from the British National Corpus Burnard (1995), together with the three preceding words in the sentence, when available. From the resulting data set, those four-grams were selected that consisted only of letters, including the apostrophe. Next, all words in these phrases were lemmatized using the celex lexical database (Baayen, Piepenbrock, & Gulikers, 1995) as follows. First, inflected words were traced back to their uninflected base form. Second, if this uninflected base form was morphologically complex, the celex parse was used to retrieve its component formatives. In this way, the phrase *a British provincial city* was associated with the set of meanings {a, Britain, ish, province, ial, city}, and the *n*-gram *abnormalities take on many* with the set of meanings {ab, norm, al, ity, take, on, many}.[2] Phrases with one or more words for which a celex parse was not available were discarded. This left us with 337,069 different phrase types, representing 562,905 phrase tokens. It is noteworthy that just 47 words — the 47 different phrase-final words of Experiment 1 of Arnon and Snider (2010) — generate no less than 337,069 different phrases covering 7494 distinct meanings.

The ‘lexicon’ of 337,069 phrases was used to calculate the weights from letters and letter bigrams to the meanings associated with the constituents of the phrases. Given a phrase as input, the weights on the connections of its letters and letter bigrams to a meaning were summed to obtain that meaning's activation. A phrasal decision latency was taken to be proportional to the log of the reciprocal of this summed activation.

Note that if the model were presented with only a single phrase in training, the weights of all of the letter unigrams and bigrams present in that phrase would be equally predictive of all of the phrases' constituent meanings. The model would, in effect, have learned an explicit 4-gram representation. However, the re-occurrence of these cues and constituents in other contexts in training will generate prediction errors, causing the cues to compete for value as predictors of the constituents. This process will increasingly discriminate relationships between specific cues and constituents, such that the orthographic cue weights will be shaped more by the distributional properties of the training corpus than by exposure to individual tokens or sequences of tokens.

In the analysis of the simulated latencies, we considered phrase pair (henceforth Pair) to be a fixed-effect factor, rather than a random-effect factor, as the phrase pairs do not constitute a random sample from a larger population of such pairs. Simply taking Arnon and Snider's set of phrase-final words and repeating their matching procedure produces a very similar, if not identical set of phrases. Because this indicates that the factor levels of Pair are repeatable, Pair was entered into the model specification as a fixed-effect factor. (Although it should be noted that virtually identical results are obtained when Pair is included as a random-effect factor.)

Following Arnon and Snider (2010), we fitted a regression model to the simulated latencies with as predictors the log-transformed frequency of the four-word phrase, the log frequency of the last word, and the log frequency of the last two words. All frequencies were calculated from the lexicon used to estimate the model's connection weight matrix, and log-transformed. It is important to note that, despite the dichotomization in Arnon and Snider (2010), the phrase frequency distribution for the 56 phrases was roughly normal in this lexicon. Stepwise backward model selection using AIC resulted in a model with only the word n-gram frequency and Pair as predictors. The slope for n-gram frequency was estimated at -0.018 ($t(27)$ =-2.189, p =0.0374). The presence of a significant effect for n-gram frequency and the absence of significant effects for the frequency of the fourth word and for the frequency of the final bigram exactly mirrors the pattern of result reported by Arnon and Snider (2010) for the empirical phrase decision latencies. Importantly, the model generating the simulated latencies is parameter-free, and driven completely by its corpus input.

Model complexity

In order to evaluate the complexity of the NDR, we compare it to an interactive activation model along the lines of McClelland and Rumelhart (1981) that includes representations for word n-grams. (Such a model with word n-gram representations has, to our knowledge, not been actually implemented.) There are two important aspects of model complexity, first, the complexity of the calculations involved, and second, the number of representations and connections required.

We first consider the complexity of the calculations. In the naive discriminative learning framework, the model's predictions are based on a single forward pass of activation, as activations of meanings are obtained by summation over incoming active connections. As mentioned above, the NDL in its current implementation does not incorporate mechanisms for conflict resolution in the case that non-targeted meanings reach high activations. However, our research thus far indicates that with this single forward pass of activation, the NDL captures the bulk of the variance in the responses that can be traced to words' lexical distributional properties.

