DISCOURSE PROCESSES
https://doi.org/10.1080/0163853X.2018.1448677

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience

Noortje J. Venhuizen, Matthew W. Crocker, and Harm Brouwer

Saarland University, Saarbrücken, Germany

ABSTRACT

The processing difficulty of each word we encounter in a sentence is affected by both our prior linguistic experience and our general knowledge about the world. Computational models of incremental language processing have, however, been limited in accounting for the influence of world knowledge. We develop an incremental model of language comprehension that constructs—on a word-by-word basis—rich, probabilistic situation model representations. To quantify linguistic processing effort, we adopt Surprisal Theory, which asserts that the processing difficulty incurred by a word is inversely proportional to its expectancy (Hale, 2001; Levy, 2008). In contrast with typical *language model* implementations of surprisal, the proposed model instantiates a novel *comprehension-centric* metric of surprisal that reflects the likelihood of the unfolding utterance meaning as established after processing each word. Simulations are presented that demonstrate that linguistic experience and world knowledge are integrated in the model at the level of interpretation and combine in determining online expectations.

Introduction

Language is processed incrementally, continuously assigning meaning to the linguistic signal on a more or less word-by-word basis (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This entails the online integration of syntactic, semantic, and pragmatic information to arrive at a mental representation that reflects utterance or discourse meaning. Moreover, for this representation to be cohesive, it may need to be augmented with "world knowledge"-driven inferences that go beyond what is literally conveyed by the linguistic input. The resultant representation is a *mental model* (Johnson-Laird, 1983) or *situation model* (van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998): a mental representation of a described situation, grounded in our knowledge about the world.[1] One of the main aims of the study of language comprehension is to understand how such a mental model is constructed on an incremental, word-by-word basis.

In addressing this question, the sentence processing and discourse/text comprehension literature have produced comprehension theories and computational instantiations thereof, with rather different foci. In the field of sentence processing, the focus has been on modeling incremental processing and explaining word-by-word fluctuations in processing difficulty; as individual words vary in the degree of influence they have on the unfolding interpretation, some words induce greater cognitive processing effort than others. A variety of processing models has been put forward to explain these fluctuations (Gibson, 1998, 2000; Hale, 2001, 2006, 2011; Jurafsky, 1996; Levy, 2008; McRae, Spivey-Knowlton, & Tanenhaus, 1998). Critically, computational instantiations of these

CONTACT Noortje J. Venhuizen ✉ noortjev@coli.uni-saarland.de Department of Language Science & Technology, Saarland University, Building C7.1, Room 1.19, Saarbrücken 66123, Germany

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hdsp.

[1] Henceforth, we will use the terms "mental model" and "situation model" interchangeably.

models are typically limited to modeling the influence of the linguistic context in which a word occurs, and hence they do not take into account the influence of the unfolding situation model. This is exemplified by recent instantiations of *surprisal theory*, a processing theory that has been particularly successful in offering a broad account of word-by-word processing difficulty (Hale, 2001; Levy, 2008).

Surprisal theory asserts that the processing difficulty incurred by a word is linearly related to its *surprisal*, a metric that is itself inversely proportional to the expectancy of a word: the less expected a word is in a given context, the higher its surprisal, and hence the greater its processing difficulty. More formally, given a sentence $w_1 \dots w_i$ so far, the surprisal of a next word w_{i+1} is defined as: $\mathrm{surprisal}(w_{i+1}) = -\log P(w_{i+1}|w_1 \dots w_i)$ (and $\mathrm{difficulty}(w_{i+1}) \propto \mathrm{surprisal}(w_{i+1})$). Surprisal estimates derived from various language models (e.g., Probabilistic Context-Free Grammars [PCFGs], Simple Recurrent Networks [SRNs], Tree-Adjoining Grammars [TAGs], and N-gram models) have been shown to correlate with behavioral metrics of word processing difficulty (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Brouwer, Fitz, & Hoeks, 2010; Demberg & Keller, 2008; Frank, 2009; Roark, Bachrach, Cardenas, & Pallier, 2009; Smith & Levy, 2008). Crucially, however, all of these instantiations of surprisal theory model comprehension as next word prediction based on linguistic experience—the probability of the word in the linguistic context, as determined by long-term experience with linguistic input—and thus offer no account of how words are combined with world knowledge into a mental model and how the unfolding mental model affects the prediction of upcoming words.

In the discourse/text comprehension literature, by contrast, the focus has traditionally been on explaining mental model construction and representation (e.g., Golden & Rumelhart, 1993; Kintsch, 1988, 1998, 2001; Myers & O'Brien, 1998; St. John, 1992; St. John & McClelland, 1990, 1992; Langston & Trabasso, 1999; van den Broek, Risden, Fletcher, & Thurlow, 1996). Models of discourse comprehension typically aim to provide a mechanistic account of how "world knowledge"-driven inferences complement the propositions conveyed by the literal linguistic input to arrive at a cohesive mental model of what a text is about. These models do not, however, explain how such a mental model is derived on an incremental, word-by-word basis; that is, processing in these models typically starts from entire sentences or sets of propositions.

In recent years, the influence of world knowledge on incremental language processing has obtained greater significance in both the empirical and the theoretical literature. On the empirical front, a wide range of findings has shown that the processing difficulty of individual words is affected by the larger discourse context and general knowledge about the world, above and beyond linguistic experience alone (see, e.g., Albrecht & O'Brien, 1993; Altmann & Kamide, 1999; Camblin, Gordon, & Swaab, 2007; Cook & Myers, 2004; Garrod & Terras, 2000; Hess, Foss, & Carroll, 1995; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Knoeferle, Habets, Crocker, & Münte, 2008; Kuperberg, Paczynski, & Ditman, 2011; Morris, 1994; Myers & O'Brien, 1998; O'Brien & Albrecht, 1992; Otten & van Berkum, 2008; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; van Berkum, Hagoort, & Brown, 1999; van Berkum, Zwitserlood, Hagoort, & Brown, 2003). As such, recent theories of text comprehension have emphasized the importance of world knowledge on incremental comprehension by arguing that the *validation* of message consistency is a central part of the comprehension process (Cook & O'Brien, 2014; Isberner & Richter, 2014; O'Brien & Cook, 2016; Richter, 2015; Singer, 2006, 2013; Singer & Doering, 2014).

Here, we offer a computational model of incremental comprehension that explicates how world knowledge and linguistic experience are integrated at the level of interpretation and combine in determining online expectations. More specifically, we present a neural network model that constructs a representation of utterance meaning on an incremental, word-by-word basis. These meaning representations are rich probabilistic mental models that go beyond literal propositional content, capturing inferences driven by knowledge about the world (cf. Frank, Koppen, Noordman, & Vonk, 2003; Golden & Rumelhart, 1993). We show that word surprisal (i.e., processing difficulty; Hale, 2001; Levy, 2008) can be naturally characterized in terms of the incremental construction of these meaning representations. Crucially, this "comprehension-centric" notion of surprisal is affected by both *linguistic experience* (the linguistic input history of the model) as well as *world knowledge* (the model's probabilistic knowledge about the world). We systematically explore this interaction between linguistic experience and world knowledge in the online processing behavior of the model. We demonstrate that, like human comprehenders, the model's surprisal values are influenced by both its linguistic experience and its knowledge about the world. That is, we show that like traditional surprisal models, our model captures surprisal effects driven by its linguistic experience, but that above and beyond that, it is able to seamlessly capture processing phenomena that require more comprehensive, possibly nonlinguistic, knowledge about the world.

Below, we will first describe the meaning representations that support comprehension in our model. Next, we describe the model's architecture and training procedure and show how the model's processing behavior naturally allows for the derivation of a "comprehension-centric" surprisal metric. We then describe the model's behavior when presented with sentences in which linguistic experience and world knowledge are manipulated. Finally, we discuss the implications of our model, as well as directions for future research.

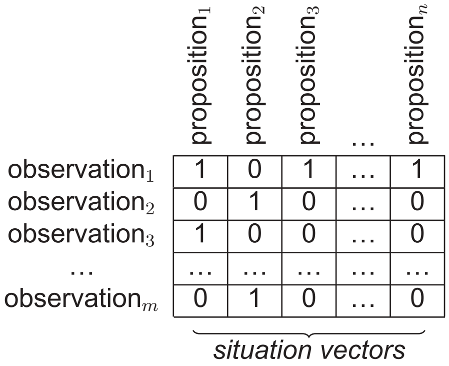
Situation models in distributed situation-state space

Golden and Rumelhart (1993) developed a framework for modeling story comprehension, which represents mental models as points in a high-dimensional state space called "situation-state space." Frank and colleagues (2003) adapted this situation-state space model by incorporating a distributional notion of propositional meaning so as to capture the dependency between individual propositions. In the resulting Distributed Situation-state Space (DSS) model, each proposition is represented as a *situation vector*: a sequence of 1's and 0's that for a given set of *observations* describes whether the proposition is the case (1) or not (0). These observations describe independent states-of-affairs in the world, defined in terms of combinations of propositions. Formally, a DSS is an $m \times n$ matrix that is constituted by a large set of m observations of states-of-affairs in the world, defined in terms of n atomic propositions (e.g., $enter(beth, restaurant)$ and $order(thom, dinner)$)—the smallest discerning units of propositional meaning. Each of the m observations in this matrix is encoded by setting atomic propositions that are the case in a given observation to 1/True and those that are not to 0/False (see Figure 1). The resulting situation-state space matrix is effectively one big truth table, in which each column represents the situation vector for its corresponding atomic proposition—that is, from a geometric perspective, a point in situation-state space.

