DRAFT, DO NOT DISTRIBUTE: Modeling Implicit Learning of Syntactic Distributions

Neal Snider
University of Rochester
Rochester, NY, USA
nsnider@bcs.rochester.edu

T. Florian Jaeger
University of Rochester
Rochester, NY, USA
fjaeger@bcs.rochester.edu

Abstract

People adapt their linguistic expectations in exquisitely sensitive ways given the properties of the environment, and there is much evidence they do this by using probabilistic inference. However, specific models of learning differ the extent to which they reflect the statistics of the environment and incorporate resource limitations such as memory. Using data from a syntactic priming experiment, which has been argued to be an epiphenomenon of implicit learning, we test two learning models: a Bayesian model that weights all evidence equally over time, and an iterative model that weights recent evidence more strongly. We compare these to a baseline proportion. The model with decay provides the best fit to the data, although the baseline proportion also explains additional variance.

1 Introduction

Language processing is affected by expectations of the upcoming signal based on experience. Sentence comprehension is affected by the frequency of words and structures, verb subcategorization biases, and verb-argument co-occurrence, among other information. Production latency and duration is affected by word frequency, the probability of a word given previous word strings, and the probability of the structure given the words in it. These findings have led to models where processing difficulty is related to how expected structures are. Connectionist models of sentence processing (Elman, 1991; MacDonald and Christiansen, 2002; Chang et al., 2006) demonstrated that models that have differential experience with linguistic material perform in a way that mirrors some of the difficulties that humans have. Multiple-constraint lexicalist models (McRae et al., 1998) and cue-based models (Bates and MacWhinney, 1989) that have weights set to match expectations from human corpus data also perform similarly to humans. What these models have in common is that they predict processing difficulty of a word or structure is inversely proportional to how expected that structure is. This is explicitly stated in recent expectation-based models (Hale, 2001; Levy, 2008), which precisely define this processing difficulty as being proportional to the information-theoretic *surprisal* (the log of the inverse probability) of the input. A key component of these models is that processing utilizes probability distributions over linguistic structures.

These probability distributions must be acquired, and there is evidence that at least some of them are acquired early in life (Snedeker and Trueswell, 2004). Infants and small children are sensitive to the frequency of words and multi-word sequences. There is also evidence that language users maintain and update their probabilistic knowledge throughout adulthood. Such adaptation seems to involve the updating of probability distributions. For example, comprehenders can not only map recently experienced exemplars of stops with different mean Voice Onset Time (VOT) on to different categories, but they also show less certainty about those categories in proportion to their variance (Clayards et al., 2008). Also listeners adapt their expectations about phonetic cues depending on the voice of the speaker (Kraljic and Samuel, 2005).

Many of the expectation-sensitive models mentioned above incorporate learning. The adaptation/learning mechanisms in these models differ in the extent to which they incorporate resource limitations. They all involve probabilistic representations, but some weight more recent evidence more strongly, which is arguably a rational strategy in some situations(Anderson and Milson, 1989). There is evidence that in object categoriza-

tion more recent exemplars are given greater consideration than distant ones (e.g. Sakamoto *et al.*, 2008). We test whether this is the strategy adopted in human language processing.

Work on adaptation of syntactic distributions during language processing (Wells et al., 2009) parallels work on skill learning and refinement, where the distributions of cues used to perform a task are learned over successive trials. Skill refinement and category learning are often studied in separate "training" and "testing" phases, but people's success on the task can also be examined throughout the experiment, as an instance of long-term priming (Mozer et al., 2007), where the amount of learning on each trial is measured as a reflection of experience gained on the previous trial or trials. Such long-term priming has also been long known to exist in sentence processing. Linguistic structures exhibit priming, in that structures are likely to be repeated in production (Bock, 1986), or are more easily comprehended (Traxler, 2008) if processed before. Priming in production is often measured using syntactic alternations: two syntactic structures whose meaning is nearly synonymous. For example, the dative alternation is exemplified as follows:

- The poor painter sold a new work to the art dealer. (PO)
- The poor painter sold the art dealer a new work. (DO)

The dative alternation involves two more or less synonymous structures (see Bresnan *et al.*, 2007 for a discussion) and a dative verb. The Prepositional Object (PO) structure has the theme argument expressed as a Noun Phrase (the sold item above, *a new work*), followed by the recipient expressed as the object of a Prepositional Phrase headed by *to* or *for* (the sellee above, *the art dealer*). The Double Object structure consists of the dative verb followed by two NPs, first referring to the recipient, and the second the theme. Priming is measured as the tendency to produce one of the two alternants (in the target trial) when it has been produced previously (in the prime trial).