By contrast, an interactive activation model requires multiple cycles in which activation flows across inhibitory and excitatory connections at several layers. The weights on the different sets of connections are set manually, such that the targeted word (or set of words, in the case of word n-grams) receives maximal activation. Unlike the NDR, the interactive activation approach does not distinguish between the functionally separate processes of response activation and response selection. Furthermore, from the perspective of naive discrimination learning, interactive activation models re-enact, albeit imperfectly, for every occasion at which an instance of a given word is processed, the learning process of distinguishing that word from all other words. Unfortunately, the model forgets what it has learned — weights are never changed — and at the next occurrence of the same word, the whole process has to be repeated.

We maintain that the combination of multiple cycles of interactive activation and the absence of learning in interactive activation models make these models computationally more complex than the NDR model.

Next, consider the requirements of the models in terms of representations and connections. The present implementation of the NDR requires 620 orthographic input units (letters and letter bigrams), 7494 meaning units, and 620 × 7494 =4,646,280 connections from orthographic units to meaning units. The total number of units and connections is 4,654,394.

For an interactive activation model with n-gram representations, we derive the following lower-bound estimates, focusing specifically on the costs that come with adding in n-gram representations. Given our lexicon, it turns out that there are 1,628,458 distinct word n-grams ($1 \leq n \leq 4$). We assume that each of these n-grams does not spell out its component words, but provides pointers to its component words. For example, given the n-gram *of France is bald*, one could assume some orthographic access representation that we denote here simply by the string `of France is bald`, but this access representation has to provide access to the meanings of the words *of, France, is*, and *bald.* For the present data set, 4,750,180 such pointers from access representations to meanings are required. Here, we ignore the possibility that each n-gram is also linked to its subordinate and superordinate n-grams. Finally, each distinct n-gram is associated with a frequency counter (or resting activation level), adding another 1,628,458 numeric representations.

We assume that words can be represented simply in terms of letters (26), without requiring letter bigrams. We also assume that, like the NDR, the interactive activation model is decompositional, and that hence the same 7494 meanings can be used to represent the meanings associated with an n-gram. In an approach in which syntactic n-grams receive separate representations, morphologically complex words should also have their own representations. Of the 1,628,458 distinct n-grams, 21,146 are distinct words (unigrams). We assume that each unigram provides links to its constituent letter representations. For the present set of 21,146 words, it turns out that 168,686 such connections from words to letters are required.

[INSERT TABLE 1 AROUND HERE]

Table 1 summarizes the counts of model units (representations, links) in an interactive activation model and in the NDR. The total count for the interactive activation model is almost twice that for the NDR. We note here that the count for the interactive activation model is a lower bound if larger n-grams are also linked to lower-order n-grams with $1 < n < 4$.

An important difference that does not emerge from these counts is that the NDR is linear in the number of meanings: For every additional meaning, exactly 620 additional links from letter unigrams and bigrams to that meaning are required, in all, 621 model units. For the interactive activation model, each additional n-gram requires an additional representation, as well as additional links to its constituent words. Since there are hundreds of millions of n-grams, an interactive activation approach will require hundreds of millions of model units at the least, whereas 10 million such units is probably a generous upper bound for the NDR. We therefore conclude that naive discriminative learning provides the simpler, more parsimoneous, and hence superior explanation of frequency effects above the (simplex) word level. In fact, we doubt that an interactive activation model will ever scale up to dealing with hundreds of millions of n-gram representations.

The same kind of reasoning applies to the comparison of NDR and memory-based learning (Daelemans & Bosch, 2005) without representational compression through information gain trees. Consider a set of unique cues $\mathcal{C}$ and a set of unique outcomes $\mathcal{O}$, and let c and o denote their respective cardinalities. The number of representations in NDR is $c + o$ and the number of connection weights is $c \cdot o$. In memory-based learning, an exemplar is a set of cues combined with a specific outcome. Suppose an exemplar has on average $n - 1$ cues, and hence n representations, and let e denote the number of exemplars. In terms of modeling entities (representations and weights), the representational complexity of NDR will be less than that of an exemplar model when the following inequality is satisfied:

$$c + o + c \cdot o < n \cdot e. \tag{6}$$

For instance, for two outcomes, and twenty cues, naive discrimination learning requires 62 modeling entities. Suppose experience is based on 1,000 exemplars (out of the theoretically possible total of $\binom{20}{5} * 2 = 31008$ exemplars), each with 5 cues and one outcome. The number of representations required for memory-based learning will therefore be 6,000. With more exemplars, this difference will be even larger. Conditional on the two models having comparable classification performance, the model with fewer representations (and in this case, simpler calculations as well) is preferable by Occam's razor.