Situation vectors capture meaning in terms of cooccurrence between *propositions*, as opposed to meaning vectors derived using other distributional methods, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), which capture word meaning in terms of *linguistic* cooccurrence. In what follows, we show that DSS-derived situation vectors offer meaning representations that effectively instantiate situation models. We will first describe how a DSS can be constructed using a "microworld" approach, and show that the DSS-derived situation vectors are inherently compositional and probabilistic (Frank, Haselager, & van Rooij, 2009; Frank et al., 2003). Then, we will illustrate how comprehension in DSS is formalized as navigation through situation-state space.

Deriving a DSS from a microworld

The observations constituting the DSS instantiate independent, episodic experiences in the world. Crucially, these experiences inherently carry systematic knowledge about the world: some combinations of events never occur, and some combinations occur more frequently than others. Within the DSS, this means that each observation must conform to specific world knowledge constraints for situation vectors to capture such dependencies between propositions, and, moreover, the complete DSS must obey the probabilistic nature of the world, according to which some (combinations of) propositions may be more likely than others. Hence, an important requirement for the derivation of



	proposition_1	proposition_2	proposition_3	…	proposition_n
observation_1	1	0	1	…	1
observation_2	0	1	0	…	0
observation_3	1	0	0	…	0
…	…	…	…	…	…
observation_m	0	1	0	…	0

Figure 1. Example of an $m \times n$ DSS matrix in which the rows represent observations, and the columns represent the situation vectors associated with the atomic propositions.

a DSS is the notion of "world knowledge." Ideally, the DSS used in our comprehension model would incorporate the same world knowledge as human comprehenders. It is computationally infeasible, however, to incorporate the entirety of a person's knowledge about the real world. To overcome this, we adopt a "microworld" strategy, in which we limit the scope of the world, rather than that of the knowledge encoded about the world; that is, we encode all relevant knowledge about a confined microworld (cf. Frank et al., 2003; Frank, Koppen, Noordman, & Vonk, 2008; Golden & Rumelhart, 1993).

The purpose of the microworld is to provide our model with world knowledge that interacts with incremental linguistic interpretation. To this end, the world knowledge captured within the microworld should be illustrative of the type of world knowledge used by human comprehenders. Because our model will be trained on a restricted set of sentences (a "microlanguage"), it suffices for the present purposes to manually construct a suitable microworld that can be used for deriving a DSS (but see the Discussion for alternative ways for deriving a DSS). That is, we define a microworld in terms of a set of propositions and cooccurrence constraints on these propositions that reflect world knowledge; these may include "hard" constraints (e.g., certain propositions never cooccur) or "soft" probabilistic constraints (e.g., certain propositions are more likely to cooccur than others). Based on such a microworld, a DSS can be derived by sampling observations in such a way that each observation satisfies the "hard" world knowledge constraints, and the entire set of observations approximately reflects the probabilistic structure of the world (see Appendix A).

The DSS employed in our model reflects a microworld that describes a (fictional) set of events taking place on an evening out. As an example of the type of world knowledge used during online comprehension, the microworld captures the information conveyed by *scripts* (or *schemata*; Rumelhart, 1975, 1977; Schank & Abelson, 1977). A script is a temporally ordered sequence of events that are typically part of the activity described by the script; for example, a script for going to the restaurant might describe a sequence of events that includes entering, reading the menu, ordering, eating, and paying. To capture this type of world knowledge, we define a microworld in terms of a set of $n = 45$ propositions constructed using the predicates *enter, ask_menu, order, eat, drink, pay*, and *leave* (see Tables 1 and 2). For instance, the atomic proposition *enter(dave,restaurant)* represents the proposition that "Dave entered the restaurant." The microworld is defined in terms of three types of world knowledge constraints: (i) some propositions are more likely (typical) than other propositions; (ii) some combinations of propositions are more likely than other combinations, and (iii) observations induce a temporal ordering on propositions.

To induce principles (i) and (ii), we sampled our observations in such a way that certain "typical" propositions and combinations of propositions occur in more observations than others. Typical

Table 1. Microworld concepts.

Class	Variable	Class members
Persons	*x*	beth, dave, thom
Places	*p*	cinema, restaurant
Foods	*f*	dinner, popcorn
Drinks	*d*	champagne, cola, water
Predicates	-	enter, ask menu, order, eat, drink, pay, leave

Table 2. Basic propositions.

Proposition	*n*
enter (*x*, *p*)	6
ask menu (*x*)	3
order (*x*, *d*), order (*x*, *f*)	15
eat (*x*, *f*)	6
drink (*x*, *d*)	9
pay (*x*)	3
leave (*x*)	3
Total	45

atomic propositions in our microworld are *order(x,water)* and *drink(x,water)* (for any person x), and typical combinations of propositions are $enter(x,restaurant) \wedge order(x,dinner)$ and $enter(x,cinema) \wedge order(x,popcorn)$. Furthermore, persons can only order one type of food in a single observation, and they tend not to order multiple drinks. Finally, we constrain persons to only enter at most one place in a single observation to make sure that each observation constitutes an unambiguous sequence of propositions. To induce principle (iii), we define observations to be "snapshots" along a timeline—that is, a single observation reflects a temporally extended sequence of propositions. Within this timeline, we make a distinction between necessary and optional propositions: necessary propositions cannot be skipped when observing a sequence, while optional propositions can be skipped. The necessary propositions in our microworld are *order* and *pay*. The proposition *order(x, d/f)* is necessary in case one of the following (combinations of) propositions holds: *drink(x,d), eat(x, f)*, $enter(x,p) \wedge pay(x)$, or $ask\_menu(x) \wedge pay(x)$. Similarly, $pay(x)$ must hold in case $order(x,d/f) \wedge leave(x)$ holds. Thus, temporal order is not explicitly encoded within our DSS, but it emerges from the way in which the observations are sampled—we will return to this point in the Discussion section.

Following the above described principles, we derived a DSS by sampling 15K microworld observations, using a nondeterministic, inference-based sampling algorithm (see Appendix A). This number was chosen empirically by verifying that the resulting DSS captured all required probabilistic dependencies defined by the constraints. To make the situation vectors suitable for application in a neural network architecture, we reduced the dimensionality of the DSS to 150 dimensions ($m = 150$, cf. Figure 1) using a dimension selection algorithm, which preserves the information encoded and maintains the probabilistic structure of the microworld (see Appendix B). This number was chosen empirically as well, by manually finding an optimal trade-off between lowering dimensionality and maintaining the (probabilistic) structure of the original 15K DSS. Below, we describe the compositional and probabilistic properties of situation vectors, and we show that the resulting DSS captures the above-described world knowledge constraints.

Compositionality and probability in DSS

In DSS, propositions are represented as situation vectors that describe for each observation whether the proposition is the case in this observation. A situation vector $\vec{s}(a)$ captures the meaning of a given proposition a in terms of its cooccurrence with other propositions in the world; two propositions are highly similar in case they occur in largely the same set of observations, and they are different in case they do not occur in the same observations. As a result, the meaning of the negation of proposition a is described by the situation vector $\vec{s}(\neg a)$ that assigns a 0 to all observations in which a is the case, and a 1 otherwise (thus resulting in a maximally different situation vector relative to $\vec{s}(a)$); this vector can be directly derived from $\vec{s}(a)$, the situation vector of a, as follows: $\vec{s}(\neg a) = 1 - \vec{s}(a)$ (Frank et al., 2009). In a similar manner, the meaning of the conjunction between propositions a and b will be described by the situation vector that assigns 1 to all observations in which both a and b are the case and 0 otherwise; this vector can be calculated by the pointwise multiplication of the situation vectors of a and b: $\vec{s}(a \wedge b) = \vec{s}(a)\vec{s}(b)$. Because the negation and conjunction operators together define a functionally complete system, the meaning of any other logical combination between propositions in situation-state space can be described using these two operations (in particular, the situation vector representing the disjunction between a and b, $\vec{s}(a \vee b)$, is defined as $\vec{s}(\neg(\neg a \wedge \neg b))$, which assigns a 1 to all observations in which either a or b is the case, and a 0 otherwise). Hence, we can combine atomic propositions into complex propositions, which can in turn be combined with other atomic and complex propositions, thus allowing for situation vectors of arbitrary complexity.

Besides their compositional nature, the distributive meaning representations inherently encode the (co)occurrence probability of propositions; on the basis of the m observations in the situation-state space matrix, we can estimate the prior probability *Pr(a)* of the occurrence of each (basic or complex) proposition a in the microworld from its situation vector $\vec{s}(a)$. We call this probability estimate the prior belief value $B(a)$ of a:

$$B(a) = \frac{1}{m}\sum_i \vec{s}_i(a) \approx Pr(a). \tag{1}$$

Figure 2 compares the prior belief values of a subset of the atomic propositions (those pertaining to *beth*) of the microworld. These values reflect the world knowledge constraints; typical propositions, e.g., *order(beth, water)*, obtain higher prior belief values than atypical propositions, e.g., *order(beth, champagne)*. Note that because of the way we construct the microworld, these values transfer to the other persons (*thom* and *dave*).