Mirroring the models of long-term priming and skill refinement, one proposed mechanism for the phenomenon of structural priming is that it is an epiphenomenon of the implicit learning of probability distributions over linguistic structures (Bock and Griffin, 2000; Chang et al., 2006; Jaeger and Snider, 2008), although this is by no means universally accepted (c.f. Pickering and Branigan, 1998) A primary piece of evidence that structural priming is a byproduct of implicit learning is that it is indeed a long-lived phenomenon, persisting over several intervening experimental trials (Bock and Griffin, 2000) (although there is some evidence of possible decay, Szmrecsanyi, 2005 , especially the lexically-driven aspects, Hartsuiker *et al.*, 2008). Another piece of evidence for the implicit learning model is that structural priming is cumulative: the more prime structures processed, the greater likelihood that structure will be repeated in the target (Kaschak et al., 2006). In the implicit learning model, this could be understood as an increased probability of using a structure the more recent evidence there is for that structure. Another prediction that the implicit learning view makes about structural priming is that, like nearly all models of learning, it should be sensitive to the comprehender's expectations about the likelihood of the structure. An essential characteristic of many learning models is that the greater the difference between what is expected and what is observed, the greater the change in expectations according to the model, in other words, more learning. The idea that priming would be sensitive to expectations about the prime is consistent with several models of language processing. The language production model of Chang et al. (2006) shows a priming effect due to the backpropagation learning mechanism of its connectionist architecture: the internal representations of the model are updated in inverse proportion to the difference between expected and observed activation. Also, ACT-R (Anderson, 1983) (which has been applied to language comprehension by (Lewis and Vasishth, 2005)) contains a learning mechanism by which the base activation of a cognitive unit is incremented upon retrieval in inverse proportion to how much it was expected. There was some indication in early experiments that structural priming might be sensitive to expectation in the form of the so-called anti-frequency effect, where less frequent structures show a decreased priming effect (Bock, 1986; Hartsuiker and Westenberg, 2000). A direct test of the expectation sensitivity of structural priming was in an experiment by Snider and Jaeger (in preparation), who showed that less expected prime structures are more likely to be repeated. Further, they showed that priming de-

pended on both ***a-priori*** and **recent** expectations. The prime was more likely to be repeated when it was unexpected given experience outside the experiment (determined by an independent rating experiment). However, it was also more likely to be repeated if the structure was unexpected given experience so far in the experiment, that is, if a structure had not been produced very frequently earlier in the experiment, then it was more likely to repeated when it was encountered than one that had been observed many times. This is evidence that priming reflects a learning process where the probability or baseline activation of syntactic structures is constantly being updated.

If structural priming is an epiphenomenon of implicit learning of syntactic distributions (as suggested in Jaeger and Snider, 2008), it provides an interesting test case to examine whether this learning is subject to memory limitations, where evidence decays or is weighted less if it occurred further in the past. In this paper, we test the predictions of two different learning models for cumulative priming behavior in a language production experiment. One model assumes that behavior directly reflects the statistics of experience, with all exemplars weighted the same regardless of how far in the past they occurred. Also, it further assumes that production behavior is based on Bayesian inference, using a Beta-binomial model. The second model assumes that experience is weighted less in a power law relation to the time since it was processed. This model is based on ACT-R and incorporates an explicit trial-by-trial learning rule with a decay function. We compare both of these to a baseline model in which processing a stimulus simply reflects the proportion of that stimulus encountered. Comparing these models will give some indication as to whether adaptation in sentence processing is like Bayesian inference, and also whether it has a memory component.