Distributional effects should not be reified

The previous section showed that n-gram frequency effects can be simulated in a model that has no n-gram representations. This demonstrates that the effect of a linguistic predictor on some behavioral measure of language processing does not provide compelling evidence for the existence of mental representations tied specifically to that predictor. To further illustrate this point, this section presents the effect of a fictitious non-linguistic predictor on lexical decision latencies and the simulation of this effect in the NDR model.

Consider a fictitious non-linguistic Boolean predictor *IsPrime*, that is defined as:

$$\text{IsPrime} = \sum_{i=1}^{n} \text{LetterPosition}_i \in P, \tag{7}$$

where P denotes the set of primes $(1, 2, 3, 5, 7, 11, \ldots)$, n is the number of letters in a word and *LetterPosition* is the position of a letter in the alphabet (e.g.; a = 1, b = 2, et cetera). For the example word *chair*, *IsPrime* is calculated as follows. First, we decompose the word into its component letters: *c*, *h*, *a*, *i* and *r*. Next, we transform these letters into their positions in the alphabet: 3, 8, 1, 9 and 18. Then, we sum these letter positions: 3 + 8 + 1 + 9 + 18 = 39. Finally, we evaluate whether or not this sum is a prime number. In this case it is not, so we set *IsPrime* to FALSE for the word *chair.* An example word for which *IsPrime* evaluates to TRUE is *bed.* The word *bed* consists of the letters *b*, *e*, *d*, which correspond to the alphabet positions 2, 5 and 4. The sum of these positions is 11, which is a prime number.

We calculated *IsPrime* for each of 1295 monomorphemic nouns in a dataset used by Baayen et al. (2011). For 302 of these nouns *IsPrime* evaluates to TRUE, whereas for 993 nouns it evaluates to FALSE. In addition, we extracted lexical decision latencies from the English Lexicon Project (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004) for all 1295 nouns.

The probability of a number being a prime number is somewhat higher for lower numbers. It is therefore important to note that *IsPrime* is not correlated with *Length* ($r < 0.001$). In addition, correlations with other common linguistic predictors are low: $|r| < 0.10$ for correlations of *IsPrime* with neighborhood density, written and spoken frequency, written-spoken frequency ratio, noun and verb frequency, noun-verb frequency ratio, family size, derivational and inflectional entropy, definite and indefinite prepositional relative entropy, number of simplex and complex synsets, mean and sum letter frequency and mean and sum bigram frequency.

The effect of *IsPrime* is significant ($t = 2.642$, $p = 0.008$) in a linear regression model on the lexical decision latencies from the English Lexicon Project (Balota et al., 2004), with longer reaction times when *IsPrime* evaluates to *true*. A permutation test confirmed the robustness of this effect. The probability of finding an effect of this or greater magnitude when randomly re-assigning the values of *IsPrime* to words is 0.005. The NDR correctly simulates the observed pattern of results, with a significant processing advantage when *IsPrime* evaluates to FALSE (($t = 2.945$, $p = 0.003$)).

It has been common practice in cognitive science to postulate representations and dedicated processes operating on these representations for experimentally observed effects. In line with this tradition, one might assume that words for which *IsPrime* is TRUE would be connected to a 'prime' representation that by default is set to expect words for which

IsPrime is FALSE. For this specific example, such a move would be particularly unconvincing. The effect is likely to arise from distributional properties of the language input space in English. Such an interpretation is supported by the fact that the effect of *IsPrime* did not replicate in Dutch.

From a discriminative learning perspective, linguistic predictors and their effects on behavioral measures can be thought of as windows on the distributional properties of the language system. Each predictor provides a unique and different window on the numerical patterns in the language input space. The successful replication of the effect of *IsPrime* in the NDR suggests that the model accurately captures those distributional properties of the language input space that are visible through the window of the predictor *IsPrime*.

Finally, we note that the point we want to make here is not that "correlation is not causation". From the perspective of causal modeling, a correlation, even if not causal, is indicative of a hidden causal variable. Consider a variable C that is the cause of both Y and X, and assume that there is no causal relation between X and Y. If X predicts Y, this is because Y and X have a common cause, C. In the case of *IsPrime*, however, there is no other lexical distributional variable that we are aware of that is confounded with *IsPrime*. *IsPrime* provides an (unexpected) window on the quantitative structure of the English lexicon, and it is this quantitative structure, which is captured only for a small part by *IsPrime*, that in combination with discrimination learning, enters into a causal relation with reaction times. Importantly, from the perspective of naive discrimination learning, all other standard predictors for reaction times (frequency, neighborhood density, morphological family size, orthographic consistency, etc.) are likewise imperfect windows on this quantitative structure, and are not causal factors.