Crucially, because of the compositional nature of situation vectors, they also inherently encode the occurrence probability of combinations of propositions. That is, the conjunction probability of

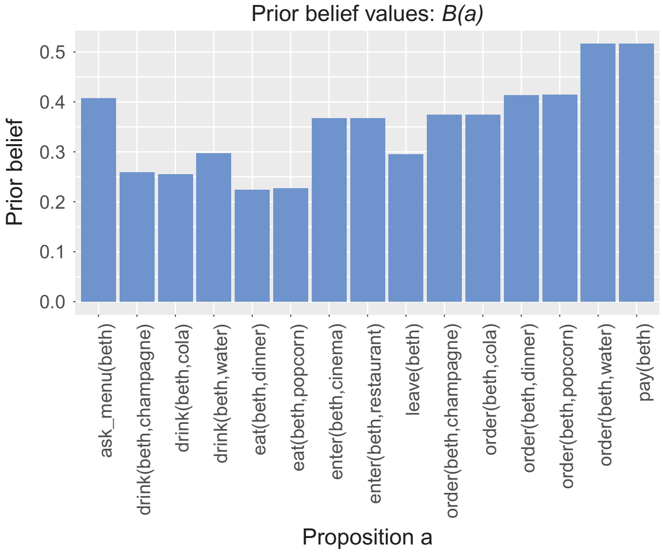


Figure 2. Prior belief values for a subset of the atomic propositions (those pertaining to *beth*) of the microworld.

the cooccurrence of two propositions a and b can be estimated by calculating the prior belief value of their conjunction vector $B(a \wedge b)$. As a result, the conditional probability of a proposition a given b can be estimated as follows:

$$B(a|b) = \frac{B(a \wedge b)}{B(b)} \approx Pr(a|b). \tag{2}$$

This means that, given a proposition b, we can infer any proposition a that depends on b in the microworld. This allows us to quantify how individual propositions are related to each other. More specifically, we can determine how much a proposition a is "understood" from b (i.e., how much proposition b contributes to the understanding of proposition a), by calculating the *comprehension score* (Frank et al., 2009):

$$comprehension(a, b) = \begin{cases} \frac{B(a|b) - B(a)}{1 - B(a)} & \text{if} B(a|b) > B(a) \\ \frac{B(a|b) - B(a)}{B(a)} & \text{otherwise.} \end{cases} . \tag{3}$$

If a is understood to be the case from b, the conditional belief $B(a|b)$ should be higher than the prior belief $B(a)$: knowing b increases belief in a. Conversely, if a is understood *not* to be the case from b, the conditional belief $B(a|b)$ should be lower than the prior belief $B(a)$: knowing b decreases belief in a. The score $comprehension(a, b)$ quantifies this by means of a value ranging from $+1$ to -1, where $+1$ indicates that proposition a is perfectly understood to be the case from b (b took away all uncertainty in a; b entails a), and a value of -1 indicates that a is perfectly understood *not* to be the case from b (b took away all certainty in a; b entails $\neg a$). Below, we will use this score to investigate how the world knowledge constraints described in the previous section are reflected in the microworld.

Situation vectors as situation models

The comprehension score defined in equation (3) defines on a scale from -1 to $+1$ how much a proposition a is understood (or: inferred) from a given proposition b. This means that we can investigate the semantic structure of the microworld by looking at the comprehension scores between individual propositions. Figure 3 shows the comprehension score for each pair of propositions a and b of a subset of the atomic propositions (those pertaining to *beth*) of the microworld. This figure shows that the DSS successfully captures the cooccurrence constraints of the microworld described above. For instance, given the proposition $b = order(beth, popcorn)$, the proposition $a = order(beth, dinner)$ is perfectly understood *not* to be the case ($comprehension(a, b) = -1$); this corresponds to the constraint that persons can only order one type of food. Moreover, from the proposition $b = enter(beth, restaurant)$, it is positively inferred that $a = order(beth, dinner)$, whereas $a = order(beth, popcorn)$ is negatively inferred; this reflects the cooccurrence constraints on entering and ordering. Finally, the temporal ordering principles are also reflected in these comprehension scores: given $b = drink(beth, champagne)$, there is a certain inference that proposition $a = order(beth, champagne)$ holds ($comprehension(a, b) = 1$), but not vice versa.

The comprehension scores in Figure 3 illustrate that the situation vectors associated with individual propositions (as well as combinations thereof) capture more than propositional meaning alone; they also capture (probabilistic) "world knowledge"-driven inferences. In other words, situation vectors are effectively rich situation models that describe the meaning of propositions in relation to each other. On a geometric interpretation, situation vectors represent points in situation-state space. Figure 4 provides a visualization of the constructed situation-state space. This figure is a three-dimensional representation of the 150-dimensional DSS (for a subset of the atomic propositions), derived using multidimensional scaling (MDS). As a consequence of the scaling procedure, there is a significant loss of information, which means that the distances

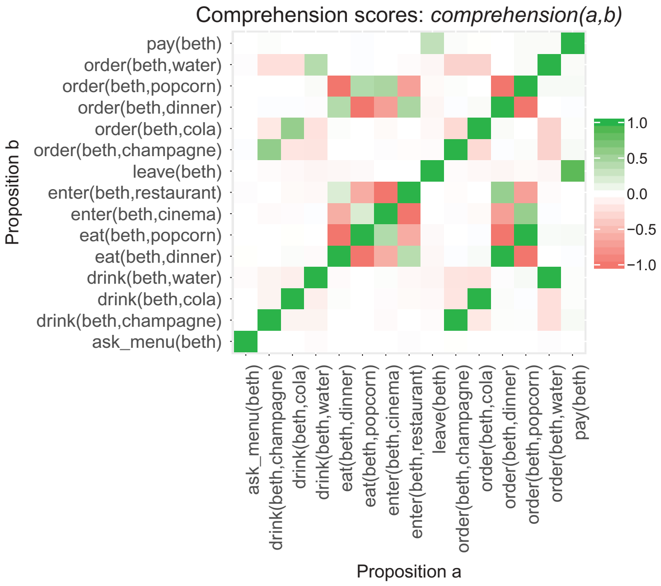


Figure 3. Comprehension score for each pair of propositions *a* and *b* of a subset of the atomic propositions (those pertaining to *beth*) of the microworld.

between the individual points do not directly map onto the distances within the DSS, and hence they should be interpreted with care. Nonetheless, this visualization serves to illustrate some interesting properties of the semantic space. In particular, it shows that within the DSS, meaning is defined in terms of cooccurrence within the observations. That is, propositions that frequently cooccur, e.g., *enter(beth,cinema)* and *order(beth,popcorn)*, are positioned close to each other in semantic space, whereas propositions that do not cooccur, e.g., *enter(beth,cinema)* and *enter (beth,restaurant)*, are positioned far away from each other. In other words, *enter(beth,cinema)* and *order(beth,popcorn)* have a higher semantic similarity in DSS than *enter(beth,cinema)* and *enter(beth,restaurant)*, as reflected by their positioning in space.

On the geometric interpretation of DSS, comprehending a linguistic utterance means navigating through situation-state space; each input word provides a cue for arriving at the utterance-final point in situation-state space, which may correspond to a single proposition, or a combination of multiple propositions.

Modeling surprisal beyond the words given

Our aim is to arrive at a model of incremental language comprehension in which expectation about upcoming words reflects both world knowledge and linguistic experience. Specifically, we present a Simple Recurrent Network (SRN) that takes sequences of words as input, and that constructs a DSS-derived situation model representation of utterance meaning on an incremental, word-by-word basis. Below, we describe the architecture and training procedure of our model. Moreover, we demonstrate that surprisal values can be directly derived from the probabilistic meaning representations induced by individual words. In the next section, we show that these values reflect the integration of linguistic experience and world knowledge in online language comprehension.

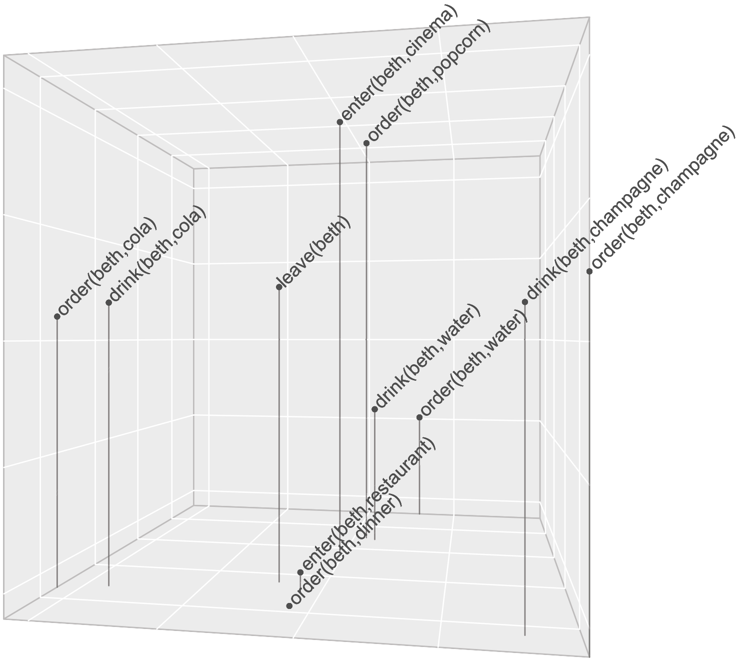


Figure 4. Visualization of the DSS into three dimensions (by means of multidimensional scaling; MDS) for a subset of the atomic propositions (the predicates *enter, order, drink*, and *leave*, applied to *beth*).

Model architecture

To develop a psychologically plausible comprehension model that incrementally maps words onto DSS representations while taking into account linguistic experience, we employ an SRN (Elman, 1990), consisting of four groups of artificial neurons: an input layer, a hidden layer, a context layer, and an output layer (see Figure 5). On the input layer, the model is presented with individual words, which are represented using "localist" word representations—vectors consisting of 21 units (reflecting the total number of words in the training data), in which a single unit (reflecting the word) is set to 1, and all other units are set to 0. We employ this basic word encoding scheme to keep the model's behaviour transparent. It should be noted, however, that nothing prevents the model from employing more realistic distributed word representations (for instance, to capture orthographic/acoustic similarity between words; see, e.g., Laszlo & Plaut, 2012). The activation of the input is "feed-forwarded" through the hidden layer to the output layer, which constitutes a DSS-derived situation vector $\vec{s}$. The hidden layer also receives input from the context layer, which is itself a copy of the previous state of the hidden layer. This context layer allows the model to take into account the previous words in a sentence when mapping word representations to situation vectors. Finally, both the hidden layer and the output layer receive input from a single bias unit, the activation value of which is always set to 1; this effectively introduces a threshold for the activation of each of the units in these layers.