2 Data

The data come from a production priming experiment provided by Snider and Jaeger (in preparation)[1]. The experiment involved priming of the dative alternation. In the experiment, the prime trials consisted of sentences that were played aloud to participants, which they were then required to repeat aloud. On target trials, a picture was presented for them to describe that was expected to elicit a dative structure. In order to investigate the effects of surprisal and cumulative learning, the experiment manipulated two other factors besides the prime structure: the *a-priori* surprisal of the prime structure given the verb (two verbs with high or low $p(PO|verb)$), and recent experience with the prime structure via a block design where primes were presented either in two blocks (all PO followed by all DO, and vice versa, see Figure 1) or alternating trial by trial. $p(PO|verb)$ was estimated by a separate norming experiment. The experiment had 45 participants and 22 items, with a total of 755 tokens in the analysis after exclusions due to (for example) non-production of a dative in the prime or target.

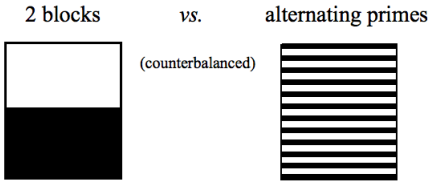

Figure 1: Recent experience was manipulated in the experiment by a blocked condition and an alternating condition

Snider and Jaeger found that surprisal interacted with priming such that less expected (more surprising) prime structures were more likely to be repeated. They argued that this effect is a sign that the phenomenon of priming is a process of implicit learning because many learning models are expectation-sensitive in this manner (although as we argue later, this prediction also holds for a Bayesian learning model without an explicit learning rule). Another key finding was that the surprisal sensitivity that they observed in the experiment occurred for both recent as well as *a-priori* experience: prime structures that were less expected given experience in the experiment so far were more likely to be repeated, just like prime structures that were less expected given the participants' experience prior to the experiment. These effects occurred in addition to general cumulative priming, where more prime structures produced previously in the experiment make that structure more likely to be produced in any given target.

In the next sections, we present two studies involving two learning models, which we apply in order to predict the cumulative priming be-

[1] We thank the authors for providing the data.

havior in this experiment. One involves iterative Bayesian learning using a Beta-binomial model, and the other uses an iterative learning rule inspired by the ACT-R learning rule.

The data were analyzed with mixed-effects logistic regression, using the same controls as in the original analysis: the main effect of prime structure, recent and a-priori surprisal of the prime, and random control factors of participant, item, and target verb. Evaluation is done by model comparison: a factor is entered into the logistic regression, and then removed to determine if it significantly reduces the likelihood of the data according to the statistical model.

3 Study 1: Beta-binomial learning model

In the original experiment, recent experience was modeled just by using the proportion of the PO alternation that had been observed up until the current trial. However, use of proportions over so few trials risks over-fitting the data (Bishop, 2006). A better approach is to use an iterative Bayesian model, where the prior and likelihood are updated on each trial as more evidence comes in. In the experiment, a choice between two categorical outcomes was assumed, so a *Beta-binomial* model was used. In a Beta-binomial model, the prior is the Beta function:

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \quad (1)$$

where μ is the probability of the outcome (PO), and the Beta function has two parameters, a and b, which are analogous to the numbers of each outcome. In this case a corresponds to the PO alternant and b corresponds to the DO alternant. Once some data has been observed, the likelihood function is just the binomial distribution:

$$\text{Bin}(\mu|m,l) = \binom{m+l}{m}\mu^m(1-\mu)^l \quad (2)$$

Where m corresponds here to the positive response, PO, and l to the DO response. In the iterative Beta-binomial model, the prior on each successive trial is just the posterior on the previous trial, following Bayes rule. Fortunately, an analytic solution exists, found by multiplying the prior by the posterior and normalizing (Bishop, 2006). So the probability distribution over probability parameters μ is a function of the number of POs observed (m), the number of DOs observed (l), and the prior parameters a and b:

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \quad (3)$$

At each trial i, the optimal value of μ is determined by finding $\text{argmax}_{\mu_i} p(\mu|m_i,l_i,a,b)$. We determine the prior parameters a and b by determining the optimal values that fit the data set using Quasi-Newton method, verified by brute force search.