Discussion

We have shown that phrasal frequency effects in the lexical or phrasal decision task can arise as a straightforward consequence of naive discriminative learning. We have also demonstrated, by means of a constructed arithmetic predictor, *IsPrime* that a predictor can be robustly significant for observed data without this significance being the consequence of mediation by particular dedicated representations. Crucially, NDR not only captures the effect of this constructed arithmetic predictor, but also points in the right direction for understanding the effect, namely, the low-level distributional properties of the orthographic-to-semantics mapping in English. We propose the same explanation for the word n-gram effect.

It is important to note that controlling for unigram, bigram, and trigram frequency when contrasting for 4-gram frequency in experiments woefully inadequate from a discriminative learning perspective. The underlying assumption of this matching procedure used by Arnon and Snider (2010) is that all that matters is relations between words and groupings of words, and that the lower-level co-occurrence patterns of sublexical units (letters, and letter bigrams) across *different* words are irrelevant. This assumption is, of course, foundational for much work inspired by formal grammars, which impose a hierarchical organization in which higher-level operations are blind to the internal constituency of lower-level units. In the learning-driven approach advocated in the present study, this assumption involves a simplification that obscures the subtle microdynamics of the mapping of form onto meaning, and that comes with the cost of having to store hundreds of millions of representations with no obvious cognitive function.

In the introduction, we suggested that an explanation of word n-gram frequency effects in terms of representations is no more plausible than an explanation of the effect of

IsPrime in terms of representations for whether the sum of alphabetic letter positions is a prime number. We are now in the position to explain this claim.

First, we have shown that the NDR model is more parsimoneous than an interactive activation model that incorporates n-gram representations. Positing n-gram representations would imply hundreds of millions of additional representations just for language. For other cognitive domains, similar combinatorial frequency effects can in all likelihood be observed as well. Since there are strong indications that the performance of naive discrimination learning is comparable to that of other classifiers, including classifiers driven by exemplars (cf. Baayen, 2011a; Janda et al., 2012), positing the existence of n-gram representations in the mind amounts to an unnecessary exponential multiplication of entities.

With our emphasis on parsimony, our approach also differs from cognitive grammar (e.g., Langacker, 1987), according to which the cognitive representational system would be massively redundant in that higher-order fully compositional representations would exist side by side with their constituents. Within the approach of naive discrimination learning, as currently implemented, it is assumed that meaning representations are restricted to those that speakers have discriminated from each other. Redundancy in the sense of having different form representations for exactly the same meaning cannot be expressed within a naive discrimination network. We note here that it is conceivable that many n-grams have their own semantic idiosyncracies, just as many derived words and compounds have meanings that are not strictly decompositional. Any n-gram with an idiosyncratic sense will require an independent meaning outcome in our model. Without sense annotations allowing us to distinguish between non-decompositional and decompositional n-grams, the modeling of the finer semantic details of word n-grams is currently not possible. Fortunately, many of the n-grams considered in the present study are fairly transparent, and for these, a decompositional representation of n-gram semantics appears to be adequate.

Second, the frequency effect is usually understood as involving some counter in the head, often implemented in the form of some resting activation level or a-priori probability. However, in discrimination learning, frequency effects are inherently contextual in nature. Baayen (2011b) shows that frequency as pure repetition, as might be modeled by a counter in the head, has very little to contribute to our understanding of lexical processing. This argument is not new, see, e.g., McDonald and Shillcock (2001), but it has all too often been ignored as an inconvenient truth.

Third, it is unclear what the functionality of n-gram representations would be. The goal of reading is understanding, and understanding involves grasping the relations between constituents. Positing a unitary phrasal representation between the constituents and the orthographic input adds another layer of complexity without adding enhanced functionality. Of course, it might be argued that an n-gram representation would constitute, or be a pointer to, a more complex semantic structure, and that the function of the n-gram representation is to save on the computations required to construct that semantic structure from scratch. From the perspective of discriminative learning — and processing — such a representation would be a crude, inaccurate approximation of the underlying pattern of associative connections that develop between forms and word meanings during learning.