The feed-forward activation of the hidden and output layers is determined for each of the individual units in these layers based on the input they receive. For a single unit j, the net input x_j is determined by the activation level y_i of each unit i that propagates to unit j, and the weight w_{ij} on the connection from i to j:

$$x_j = \sum_i y_i w_{ij}. \tag{4}$$

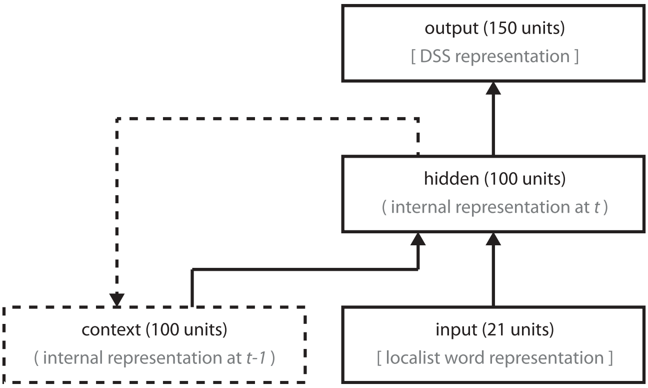

output (150 units)
[DSS representation]
hidden (100 units)
(internal representation at t)
context (100 units)
(internal representation at t-1)
input (21 units)
[localist word representation]

Figure 5. Model architecture. Boxes represent groups of artificial neurons, and solid arrows between boxes represent full projections between the neurons in a projecting and a receiving group. Prior to feed-forward activation at time-step t, the context group receives a copy of the activation pattern of the hidden layer at time-step $t - 1$. The hidden and output layer also receive input from a single bias unit (omitted in this figure), the activation value of which is always set to 1.

The activation level y_j of unit j, then, is defined as follows:

$$y_j = \frac{1}{1 + e^{-x_j}}. \tag{5}$$

Given a localist word representation on the input layer, therefore, feed-forward activation involves calculating the activation level of the units in the hidden layer based on the activation in the input layer, the weights between the input and hidden layer, the activation in the context layer and the associated weights, as well as the bias associated with the hidden layer. The activation level of the units in the hidden layer, then, can be used in combination with the weights between the hidden and output layer, as well as its associated bias, to calculate the activation level of the units in the output layer (i.e., the vector $\vec{s}$ constituting the DSS-derived situation model). The weights themselves are determined by training the model to map sequences of localist word representations to situation vectors.

Training the comprehension model

During the training phase, the model is presented with activation patterns on the input layer (i.e., localist word representations) and on the output layer (DSS-derived situation vectors). That is, using the feed-forward dynamics outlined above, the model processes a sequence of words constituting a sentence and produces an activation pattern representing sentence meaning. Next, it is determined how well the model recovers the intended sentence meaning; that is, how much the produced activation pattern differs from its target activation pattern (the DSS situation vector presented at the output layer). The resulting error signal is then "backpropagated" through the model (Rumelhart, Hinton, & Williams, 1986),[2] and on the basis of this signal, each of the weights is slightly adjusted, such that the model's error will be reduced upon the next encounter of this sentence.

The training items presented to the model are sentences combined with a situation vector representing the meaning conveyed by that sentence. Presentation of a single training item therefore means presenting the model with localist representations of each of the words in the sentence together with the (sentence-final) situation vector. The input sentences to the model are

[2] We here employ a variation on the standard backpropagation algorithm, called bounded gradient descent (Rohde, 2002).

Table 3. Grammar of the language used for training. Optional arguments are in square brackets, and different instantiations of a rule are separatedusing the pipe symbol. Variable $V \in \{enter, menu, order, eat, drink, pay, leave\}$ denotes verb types.

Head		Body
S	→	NP$_{person}$ VP$_{V}$ [CoordVP$_{V}$]
NP$_{person}$	→	beth | dave | thom
NP$_{place}$	→	the cinema | the restaurant
NP$_{food}$	→	dinner | popcorn
NP$_{drink}$	→	champagne | cola | water
VP$_{enter}$	→	entered NP$_{place}$
VP$_{menu}$	→	asked for the menu
VP$_{order}$	→	ordered NP$_{food}$ | ordered NP$_{drink}$
VP$_{eat}$	→	ate NP$_{food}$
VP$_{drink}$	→	drank NP$_{drink}$
VP$_{pay}$	→	paid
VP$_{leave}$	→	left
CoordVP$_{enter}$	→	and VP$_{menu}$ | and VP$_{order}$ | and VP$_{leave}$
CoordVP$_{menu}$	→	and VP$_{order}$ | and VP $_{leave}$
CoordVP$_{pay}$	→	and VP$_{order}$ | and VP $_{leave}$

derived using a basic grammar, which is shown in Table 3. The grammar defines two types of sentences: simple sentences (NP VP), describing basic propositions in the microworld, and coordinated sentences (NP VP CoordVP), describing combinations between basic propositions. The semantics of these sentences are defined accordingly: the simple sentence "dave drank cola" is associated with the situation vector described by *drink(dave,cola)*, and the coordinated sentence "beth entered the restaurant and ordered dinner" is associated with the situation vector described by *enter(beth,restaurant)* ∧ *order(beth,dinner)*. The grammar combines 21 words into a total of 117 different sentences, each describing a unique situation. To induce differential linguistic experience in the model, some of these sentences are encountered more often than others during training. The frequent sentences are divided into two groups, highly frequent sentences (×9): "NP$_{person}$ ordered dinner," "NP$_{person}$ ate popcorn," "NP$_{person}$ ordered champagne," "NP$_{person}$ drank water"; and relatively frequent sentences (×5): "NP$_{person}$ ordered cola," "NP$_{person}$ drank cola." These frequencies make sure that there are no overall frequency differences between the target and control words for the experiments described in the next section. The total training set consisted of 237 sentences.

Prior to training, weights were randomly initialized in the range $(-.5, +.5)$. For each training item $[\langle\vec{w}_1, \ldots, \vec{w}_n\rangle, \vec{s}]$ (word sequence, meaning), error was backpropagated after each word, using a zero error radius of 0.05, meaning that no error was backpropagated if the error on a unit fell within this radius. Training items were presented in permuted order, and weight deltas were accumulated over epochs consisting of all training items. At the end of each epoch, weights were updated using a learning rate coefficient of 0.2, which was scaled down with a factor of 0.9 after each block of 500 epochs, and a momentum coefficient of 0.9. Training lasted for 5000 epochs, after which the mean squared error was 0.33. The overall performance of the model was assessed by calculating the cosine similarity between each sentence-final output vector and each target vector for all sentences in the training data. All output vectors had the highest cosine similarity to their own target (mean $= .99$; sd $= .01$), indicating that the model successfully learned to map sentences onto their corresponding semantics. An additional way to quantify performance is to compute for each sentence how well the intended target is "understood" from the output of the model: $comprehension(\vec{s}_{target}, \vec{s}_{output})$; see equation (3). The average comprehension score over the entire training set was 0.89, which means that after processing a sentence, the model almost perfectly infers the intended meaning of the sentence.

Online surprisal from situation models

Our comprehension model constructs an interpretation of an utterance on an incremental, word-by-word basis. After processing an entire utterance, the activation pattern at the output layer of the model forms a situation vector representing the meaning of that utterance: a point in situation-state space. Crucially, the model will produce a vector representing a point in situation-state space after each word of a sentence. This means that during comprehension, the model navigates situation-state space to ultimately arrive at the meaning of an utterance. The intermediate points in state space are effectively accumulations of evidence about the sentence-final utterance meaning and may not exactly correspond to situation vectors of (combinations of) atomic propositions; that is, one may envision these points to lie at the crossroads of several potential sentence-final utterance meanings consistent with the sentence so far.

Like all points in situation-state space, each intermediate point $\vec{s}_i$, as determined by a sequence of words $w_1 \ldots w_i$, inherently carries its own probability in the microworld. This means that we can study what the model "understands" at each word of a sentence by computing a comprehension score $comprehension(a, \vec{s}_i)$ for any proposition a; see equation (3). Figure 6 shows the word-by-word comprehension scores for the sentence "Beth entered the restaurant and ordered champagne" with respect to 6 selected propositions. This figure shows that by the end of the sentence, the model has understood its meaning: $enter(beth, restaurant) \wedge order(beth, champagne)$. Critically, it arrives at this interpretation on an incremental, word-by-word basis. For instance, at the word "restaurant," the model commits to the inference *enter(beth,restaurant)*, which rules out *enter(beth,cinema)* (because these propositions never cooccur). That is, the model navigates to a point in DSS that is close to the vector corresponding to *enter(beth,restaurant)* and far away from the vector corresponding to *enter(beth,cinema)*. This also leads the model to infer that *order(beth,dinner)* is likely (ordering dinner is something that is often done in a restaurant). At the word "champagne," the model draws the inference that *order(beth,champagne)*, and it slightly lowers its belief that *order(beth, water)* (although it does not completely rule it out). Finally, no significant inferences are drawn about the unrelated proposition *leave(dave)*.