The Beta-binomial model implicitly predicts a positive correlation between surprisal and priming, just like mechanistic models with an explicit expectation-sensitive update rule. This is due to the fact that, when one of the experience dimensions ($m+a$, $l+b$) is high and the other low, the function defined by $\text{argmax}_{\mu_i} p(\mu|m_i,l_i,a,b)$ has a higher slope in the dimension of lower occurrences than in the dimension of higher occurrences. For example, suppose at one point in the experiment, the PO alternant is highly expected, having occurred 11 times while the DO only occurred once, then $\text{argmax}_{\mu_i} p(\mu|m_i = 11, l_i = 1, a = 1, b = 1) = 0.916$. Observing one more of the (less expected) DO changes the argmax probability to 0.846, a difference of 0.07, while observing one more PO after having observed 11, changes the the argmax probability to 0.923, for a difference of only 0.007. Therefore, observing a less expected prime structure increases that structure's probability more than observing an expected prime structure.

3.1 Results

In order to determine the prediction of the Beta-binomial model for cumulative priming, one must determine the parameters of the prior distribution, a and b. We set these parameters by optimizing the mixed effects regression model of the data with respect to a Beta-binomial model of cumulative priming with these parameters, including the controls from the original experiment. Optimization was done by Quasi-Newton method (using the *optim* function in *R*, with subsequent brute-force verification), and yields optimal parameters of $a = 1$ and $b = 1$, which corresponds to a uniform prior where $p(PO) = 0.5$ (interestingly, a prior that reflects the proportion of PO and DO found in spoken corpora by Bresnan *et al.*, 0.2 and 0.8 respectively, yields a much worse fit).