Discriminative learning serves to minimize the uncertainty between a set of cues ("the learner") and a set of predicted outcomes (what is learned). Just as this approach to learning provides its own interpretation of n-gram effects, it also offers its own perspective on what the processes of language understanding involve, and what a model of linguistic understanding ought to look like. These perspectives are very different to those of traditional cognitive science. The logic of discrimination learning sits ill with the idea that linguistic signals convey meanings in words in much the same way as trains convey goods (Ramscar et al., 2010). Instead, since learning is a discriminative process, and since linguistic representations are learned discriminatively, we assume that understanding is a discriminative process as well. From this perspective, meaning is not something encoded in and decoded from language signals, but rather languages are codes in an information theoretic sense, with language signals acting to gradually reduce uncertainty about the message being transmitted, and with understanding being created by the listener/reader in the course of reconstructing the message contained in a signal (Shannon, 1948, 1956)

From a discriminative learning perspective, meaning is distributed across signals, and comprehension is a predictive process that involves the probabilistic activation of learned, experiential cues that reduce uncertainty regarding a signal and an intended message. Understanding is thus a dynamic, contextual process (Ramscar, Dye, & Klein, 2012), in which the relationship between the meaning of a message and the constituent parts of a signal depends upon the degree to which experience has discriminated them across contexts (Arnon & Ramscar, 2012), in exactly the same way that "representations" of n-grams are more or less represented as a function of experience. Cues learned in this way thus capture far more than just the associative links between elementary meanings and forms. While they can be sensitive to form patterns that may be real but arbitrary, such as `IsPrime`, a fully fledged discriminative learning model will also capture patterns that are predictive at higher levels of abstraction, such as those that are normally thought of in terms of representations of constructions, scripts, or frames.

The conceptual approach we take here is thus very different to that adopted in most traditional connectionist models of language. Influential models, such as the model for the past tense (Rumelhart & McClelland, 1986) and the triangle model (Harm & Seidenberg, 2004), are built around banks of subsymbolic units that have a function similar to that of representations in symbolic theories. For instance, the core of the past-tense model maps distributed representations of present-tense forms onto distributed representations of past-tense forms. In doing so, they adopt the hypothesis that past tense forms are indeed derived from present tense forms, as first argued by (Bloch, 1947), and subsequently by (Chomsky & Halle, 1968) and following work in the generative tradition. However, children do not learn their first (native) language by memorizing verb conjugations by rote, as in a classroom. Instead, they must learn to discriminate linguistic mappings in context. From the perspective of naive discrimination learning, past-tense forms are generated from their semantics, rather than by a phonological transformation of the present tense form (see, e.g., Ramscar & Yarlett, 2007; Tabak, Schreuder, & Baayen, 2010), such that in a naive discrimination learning approach to the production of tensed forms, the semantics of the verb as well as the semantics of tense are assumed to drive the selection of the appropriate articulatory gestures for speech production (see Hendrix, Ramscar, & Baayen, 2012, for a naive discrimination approach to word naming).

Connectionist models often seek to explain higher-order generalizations in patterns of activation over hidden units (see, e.g. Elman, 1990; McClelland & Rumelhart, 1986). The NDR network presented in the present paper, by contrast, is not intended to account for all higher-order phenomena. Although our model is highly sensitive to subtle co-occurrence patterns and their consequences for comprehension, we do not claim that the present simple network provides a comprehensive account of the full understanding of n-grams and the syntactic relations between the words in these n-grams. Complementation by additional cognitive structures is required here. Hovewer, because learning and processing are theoretically contiguous in discrimination learning, the architecture of discrimination networks can remain relatively simple when compared to connectionist models in which learning and processing are less closely aligned (Elman, 1990).

Whereas most connectionist models use back-propagation to estimate connection weights — a technique that has been criticized for being psychologically and neurobiologically implausible (see e.g., Crick, 1989; Murre, Phaf, & Wolters, 1992; OReilly, 1998, 2001) — the learning network we describe directly maps input units onto outcomes, without one or more intervening layers of hidden units. Thus as well as being more closely aligned with neurobiological approaches to learning (Schultz, 2006) than many connectionist models, our model is also much more theoretically transparent: For each outcome, we can assess straightforwardly what the individual contributions of the input cues to its activation are.