Utterance comprehension in the model thus involves word-by-word navigation of situation-state space, gradually moving from a point of relative indecision about which propositions are inferred to be the case, to a point capturing the meaning of the utterance. Critically, this navigational process is guided by the "linguistic experience" of the model; it learns that certain sequences of words are more frequent in the training data, which means that during utterance comprehension it will move to a point in situation-state space that reflects the meaning of the expected continuation of the utterance. Thus, each incoming word effectively provides the

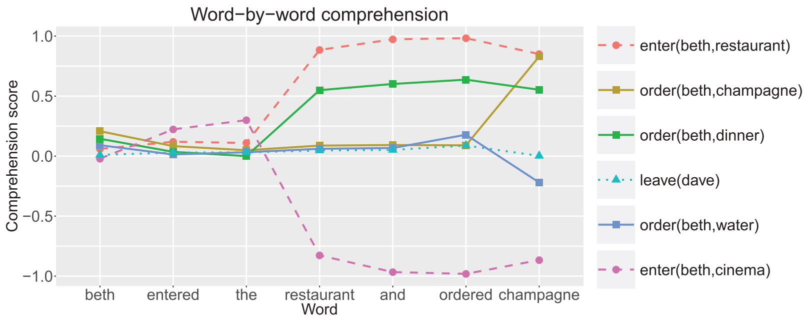


Figure 6. Word-by-word comprehension scores of selected propositions for the sentence "Beth entered the restaurant and ordered champagne" with the semantics: $enter(beth, restaurant) \wedge order(beth, champagne)$.

model with a cue toward the sentence-final utterance meaning. More formally, each incoming word w_{i+1} induces a transition from the current point in state-space $\vec{s}_i$ to the next $\vec{s}_{i+1}$. In case the point $\vec{s}_{i+1}$ is close to point $\vec{s}_i$ in the DSS, the transition induced by word w_{i+1} is small, and hence the word itself is unsurprising. If, on the other hand, point $\vec{s}_{i+1}$ is far away from point $\vec{s}_i$, the transition induced by word w_{i+1} is big, which means that this word is highly surprising. We can formalize this notion of surprisal by exploiting the probabilistic nature of the situation vectors. Following equation (2), we can estimate the conditional probability $P(\vec{s}_{i+1}|\vec{s}_i)$ of state $\vec{s}_{i+1}$ given state $\vec{s}_i$, which allows us to quantify the level of surprise in the transition incurred by word w_{i+1}:

$$s_{onl}(w_{i+1}) = -\log P(\vec{s}_{i+1}|\vec{s}_i). \tag{6}$$

To investigate the influence of linguistic experience and world knowledge on this online surprisal metric, we compare it to two baseline surprisal metrics: offline *linguistic* surprisal and offline *situation* surprisal. Both of these metrics are offline metrics, as they are estimated from the training data, rather than from the processing behavior of the model itself. Offline linguistic surprisal reflects linguistic experience and is straightforwardly estimated from the sentences on which the model is trained (cf. Hale, 2001; Levy, 2008):

$$\begin{aligned} s_{ling}(w_{i+1}) &= -\log P(w_{i+1}|w_{1,\ldots,i}) \\ &= -\log \tfrac{P(w_{1,\ldots,i+1})}{P(w_{1,\ldots,i})}. \end{aligned} \tag{7}$$

If a word w_{i+1} frequently occurs after words $w_1 \ldots w_i$, therefore, its conditional probability will be high, and its surprisal low (and vice versa). Crucially, this linguistic surprisal metric is not influenced by the world knowledge contained within situation vectors; it solely derives from the distribution of word sequences on which the model is trained.

Offline situation surprisal, in turn, reflects world knowledge, and it is also estimated from the training data (cf. Frank & Vigliocco, 2011). Rather than from the training sentences, however, it is derived from the (150-dimensional) situation vectors corresponding to those sentences. The situation vector for a sequence of words $w_1 \ldots w_i$, $sit(w_{1,\ldots,i})$, is derived by taking the disjunction of the semantics of all sentences that are consistent with this prefix. For instance, the situation vector of the prefix 'Dave drank' is defined as $sit(Dave\ drank) = \vec{s}(drink(dave, water) \vee drink(dave, cola) \vee drink(dave, champagne))$. The offline situation surprisal induced by a next word is then defined as follows:

$$s_{sit}(w_{i+1}) = -\log P(sit(w_{1,\ldots,i+1})|sit(w_{1,\ldots,i})). \tag{8}$$

If an incoming word w_{i+1} leads to a situation vector that is highly likely given the situation vector for the disjunctive semantics consistent with the words $w_1 \ldots w_i$, therefore, its conditional probability—which is estimated through its conditional belief—will be high, and its surprisal low, and vice versa. This offline situation surprisal metric is independent of linguistic experience; it is only sensitive to probabilistic world knowledge encoded within the DSS.

Evaluating model behavior

As described above, the main aim for the development of our incremental comprehension model is to show that surprisal theory can explain effects that go beyond linguistic experience, when defined with respect to the unfolding meaning representations in a "comprehension-centric" model of online comprehension. In this section, we evaluate our online surprisal metric by investigating how the model responds to different manipulations of linguistic experience and world knowledge. In particular, we perform three types of simulations, which together provide insight into the sources underlying DSS-derived online surprisal:

(1) Manipulation of linguistic experience only: The model is presented with sentences that differ in terms of their occurrence frequency in the training data, but that keep the microworld probabilities constant (i.e., equal situation surprisal).
(2) Manipulation of world knowledge only: The model is presented with sentences that occur equally frequent in the training data, but they differ with respect to their probabilities within the DSS (i.e., equal linguistic surprisal).
(3) Manipulation of both linguistic experience and world knowledge: To investigate the interplay between linguistic experience and world knowledge, we present the model with sentences in which the linguistic experience and world knowledge are in conflict with each other (i.e., linguistic experience dictates high surprisal, whereas world knowledge dictates low surprisal, and vice versa).

In what follows, we investigate how these manipulations affect online surprisal in our model, by contrasting it with the two offline surprisal metrics presented above.

Manipulation of linguistic experience

Existing computational models focus on word surprisal effects in which the expectancy of a word is computed on the basis of prior linguistic experience of having seen that word in a particular linguistic context (e.g., in terms of the preceding part-of-speech tags: Boston et al., 2008; Demberg & Keller, 2008; Frank, 2009; or in terms of the preceding words: Brouwer et al., 2010; Roark et al., 2009; Smith & Levy, 2008). We therefore begin by examining whether our online surprisal metric reflects manipulations of linguistic experience, as measured by offline linguistic surprisal (see equation (7)), in sentences where only the linguistic context modulates expectancy.

Figure 7 shows the surprisal effects, that is, the difference in surprisal between a target and control condition, for the contrast "NP$_{person}$ ordered *popcorn* [Target]/*dinner* [Control]." In the training data, the sentence "NP$_{person}$ ordered popcorn" is less frequent than the sentence "NP$_{person}$ ordered dinner" (although the unigram frequencies for the words "popcorn" and "dinner" are the same). The unexpectedness of the word "popcorn" in this context is reflected by a positive effect on offline linguistic surprisal for "popcorn" relative to "dinner." On the other hand, there is no preference for *order*(*x*,*popcorn*) over *order*(*x*,*dinner*) in the microworld (see Figure 2). As a result, the offline situation surprisal metric (see equation (8)) shows that the difference between the surprisal values for "popcorn" and "dinner" is negligible (the minuscule bias toward a positive effect is attributable to noise resulting from dimension selection of the DSS). Critically, our online surprisal metric follows the offline linguistic surprisal metric in predicting that "popcorn" is more surprising than "dinner" in this context, demonstrating that it is sensitive to linguistic experience.

Because our online surprisal metric is derived from the DSS situation vectors constructed by the model, the observed effect demonstrates that linguistic experience influences how the model navigates DSS on a word-by-word basis; that is, after processing the word "ordered," the model has navigated to a point in DSS that is closer to the point representing $order(x, dinner)$ than to the point representing $order(x, popcorn)$. Indeed, after processing "beth ordered," the model "understands" $order(beth, dinner)$ (comprehension score = .33) to a larger degree than $order(beth, popcorn)$ ($-$ .32). This preference stems directly from the fact that the model has encountered utterances with the former semantics more often than the latter, as it has seen more sentences describing someone ordering dinner than someone ordering popcorn. Although there is no difference in preference for the two alternatives encoded within world knowledge, the online comprehension model uses its linguistic experience as a guide toward the sentence-final meaning it has seen more frequently during training.

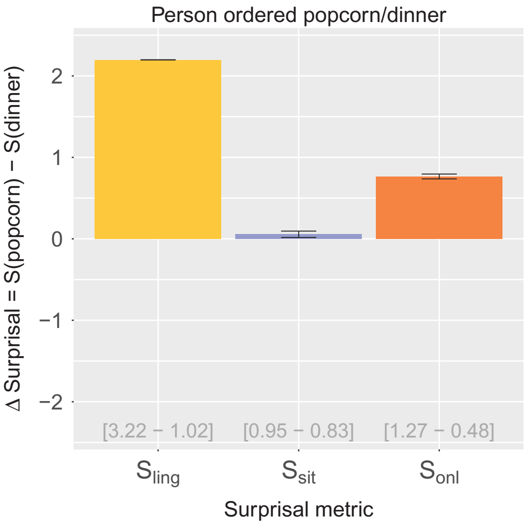


Figure 7. Effects of linguistic experience. Average surprisal differences for the contrast "*NP* $_{person}$ *ordered popcorn* [T]/*dinner* [C]" for the three surprisal metrics: linguistic surprisal (S_{ling}), situation surprisal (S_{sit}) and online surprisal (S_{onl}). Error bars show standard errors ($n = 3$). Individual means are shown in brackets (T − C).