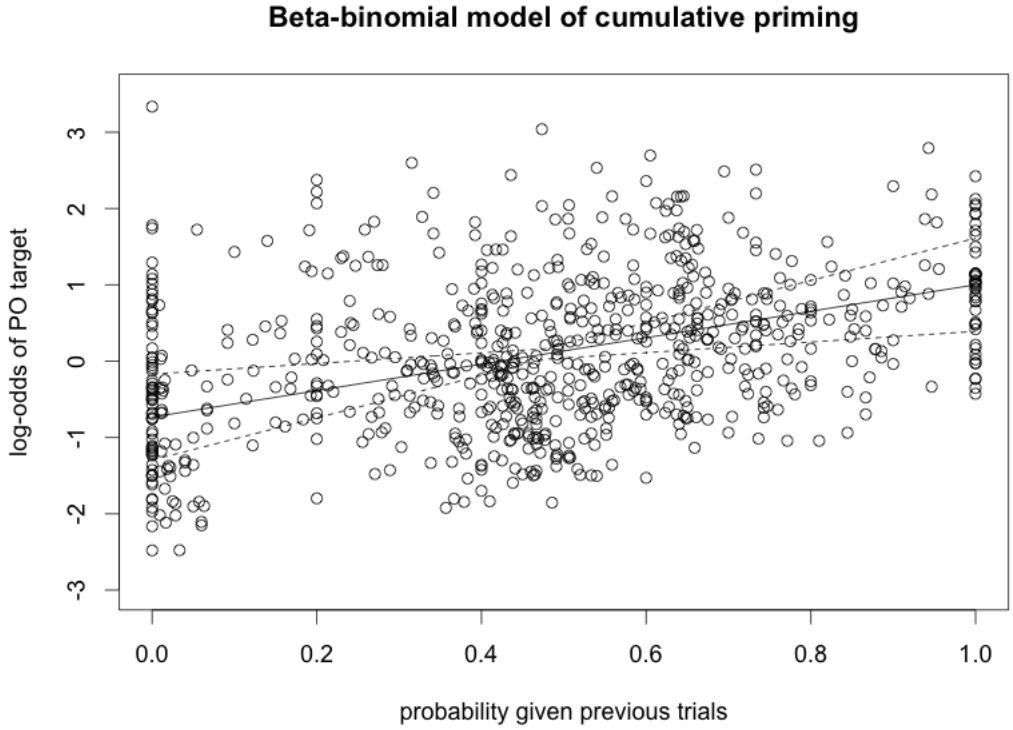

Figure 2: Beta-binomial model

The Beta-binomial model does significantly predict cumulative priming behavior: the more probable a PO given experience so far in the experiment, according to equation (3), the more likely a PO is to be produced on the target trial (χ^2 = 5, p = .025). The prediction of the model for both the Beta-binomial is compared with the data from the experiment in Figures (2). The success of the Beta-binomial model indicates that human learning is to at least some extent weights exemplars of experience equally, in a manner consistent with Bayesian inference. In the next section, we test whether the model with an iterative learning rule and weighting of experience provides a better fit to the data.

4 Study 2: Weighted learning model

The learning model we test here uses a metaphor of activation of representational nodes, where activation increases with repeated exposure to a stimulus, but activation from units encountered further in the past is given lower weight (or decays). The decay follows a power law function, similar to the ACT-R Base-Level Learning Equation (Anderson and Lebiere, 1998). We also allow for the activation of a unit to be explicitly proportional to its surprisal, with more surprising structures (given recent exposure) to be given more activation. Surprisal based on recent experience is defined as follows:

$$Surp(PO)_i = -\log p(PO|\text{items } < i) \quad (4)$$

In order to weight the surprisal-sensitive activation, we allow activation to be a linear interpolation of a baseline level from having encountered the structure and the surprisal from having encountered that structure. The baseline activation from processing a prime is:

$$BaseAct(PO)_i = \begin{cases} 1 & \text{if } i = PO \\ -1 & \text{if } i = DO \end{cases} \quad (5)$$

The surprisal-sensitive component of activation is:

$$SurpAct(PO)_i = \quad (6)$$

$$\begin{cases} Surp(PO)_i & \text{if } i = PO \\ -1 \cdot Surp(DO)_i & \text{if } i = DO \end{cases}$$

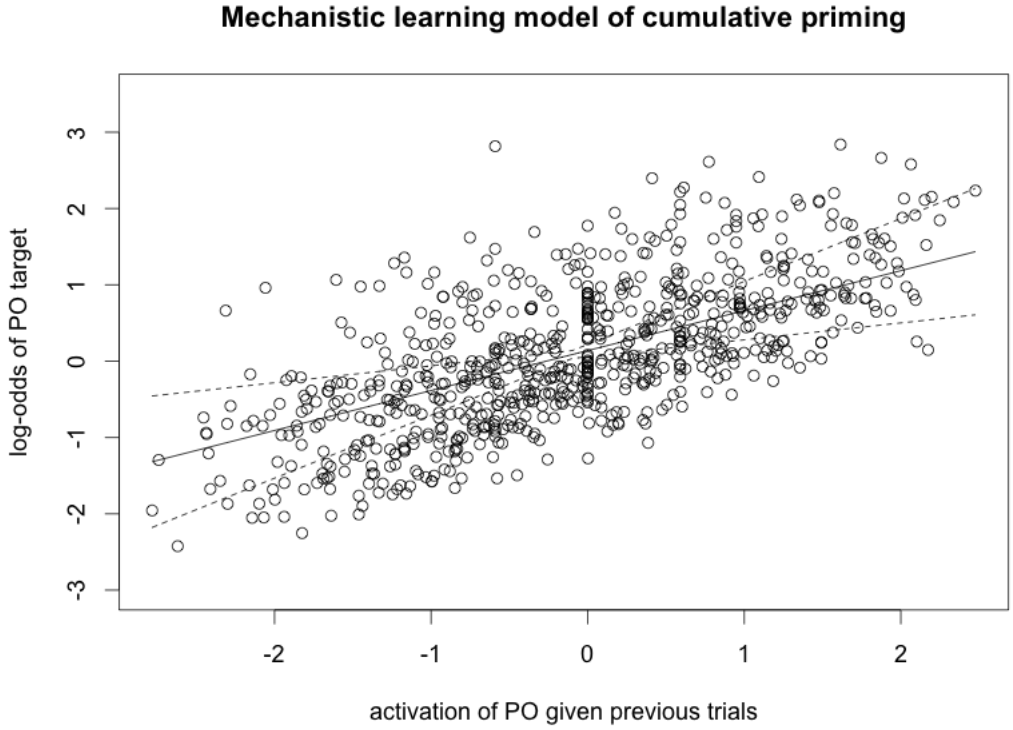

Figure 3: Iterative learning model with weighted experience

A parameter σ represents the weight of surprisal versus the baseline in determining the total activation, which is as follows:

$$\begin{aligned} Act(PO)_i = &(1-\sigma)BaseAct(PO)_i \\ &+ \sigma SurpAct(PO)_i \end{aligned} \tag{7}$$

The activation of PO increases if a PO is encountered, and decreases if a DO is encountered. If $\sigma = 1$, then the surprisal of the structure alone determines its activation with no baseline increase for having encountered the structure. If $\sigma = 0$, then surprisal has no effect on activation, encountering a PO increases activation by 1, and a DO decreases it by 1. Values of σ between 0 and 1 reflect intermediate weightings of baseline and surprisal-sensitive activation.

In order to determine the cumulative activation of the PO structure given previous experience at any target point, the activation from each encountered instance is summed, but each encounter is weighted by its temporal distance from the target to the power $-\delta$:

$$\begin{aligned} &TotalAct(PO)_{targ} = \\ &\sum_{i=1}^{\text{targ}} Act(PO)_i(\text{targ} - i)^{-\delta} \end{aligned} \tag{8}$$

In order to compare the performance of the mechanistic learning model with that of the Beta-binomial, we will first determine the best fitting parameters for surprisal weight (σ) and decay (δ). We will then compare this mechanistic model prediction for cumulative priming to the Beta-binomial model.

4.1 Results

As with the previous model, we set the parameters by optimizing the mixed effects regression model of the data with respect to the weighted learning model of cumulative priming with these parameters. Optimization was done by Quasi-Newton method (using the *optim* function in *R*, with subsequent brute-force verification). We found that the optimum weight of surprisal was $\sigma = 0.635$, and the optimal decay coefficient was $\delta = .119$. In the analysis, the controls are the same as in Study 1.

The weighted model is also a significant predictor of cumulative priming behavior: the more activation the PO node has given previous exposure with recent exemplars weighted higher, the more likely a PO is to be produced in the target (χ^2 = 19, $p < .001$). The prediction for the weighted model is compared with the data from the experiment in Figure (3). The weighted model has a higher χ^2, which may indicate that it is a stronger predictor of behavior in the experiment, a hypothesis we test directly in the next section.

4.2 Comparing the two models

We show above that the models both with and without experiential weighting predict cumulative priming behavior, at least taken on their own. However, it is possible that they explain some of the same variance, so we performed another analysis that compared the two directly in the same statistical model. We find that the weighted learning model is the only significant predictor of cumulative priming ($\chi^2 = 14.6, p < .001$), with the Beta-binomial model explaining no additional variance ($\chi^2 = .6$, $p = .43$). This further argues that the participants in the experiment did indeed weight recent exemplars more strongly. However, it is still not yet clear if this is a failure of models that weight exemplars equally, or just ones that do it via a Beta-binomial. As a further test, we also examined the effect of a baseline model that simply represents the likelihood of the PO structure as the proportion of PO structures processed so far. Comparing the weighted model with this baseline, yields the result that both the weighted model ($\chi^2 = 7.2$, $p = .007$), and if the baseline model ($\chi^2 = 5.7$, $p = .02$) predict the participants' production in the experiment. This indicates that there is a component to learning that weights all exemplars equally, as well as a component that weights recent exemplars more strongly.

We cannot draw strong conclusions about the comparison of the two models because we only compared them using the parameters that best fit when taken alone, rather than comparing them both over a range of parameters. However, we are most interested in comparing these two exact models, and they both have the same number of parameters (two). In future work, we will compare across a greater range of the parameter space.