The pervasive role of prediction in comprehension (MacDonald, Pearlmutter, & Seidenberg, 1994; MacDonald & Seidenberg, 2006; Chang, Dell, & Bock, 2006; Levy, 2008; Ramscar et al., 2010; Balling & Baayen, 2012), when taken together with the degree to which the acquisition of the representations that ultimately facilitate comprehension can both be accurately modeled and predicted by discriminative learning models (Ramscar et al., 2010; Ramscar, Dye, Popick, & O'Donnell McCarthy, 2011; Arnon & Ramscar, 2012; Ramscar, Dye, & Hubner, 2012; Ramscar, Dye, & Klein, 2012; Ramscar, Dye, & McCauley, 2012), offer strong support for a theoretical approach in which all linguistic processing is conceptualized in predictive, discriminatory terms. Furthermore, human behavior in sequence learning tasks approximates more closely the performance of the Rescorla Wagner model than an SRN (Gureckis & Love, 2010). This suggests that the simpler architectures this conceptualization of language encourages offer genuine psychological and linguistic insight, in addition to parsimony.

Although the NDR does not make use of subsymbolic features (both the letter unigrams and bigrams, as well as the meanings, are represented as localist symbolic representations), we note here that the NDR shares with connectionist models the fundamental assumption that words are not atomic islands, and that subword distributional properties within n-grams co-determine learning and uncertainty reduction.

Independent empirical evidence for the importance of between-word orthographic similarities in reading — which allow the n-gram frequency to be captured by our model — is provided by the effects of relative entropy reported (and modelled) by Baayen et al. (2011) for English prepositional phrases. The basic phenomenon is that the reading of a word presented in isolation is co-determined by the distribution of contexts in which it is used. For an English word such as *table*, response latencies in visual lexical decision are co-determined by the probability distribution of the prepositional phrases in which *table* occurs (*on the table, above the table, under the table, in the table, . . .*) and the corresponding probability distribution of these phrases averaging across all nouns (*on the X, above the X, under the X, in the X, . . .*). The distance between the two probability distributions, gauged by means of the relative entropy measure, predicts response latencies to isolated nouns, such that a greater relative entropy corresponds with longer response latencies. The more atypically a noun makes use of its prepositional potential, the more difficult it is to read. The NDR captures this effect of prepositional entropy precisely because the input from which its weights are estimated consists of word n-grams rather than isolated words. (At the word level, exactly the same phenomenon is observed for Serbian case inflected nouns (see also Milin et al., 2009): The more atypical the probability distribution of a noun's case endings, the longer it takes to read that noun, irrespective of the context in which the noun is used. In this case, because the case endings are realized on the stem, rather than as separate words, the NDR can capture this effect simply on the basis of weights estimated from word unigrams.) In summary, we have independent evidence that naive discrimination learning is best trained on word n-grams. Given such a training regime, a word n-gram effect follows for free.

The approach we are pursuing is also quite different from other current non-connectionist computational models of language processing. At first sight, it would seem that our analysis of word n-grams dovetails well with the claim of Frank and Bod (2011) that the human processing system would be insensitive to hierarchical structure. After all, the NDL model estimates the activation of an n-gram as the sum of the activations of the unordered set of its meanings. However, since the eye needs several fixations for longer n-grams, partial order information can be registered and used to construct hierarchical or simply sequential representations, if so desired. More important, however, is that the grammatical formalisms take the word as an atomic unit for prediction and the calculation of measures such as surprisal, whereas in our approach, subword properties such as letters and letter bigrams across the different words in an n-gram crucially co-determine learning and prediction.

We are also well aware that the information carried by n-grams (as sets of features, or as tree fragments), are crucial for state-of-the-art theories in memory-based learning and parsing (see, e.g., Bod, 2006; Daelemans, Bosch, & Zavrel, 1999; Daelemans & Bosch, 2005). Memory-based learning has also been successfully applied to morphological processing (Krott, Baayen, & Schreuder, 2001; Krott, Schreuder, & Baayen, 2002; Ernestus & Baayen, 2003, 2007), which suggests to us that exemplar algorithms capture crucial aspects of the quantitative structure driving human processing. In fact, different quantitative approaches tend to capture the same quantitative structure with roughly the same accuracy (see Ernestus & Baayen, 2003, for an example), which suggests that the different models must be mathematically equivalent.