It is perhaps worth emphasizing that this means of capturing surprisal based on linguistic experience is fundamentally different from previous models that have demonstrated that these effects can be modeled using SRNs trained for next word prediction (Frank, Otten, Galli, & Vigliocco, 2015). Those models are best viewed as language models (closer to N-gram models) that learn to estimate the probability of a word in its linguistic context. By contrast, our model uses linguistic experience to prefer moving toward more frequently seen situation vectors, with online surprisal reflecting the extent to which the meaning described by the input word is consistent with the situation vector already constructed.

Manipulation of world knowledge

While experience-based surprisal estimates determined from corpora have proven to be good predictors of processing effort, there is evidence suggesting that human estimates of a word's probability in context—as determined using Cloze procedures—not only differ systematically from corpus-based estimates, but are in fact better predictors of processing effort (Smith & Levy, 2011). Smith and Levy argue that whereas N-grams approximate "true" probabilities, Cloze probabilities reflect "subjective" probabilities. This raises the question of why subjective probabilities are better estimates of processing effort. One explanation for this is that Cloze tasks, which involve participants deciding which word(s) fit best in a provided context, reflect not only the person's linguistic experience but the probabilities of the resultant meaning with respect to their general knowledge—that is, their probabilistic experience with both language *and* the world. In fact, there is ample evidence that the cognitive effort associated with processing a word is crucially dependent on rich discourse-level situation models that are driven by world/script knowledge (e.g., Albrecht & O'Brien, 1993; Camblin et al., 2007; Delogu, Drenhaus, & Crocker, 2018; Hess et al., 1995; Kuperberg et al., 2011; Metusalem et al., 2012; Myers & O'Brien, 1998; O'Brien & Albrecht, 1992; Otten & van Berkum, 2008; van Berkum et al., 2005, 1999, 2003). Several studies conducted by van Berkum and colleagues, for example, have demonstrated how prior discourse context strongly influences which critical word will be expected (or not) in a subsequent sentence. Words that are not supported by the broader situation model cause increased processing effort, as reflected by both an increase in N400 amplitude and reading time, in a manner that no surprisal model driven by linguistic experience alone is able to explain (Otten & van Berkum, 2008; van Berkum et al., 2005). Indeed, Hagoort and colleagues provide evidence that established knowledge about the world, for example, that Dutch trains are yellow, results in increased processing difficulty when comprehenders are presented with the sentence “Dutch trains are *white*” (Hagoort, Hald, Bastiaansen, & Petersson, 2004). It seems unlikely that this is because of hearing or reading about Dutch trains being yellow but rather due to this being a more likely meaning, given one’s experience with the world.

To determine whether our model can account for such findings, we investigate utterances in which only world knowledge affects expectancy. Figure 8 shows the effects on the surprisal metrics for the contrast “NP$_{person}$ entered the cinema and ordered popcorn/dinner” (left panel), and “NP$_{person}$ entered the restaurant and ordered *popcorn*/*dinner*” (right panel). All four sentences occur equally frequently in the training data, which is reflected in the absence of any effect on offline linguistic surprisal. According to world knowledge, however, it is more likely to order popcorn than dinner in the cinema, whereas in the restaurant it is the other way around. These preferences are reflected in offline situation surprisal, which is higher for dinner when ordered in the cinema (left panel), and higher for popcorn when ordered in the restaurant (right panel). In both cases, our online surprisal metric follows offline situation surprisal, showing that it is sensitive to world knowledge.

The influence of world knowledge on our online surprisal metric directly stems from the processing dynamics of the model. As the model navigates the DSS on a word-by-word basis, it has already encountered several choice points before arriving at the point in space after processing the prefinal word “ordered.” The trajectory for all four sentences is the same until the word “cinema”/”restaurant,” after which the model will navigate to different points in DSS. The word “cinema” will navigate the model to a point that renders ordering popcorn more likely than ordering dinner, and the other way around for “restaurant” (see also Figure 4). After processing the prefinal word “ordered,” the model is still at two different points in situation-state space, depending on having processed “cinema” or “restaurant.” As a consequence, the target words “popcorn” and

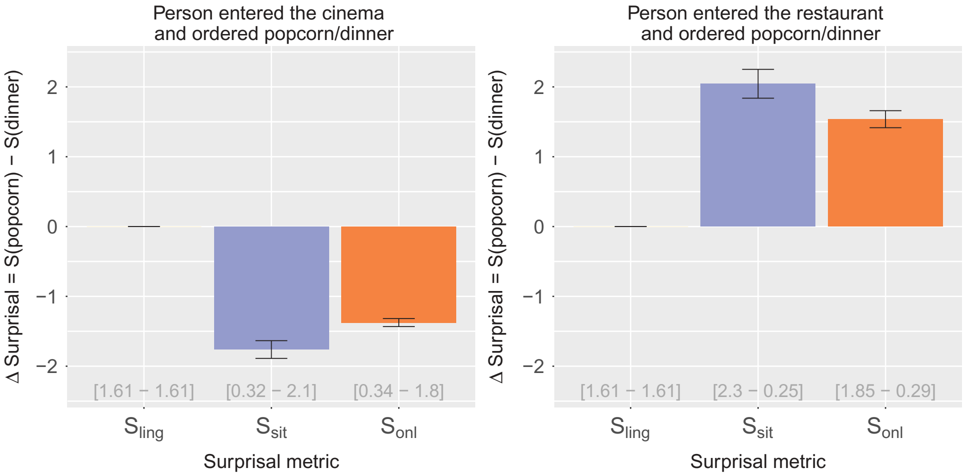


Figure 8. Effects of world knowledge. Average surprisal differences for the contrasts “*NP* $_{person}$ *entered the cinema and ordered popcorn* [T]/*dinner* [C]” (left), and “*NP* $_{person}$ *entered the restaurant and ordered popcorn* [T]/*dinner* [C]” (right) for the three surprisal metrics: linguistic surprisal (S_{ling}), situation surprisal (S_{sit}) and online surprisal (S_{onl}). Error bars show standard errors ($n = 3$). Individual means are shown in brackets (T − C).

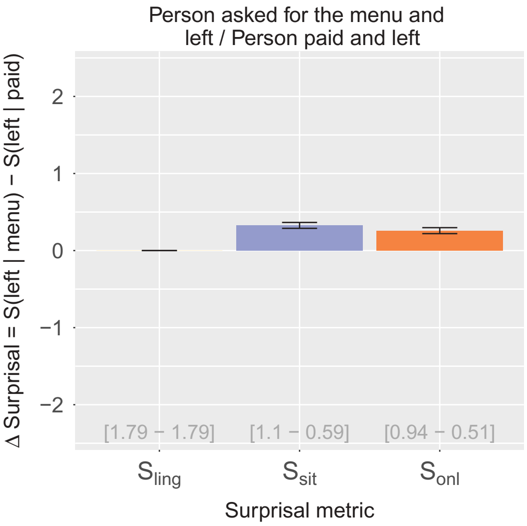


Figure 9. Script-driven effects. Average surprisal differences for the contrast "*NP* $_{person}$ *[asked for the menu and left]* [T]/*[paid and left]* [C]" for the three surprisal metrics: linguistic surprisal (S_{ling}), situation surprisal (S_{sit}), and online surprisal (S_{onl}). Error bars show standard errors ($n = 3$). Individual means are shown in brackets (T − C).

"dinner" will be integrated into different contexts. These different contexts drive the observed effects for offline situation surprisal and online surprisal, which reflect an interaction between context and what is expected.

The above-described effects can be seen as an example of script-driven surprisal: whether ordering popcorn is more surprising than ordering dinner depends on whether the model pursues a "cinema-script" or a "restaurant-script." The influence of script knowledge on surprisal is even more evident, however, when we consider the temporal ordering of events within a script. Figure 9 shows the effects on the surprisal metrics for the contrast "NP_{person} asked for the menu and *left*" versus "NP_{person} paid and *left*." Again, these sentences occur equally frequently in the training data, as is reflected in the absence of any offline linguistic surprisal effect. Now, because of the way in which we sampled temporally dependent events within the observations constituting the DSS, *pay(x)* and *leave(x)* cooccur more frequently than *ask_menu(x)* and *leave(x)*. World knowledge, therefore, dictates that it is more likely to leave after paying than to leave after asking for the menu. This is reflected in both the offline situation surprisal metric, as well as in the online surprisal metric. In our model, the concept of script knowledge is thus inherent to the way in which world knowledge is encoded in the DSS.

Interplay between linguistic experience and world knowledge

In the previous sections, we have shown that our online surprisal metric is independently sensitive to both linguistic experience and world knowledge. Although these surprisal effects stem from different information sources (i.e., from the frequency differences in the model's input and the probability distribution within the DSS representations, respectively), the model nonetheless predicts that these sources interact with each other. This becomes particularly clear in situations in which linguistic experience and world knowledge contradict each other.