5 Conclusion

The models presented here provide evidence for two types of learning during language processing, and it the first application of these two learning models to syntactic priming, to our knowledge. While the model that includes an explicit learning rule and weights recent experience more highly provides the best fit to the data, some of the variation in the experiment is also captured by the raw proportions processed in the experiment. The predictive power of the weighted learning model provides further evidence that people do weight recent experience more strongly. This is arguably a rational strategy because recent experience is more likely to be representative of the current situation, especially in artificial task-driven environments like laboratory experiments. There is also a component to learning that takes all experience into account equally. This opens the question of whether the best model of learning in human language processing might be a model that used Bayesian inference, but weighted evidence by recency.

Another interesting point is that the best-fitting Beta binomial has a prior that equally weights the PO and DO structures. This is despite the fact that these do not occur in equal proportion in participants' experience outside of the experiment. It may indicate that participants do not treat the experiment as being representative of their previous experiment, and instead as a domain unto itself. However, this cannot be the full explanation because the original experiment found effects of surprisal of the prime given experience prior to the experiment.

This paper also illustrates the usefulness of computational modeling of laboratory data. The models make processing and learning hypotheses explicit, and the laboratory data can be designed to include sufficient data from the key cases on which the hypotheses differ to tease them apart, such as the high and low proportions as well as distant and recent experience in the experiment here.

References

J.R. Anderson and C. Lebiere. 1998. *The atomic components of thought*. Lawrence Erlbaum.

J.R. Anderson and R. Milson. 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.

J.R. Anderson. 1983. *The Architecture of Cognition*. Lawrence Erlbaum Associates.

E. Bates and B. MacWhinney. 1989. Functionalism and the competition model. *The crosslinguistic study of sentence processing*, pages 3–73.

C.M. Bishop. 2006. *Pattern recognition and machine learning*. Springer New York:.

K. Bock and Z.M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning. *Journal of Experimental Psychology: General*, 129(2):177–192.

J.K. Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

J. Bresnan, A. Cueni, T. Nikitina, and H. Baayen. 2007. Predicting the dative alternation. *Proceedings of the Royal Netherlands Academy of Science Workshop on Foundations of Interpretation. Amsterdam*.

F. Chang, G.S. Dell, and K. Bock. 2006. Becoming syntactic. *Psychological Review*, 113(2):234–272.

M. Clayards, M.K. Tanenhaus, R.N. Aslin, and R.A. Jacobs. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809.

Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2-3):195–225.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, pages 159–166.

R.J. Hartsuiker and C. Westenberg. 2000. Word order priming in written and spoken sentence production. *Cognition*, 75(2):27–39.

R.J. Hartsuiker, S. Bernolet, S. Schoonbaert, S. Speybroeck, and D. Vanderelst. 2008. Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58(2):214–238.

T. F. Jaeger and N. Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

M.P. Kaschak, R.A. Loney, and K.L. Borreggine. 2006. Recent experience affects the strength of structural priming. *Cognition*, 99(3):73–82.

T. Kraljic and A.G. Samuel. 2005. Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51(2):141–178.

R. Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

R.L. Lewis and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–420.

Maryellen C. MacDonald and Morten H. Christiansen. 2002. Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109(1):35–54.

K. McRae, M.J. Spivey-Knowlton, and M.K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

M.C. Mozer, S. Kinoshita, and M. Shettel. 2007. Sequential dependencies in human behavior offer insights into cognitive control. *Integrated models of cognitive systems*, pages 180–193.

M.J. Pickering and H.P. Branigan. 1998. The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language*, 39(4):633–651.

Y. Sakamoto, M. Jones, and B.C. Love. 2008. Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & cognition*, 36(6):1057.

J. Snedeker and J.C. Trueswell. 2004. The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing* 1. *Cognitive Psychology*, 49(3):238–299.

Benedikt M. Szmrecsányi. 2005. Language users as creatures of habit: a corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, 1:113–149.

M.J. Traxler. 2008. Lexically independent priming in online sentence comprehension. *Psychonomic Bulletin and Review*, 15(1):149.

J.B. Wells, M.H. Christiansen, D.S. Race, D.J. Acheson, and M.C. MacDonald. 2009. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, 58(2):250–271.