Interestingly, exemplar-based models are mathematically isomorphic with kernel methods (Jäkel et al., 2009), which in turn are a superclass of perceptron models. In other words, insights captured by exemplar models can be rephrased in terms of learning models, and vice versa, which in turn suggests that the presence of n-grams in exemplar-based models might better be viewed as being necessary for the implementation of such models, as opposed to their being an intrinsic theoretical necessity. A similar point is made by Keuleers (2008) with respect to memory-based learning and rule-induction as proposed by Albright and Hayes (2003). Given mathematical equivalence and equivalence in prediction accuracy, the attractiveness of a theory as a model of cognitive processing will depend on its simplicity and neurocomputational plausibility. We believe the NDR model meets these criteria better than other mathematically equivalent models.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*, 292–305.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Baayen, R. H. (2011a). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*, 295–328.

Baayen, R. H. (2011b). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*, 436–461.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, *36*, 94–117.

Baayen, R. H., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Baayen, R. H., Janda, L., Nesset, T., Dickey, S., Endresen, A., & Makarova, A. (2013). Making choices in slavic: Pros and cons of statistical methods for rival forms. *Journal of Slavic Linguistics*, in press.

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding.* (pp. 257–270). Amsterdam/Philadelphia: Benjamins.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., & 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., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (cd-rom)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Balling, L., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*, 80–106.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, *133*, 283–316.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

Bloch, B. (1947). English verb inflection. *Language*, *23*, 399–418.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*(3), 291-320.

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age — evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, *1*(1), 175–188.

Bruza, P., Kitto, K., Nelson, D., & McEvoy, C. (2009). Extracting Spooky-activation-at-a-distance from Considerations of Entanglement. *Quantum Interaction*, 71–83.

Buch, A. (2011). *Linguistic spaces: Kernel-based models of natural language (doctoral dissertation)*. Tübingen.

Burnard, L. (1995). *Users guide for the British National Corpus.* Oxford university computing service: British National Corpus consortium.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psycological Review*, *113*, 234–272.

Chomsky, N., & Halle, M. (1968). *The sound pattern of english.* New York: Harper and Row.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Dabrowska, E. (2000). From formula to schema: the acquisition of english questions. *Cognitive Linguistics*, *11*(1/2), 83–102.

Daelemans, W., & Bosch, A. Van den. (2005). *Memory-based language processing.* Cambridge: Cambridge University Press.

Daelemans, W., Bosch, A. Van den, & Weijters, A. (1997). IGTree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, *11*, 407-423.

Daelemans, W., Bosch, A. Van den, & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine learning, Special issue on natural language learning*, *34*, 11–41.

Daelemans, W., Zavrel, J., Sloot, K. Van der, & Bosch, A. Van den. (2007). *TiMBL: Tilburg Memory Based Learner Reference Guide. Version 6.1* (Technical Report No. ILK 07-07). Computational Linguistics Tilburg University.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 1–36.

Ernestus, M., & Baayen, R. (2007). Paradigmatic effects in auditory word recognition: The case of alternating voice in Dutch. *Language and Cognitive Processes*, *22*, 1–24.

Ernestus, M., & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, *79*, 5–38.

Ernestus, M., & Baayen, R. H. (2011). Corpora and exemplars in phonology. In J. A. Goldsmith, J. Riggle, & A. C. Yu (Eds.), *The handbook of phonological theory (2nd ed.)* (pp. 374–400). Oxford: Wiley-Blackwell.

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

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518-565.

Gureckis, T. M., & Love, B. C. (2010). Direct associations or internal transformations? exploring the mechanisms underlying sequential learning behavior. *Cognitive Science*, *34*, 10-50.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Hendrix, P., Ramscar, M., & Baayen, R. (2012). Ndra: a single route model of reading aloud based on discriminative learning. *Manuscript University of Tübingen.*

Jäkel, F., Schölkopf, B., & Wichmann, F. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*, 381–388.

Kamide, Y., Altmann, G., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal* *of Memory and Language*, *49*(1), 133–156.

Keuleers, E. (2008). *Memory-based learning of inflectional morphology.* Antwerp: University of Antwerp.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(1), 51–93.

Krott, A., Schreuder, R., & Baayen, R. (2002). Linking elements in dutch noun-noun compounds: constituent families as predictors for response latencies. *Brain and Language*, *81*, 708–722.