Figure 10 shows the effects on the surprisal metrics for the contrasts "NP_{person} ordered *cola/water*" (left), and "NP_{person} ordered *champagne/water*" (right). According to world knowledge, ordering water is more likely than ordering either cola or champagne, and there is no difference in likelihood between ordering cola and champagne (see Figure 2). This is reflected

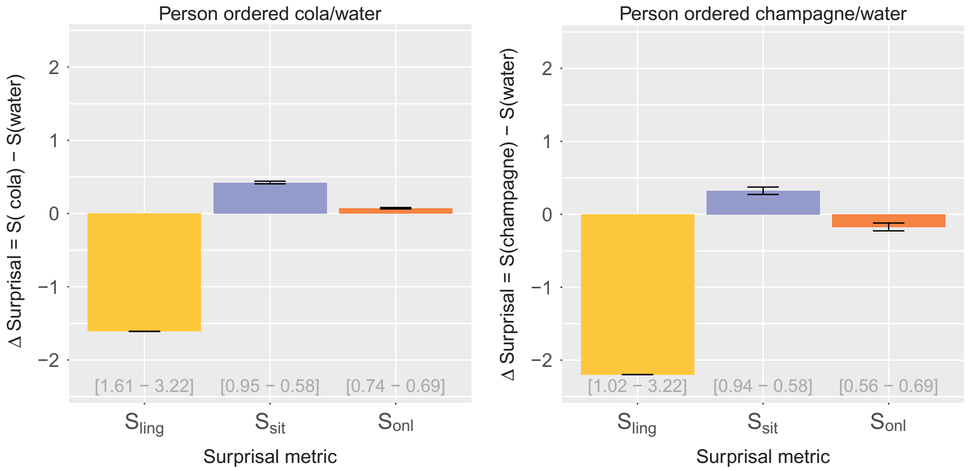


Figure 10. Interplay between linguistic experience and world knowledge. Average surprisal differences for the contrasts "*NP_{person} ordered cola* [T]/*water* [C]" (left), and "*NP_{person} ordered champagne* [T]/*water* [C]" (right) for the three surprisal metrics: linguistic surprisal (S_{ling}), situation surprisal (S_{sit}) and online surprisal (S_{onl}). Error bars show standard errors ($n = 3$). Individual means are shown in brackets (T − C).

in offline situation surprisal, which is higher for cola relative to water, as well as for champagne relative to water (the small difference in situation surprisal between "cola" and "champagne" is again attributable to dimension selection). In the linguistic experience, in turn, the picture is reversed. The model encounters the sentence "NP_{person} ordered *cola*" 5 times more often than "NP_{person} ordered *water*." This is reflected in higher offline linguistic surprisal for latter relative to the former. Our online surprisal metric, however, follows offline situation surprisal in that "cola" is more surprising than "water." Indeed, this effect in online surprisal is relatively small. Crucially, when linguistic experience is strengthened, the effect on online surprisal reverses and follows offline linguistic surprisal. That is, the model encounters the sentence "NP_{person} ordered *champagne*" 9 times more often than "NP_{person} ordered *water*," and here the online metric follows linguistic surprisal in that "water" is more surprising than "champagne." These results show that in integrating linguistic experience and world knowledge in online comprehension, our model balances cues from these different information sources, depending on their relative strengths.

Experimentally, it is extremely difficult to contrast these two types of information sources in a controlled manner. Nonetheless, studies that aim to elucidate the interplay between linguistic experience and world knowledge do exist. Hald, Steenbeek-Planting & Hagoort (2007), for example, contrasted stereotypical and nonstereotypical sentences (e.g., "The city Venice has very many *canals/roundabouts* […]") in contexts that either support or violate world knowledge (e.g., a context describing either gondola tours or traffic flow in Venice). They found that in the world knowledge violation context, the different target words did not show any differential processing difficulty. This result indicates that default expectations about the world and the linguistic signal may be overridden by the situation model that is constructed online as linguistic input comes in. Nieuwland & van Berkum (2006) provide further evidence in this direction. They found that in light of a supporting situation model (e.g., a romance between two peanuts), animacy-violating sentences such as "the peanut was in *love*" led to less processing difficulty than nonviolating sentences such as "the peanut was *salted*," suggesting that the unfolding situation model may override linguistic expectations. Indeed, studies like these offer at least preliminary support for our model's prediction that linguistic experience and world knowledge interact during sentence processing and may in some cases even cancel each other out.

Discussion

We have presented a model of language comprehension that constructs a rich representation of utterance meaning—a situation model—on a word-by-word basis. The probabilistic nature of these representations straightforwardly supports the computation of surprisal in a manner that integrates both probabilities determined by linguistic experience and probabilistic knowledge about the world. Crucially, these probabilistic information sources affect processing independent of each other; that is, effects of linguistic experience stem from the frequency of observing a given situation model representation, whereas probabilistic knowledge about the world is contained within these representations. Below, we discuss the implications of such a view on comprehension and surprisal, and we describe how the predictions made by the model can be empirically tested. Furthermore, we present an evaluation of the meaning representations employed by the model.

Comprehension as situation-state space navigation

We have shown that during online comprehension in the model, the different sources of information are both manifest in a single layer of meaning representation, namely, the situation vectors on the output layer. Comprehension, in our model, involves navigating through situation-state space—each incoming word provides a cue for the model toward the utterance-final meaning representation. The way in which the model navigates through this semantic space is determined by the input it has seen during training (~linguistic experience), as well as by the way in which the semantic space itself is organized (~world knowledge). That is, if the model's linguistic experience indicates that a particular continuation is more likely than another, it will move toward a point in situation-state space that is closer to the meaning of the former continuation than that of the latter. Moreover, the model's prior knowledge about the world may dictate that certain meanings are more related to each other than others, resulting in them being closer in semantic space.

To further elucidate the model's comprehension process, Figure 11 presents a repetition of Figure 4, with the addition of the model's word-by-word output for the sentences *"beth ordered champagne/cola/water"* and *"beth left."* The blue arrows indicate how the model navigates through situation-state space on a word-by-word basis, with labels indicating the online surprisal value for a given transition.[3] This figure illustrates how the model assigns intermediate points in situation-state space to sentence-internal words, and approximates propositional meanings for sentence-final words. In particular, it shows that at the word *"ordered,"* the model navigates to a point in space that is in between the meanings of the propositions reflecting the different orders (cola, water, champagne, dinner, and popcorn) but closer to some of these (closest to champagne), as determined by world knowledge and linguistic experience.

This perspective on comprehension sheds an interesting light on recent insights in the text comprehension literature on the role of validation in the comprehension process. Validation describes the process of evaluating consistency of incoming linguistic information with the previous linguistic context and general knowledge about the world (Singer, 2013). Within the RI-Val model of comprehension, Cook and O'Brien (2014) take validation to be one of the three central processes of comprehension (together with activation/resonance and integration). These processes are assumed to operate in a parallel but asynchronous manner; validation starts only after the process of integration has begun (see also O'Brien & Cook, 2016). Based on a range of empirical findings, Richter (2015) argues that validation is perhaps even more closely interwoven with integration,

[3] These surprisal values are calculated for the depicted word-by-word transitions, and they may differ slightly from the averaged surprisal values shown in Figure 10.

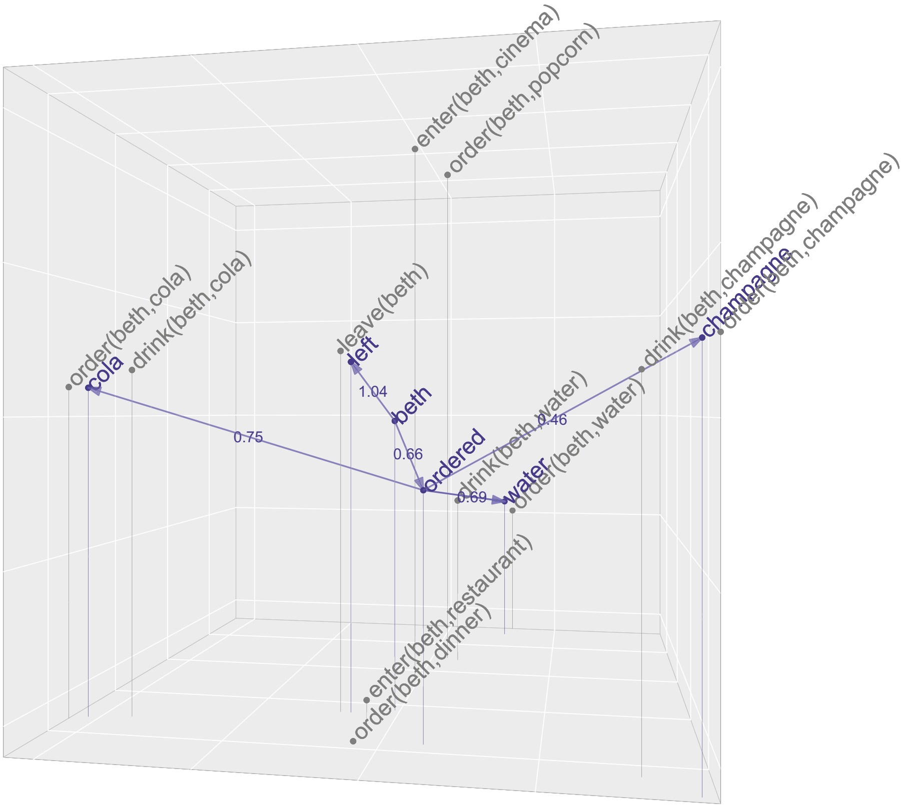


Figure 11. Visualization of the DSS into three dimensions (by means of multidimensional scaling; MDS) for a subset of the atomic propositions (the predicates *enter, order, drink*, and *leave*, applied to *beth*). Highlighted points show the word-by-word output of the model for the sentences *"beth ordered champagne/cola/water"* and *"beth left"*. The arrows reflect the DSS navigation and are labeled with the word-by-word online surprisal values.

because the result of either process depends on the other (see also Isberner & Richter, 2014). The comprehension process implemented in our model is in line with this perspective; because world knowledge is inherently encoded into the meaning representations that the model constructs, it simultaneously performs integration of novel information and validation with respect to world knowledge during comprehension. As such, we believe future simulations with our model may further elucidate the role of validation in the comprehension process.