Langacker, R. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

MacDonald, M. C., Pearlmutter, N., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics, 2nd edition* (pp. 581–611). London: Elsevier.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part i. an account of the basic findings. *Psychological Review*, *88*, 375-407.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models.* Cambridge, Mass.: MIT Press.

McDonald, S., & Shillcock, R. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–323.

Milin, P., Filipović Durdević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 50–64.

Montague, P., Hyman, S., & Cohen, J. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760–767.

Morton, J. (1968). A retest of the response-bias explanation of the word frequency effect. *The British Journal of Mathematical and Statistical Psychology*, *21*, 21-33.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1271–1278.

Mulder, K., Dijkstra, T., Schreuder, R., & Baayen, R. (2013). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Under revision.*

Murre, J. M. J., Phaf, R. H., & Wolters, G. (1992). Calm: Categorizing and learning module. *Neural Networks*, *5*, 55–82.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189-234.

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

OReilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Science*, *2*, 455–462.

OReilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural Computation*, 1199–1242.

Ramscar, M., Dye, M., Gustafson, J., & Klein, J. (in press). Dual routes to cognitive flexibility: learning and response conflict resolution in the dimensional change card sort task. *Child Development*.

Ramscar, M., Dye, M., & Hubner, M. (2012). When the fly flied and when the fly flew: how semantics can make sense of inflection. *Language and Cognitive Processes*, *in press*.

Ramscar, M., Dye, M., & Klein, J. (2012). Children value informativity over logic in word learning. *Psychological Science*, *in press*.

Ramscar, M., Dye, M., & McCauley, S. (2012). Expectation and error distribution in language learning: The curious absence of "mouses" in adult speech. *Language*, *accepted*.

Ramscar, M., Dye, M., Popick, H. M., & 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). Available from `e22501.doi:10.1371/journal.pone.0022501`

Ramscar, M., Suh, E., & Dye, M. (2011). A steep price to pay? on the costs and benefits of learning relative pitch. In *Proceedings of the 33rd meeting of the cognitive science society.* Boston, MA.

Ramscar, M., & 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., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909-957.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing. explorations in the microstructure of cognition. Vol. 2: Psychological and biological models* (p. 216-271). Cambridge, Mass.: The MIT Press.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, *57*, 87-115.

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

Shannon, C. (1956). The bandwagon. *IEEE Transactions Information Theory 2*, *3-3*.

Skousen, R. (2000, August). *Analogical modeling and quantum computing.* Los Alamos National Laboratory <http://arXiv.org>.

Tabak, W., Schreuder, R., & Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon*, *5*(1), 22–46.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, Mass.: Harvard University Press.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

Van Berkum, J., Brown, C., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467.

Wagner, A., & Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64–99). New York: Appleton-Century-Crofts.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*, 49–63.

Yeung, N., Nystrom, L., Aronson, J., & Cohen, J. (2006). Between-task competition and cognitive control in task switching. *The Journal of Neuroscience*, *26*(5), 1429–1438.

Author Note

This research was funded in part by an Alexander von Humboldt research award to the first author by the Alexander von Humboldt foundation.

Notes

[1] Since stipulating 'holistic storage' for n-grams may seem somewhat crude, an at first sight less bold claim would seem more attractive, namely, that speakers would "only" store co-occurrence information. This alternative is no less problematic, however. The reason is that the frequency information stored for an n-gram must be identifiable as belonging to that n-gram. In current computational architectures, this can only be accomplished by means of an access key for the frequency information that is triggered only by the words in the n-gram, in exactly the order in which they appear in that n-gram. In other words, for the n-gram's frequency to be available at the appropriate points in time, a holistic representation for the n-gram mediating access to the frequency information is required. Hence, the alternative formulation has no advantage in terms of storage requirements.

[2] Note that plurality is not represented in the CELEX parse of the n-grams. While this coding scheme proved sufficient for the current simulations, more realistic meaning representations would include plurality information.

Table 1

*Model complexity evaluated in terms of counts of representations and connections, for an interactive activation model (*IA*) and the* NDR *model.*

	IA	NDR
n-gram representations	1628458	0
n-gram-to-word links	4750180	0
n-gram frequency counters	1628458	0
letters	26	27
letter bigrams	0	593
meanings	7494	7494
word-to-meaning links	21146	0
word-to-letter links	168686	0
orthography-to-meaning links	0	4646280
total	8204448	4654394