Surprisal indexes change in situation models

The model presented in this work differs crucially from existing implementations of surprisal, in that it is not the likelihood of the word in context *per se* that determines surprisal, but rather the consequences of that word for the unfolding probabilistic situation model. Instantiating surprisal as a function of interpretation-level state transition has several implications for how we conceptualize the expectedness of individual words. In traditional surprisal models, the surprisal induced by a word derives *directly* from the words that precede it (e.g., Frank et al., 2015; Hale, 2001; Levy, 2008). In the present model, however, words only *indirectly* determine surprisal; that is, words are cues or instructions to meaning-level state-transitions, which in turn determine surprisal. Moreover, the expectedness of a cue/instruction is an *implicit* function of linguistic experience, knowledge about the world, and the current state of interpretation (the current point in DSS). Hence, there is no *explicit* predictive processing taking place in our model—that is, no construction of mental representations that go beyond the current input—rather, the model moves to a state that reflects a weighted average of the meanings the sentence could have. In extreme cases, where meaning and experience heavily conspire toward a small number of continuations, the model will expect specific cues (i.e., words), thus explaining effects of lexical prediction (DeLong, Urbach, & Kutas, 2005; van Berkum et al., 2005). This state-driven notion of implicitly predicting potential upcoming information is consistent with the concepts of *readiness* and *resonance* that are central to the text comprehension (e.g., Gerrig & McKoon, 1998; Gerrig & O'Brien, 2005) and human memory literature (e.g., Anderson, 1990; Ratcliff & McKoon, 1988).

Another important implication of the presented perspective on surprisal involves its mapping to Event-Related Potentials (ERPs). In the electrophysiological domain, surprisal has often been linked to the N400 component (Delaney-Busch, Morgan, Lau, & Kuperberg, 2017; Delogu, Crocker, & Drenhaus, 2017; Frank et al., 2015), the amplitude of which is inversely related to the expectancy of a word; the less expected a word, the higher N400 amplitude (Kutas, Lindamood, & Hillyard, 1984). Although the N400 component was previously taken to be an index of integrative processing (i.e., the updating of an unfolding utterance representation), it has recently been linked to the process of lexical retrieval, which is facilitated by lexical and contextual priming; that is, N400 amplitude is reduced if word-associated conceptual knowledge is preactivated in memory (Brouwer, Fitz, & Hoeks, 2012; Kutas & Federmeier, 2000; Lau, Phillips, & Poeppel, 2008; van Berkum, 2009, 2012). Brouwer et al., (2012) adapt this perspective on the N400 as part of the Retrieval-Integration hypothesis and argue that the process of integrating word meaning with the unfolding utterance representation is instead reflected in P600 amplitude, a late positive deflection of the ERP signal. Under this view, our comprehension-centric formalization of surprisal is predicted to be reflected in the P600; surprisal is a measure of how likely a transition is from one interpretative state to the next, and P600 amplitude is a reflection of the neurophysiological processing involved in this transition (see also Brouwer, 2014; Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Crocker, Knoeferle, & Mayberry, 2010).

Testing the model's predictions

The proposed computational model and the comprehension-centric formalization of surprisal that it instantiates lead to two overarching predictions: first, both linguistic experience and world knowledge contribute to the determination of surprisal, and second, surprisal reflects meaning-level probabilistic state-transitions rather than simply lexical expectations. Testing these predictions critically involves mapping surprisal estimates onto empirical correlates of linguistic processing. In the behavioral domain, surprisal is typically correlated with reading times (Hale, 2001; Levy, 2008). With respect to the electrophysiological domain, we believe the link between surprisal and the P600 as outlined in the previous section offers a particularly promising direction for investigating world knowledge-driven effects on surprisal.

Previous studies into the electrophysiological correlates of surprisal have mainly focused on the N400 component (Delaney-Busch et al., 2017; Frank et al., 2015). Frank and colleagues (2015), for instance, report a reliable correlation between N400 amplitude and surprisal. Critically, however, the Retrieval-Integration account predicts that an increase in N400 amplitude typically cooccurs with an increase in P600 amplitude (Brouwer et al., 2017, 2012; Brouwer & Hoeks, 2013; Brouwer & Crocker, 2017). An increase in P600 amplitude, on the other hand, does not necessarily cooccur with an increase in N400 amplitude (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007; for reviews). A particularly strong test of the predicted link between surprisal and P600 amplitude, therefore, comes from experimental designs in which only a P600 is predicted to occur. Hoeks, Stowe, and Doedens (2004), for instance, found that Dutch sentences such as, "De speer heeft de atleten *geworpen*" (lit: "The javelin has the athletes *thrown*") produce only a P600-effect relative to "De speer werd door de atleten *geworpen*" (lit: "The javelin was by the athletes *thrown*"). Assuming these sentences do indeed result in different surprisal, as suggested by the results of an offline rating task in which the former were rated as significantly more difficult than the latter, a link between surprisal and the P600 would be supported above and beyond any link between surprisal and the N400. That is, although word-induced N400 amplitude *indirectly* reflects surprisal in terms of the degree to which an incoming concept is expected given the unfolding situation model, the P600 *directly* reflects the surprisal incurred by updating the situation model with this incoming information. Hence, we believe that, in addition to reading time studies, ERP paradigms that focus on the P600 as a dependent measure of surprisal allow for the most straightforward testing of the predictions of our model.

Evaluation of the DSS-derived meaning representations

We have shown that the DSS representations can be successfully employed to capture world knowledge-driven inferences, such as script-driven effects. We induced this knowledge by sampling the observations in the DSS in such a way that certain propositions cooccur more often than others, and such that individual observations reflect a temporally extended sequence of propositions. Whereas we have shown that this approach suffices for capturing simple temporal dependencies between events, it also has its limitations, for instance, observations are constrained to only incorporate a single *enter* event to prevent ambiguity. That is, if *enter(beth,bar), leave(beth,bar), enter(beth,restaurant)* and *leave(beth,restaurant)* all hold in a single observation, it is impossible to tell whether Beth first entered the bar or the restaurant using the current scheme. Importantly, however, this is the result of the way in which we chose to encode time in DSS, and not a limitation of the DSS formalism in itself. That is, different representational schemes may be employed to encode a more elaborate notion of temporal dependency. This could be achieved by associating each atomic proposition with an explicit notion of *aspect* (e.g., past, present, future), or by introducing *variables* (cf. neo-Davidsonian event semantics; Davidson, 1967; Parsons, 1990). We aim to explore these different approaches in future work.

Furthermore, the DSS representations employed in the current model are derived using a "microworld" strategy, which means that the world knowledge that can be captured is limited to that available in the confined microworld (cf. Frank et al., 2009, 2003). It must be noted, however, that none of the presented machinery or results hinges upon this choice—it primarily serves the controlled investigation of the model's performance. In future work, we aim to investigate different ways of deriving a DSS from empirical data, for example, by making use of semantically annotated corpora (e.g., the Groningen Meaning Bank; Bos, Basile, Evang, Venhuizen, & Bjerva, 2017), or crowd-sourced data describing world knowledge (see, e.g., Elman & McRae, 2017).

As the human language comprehension system is implemented in the neural hardware of the brain, a central aim of the study of language is to understand how the computational principles and representations underlying language comprehension are implemented in neural hardware. Brouwer, Crocker, and Venhuizen (2017) have recently argued that the DSS model can serve as a framework for such a *neural semantics*; it offers neurally plausible meaning representations that directly reflect experience with the world—in terms of observations over meaning-discerning atoms—and complex meaning can be directly derived from these atoms. Moreover, the meaning representations inherently carry probabilistic information about themselves and their relation to each other. As demonstrated in this article, these representations can be constructed on a word-by-word basis in a computational model of language processing. Hence, the solution that our model offers for integrating linguistic experience with influences of world knowledge on word processing difficulty is one that is *linguistically* plausible (it provides a rich framework to encode both sentence meaning and knowledge about the world), *psychologically* plausible (it explains behavior by indexing surprisal as a consequence of comprehension), as well as *neurally* plausible (meaning is computed and represented using neural hardware).

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG), SFB/CRC 1102 "Information density and linguistic encoding" (Project A1) awarded to MWC.

References

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1061–1070.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi:10.1016/S0010-0277(99)00059-1

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. doi:10.1016/j.brainresrev.2008.05.003

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., & Bjerva, J. (2017). The Groningen meaning bank. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 463–496). Dordrecht, Netherlands: Springer Netherlands.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Brouwer, H. (2014). *The electrophysiology of language comprehension: A neurocomputational model* (doctoral dissertation). Groningen, Netherlands. Retrieved from http://hdl.handle.net/11370/32f172dc-8ee5-42bf-a91f-c2406398c019

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, *8,*1327. doi:10.3389/fpsyg.2017.01327

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*, 1318–1352. doi:10.1111/cogs.12461

Brouwer, H., Crocker, M. W., & Venhuizen, N. J. (2017). Neural semantics. In M. Wieling, M. Kroon, G. van Noord, & G. Bouma (Eds.), *From semantics to dialectometry: Festschrift in honor of John Nerbonne* (vol. 32, pp. 75–83). London, UK: College Publications.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. doi:10.1016/j.brainres.2012.01.055

Brouwer, H., Fitz, H., & Hoeks, J. C. (2010). *Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory*. Proceedings of the 2010 workshop on cognitive modeling and computational linguistics (pp. 72–80). Uppsala, Sweden.

Brouwer, H., & Hoeks, J. C. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, *7*(758), 1–12. doi:10.3389/fnhum.2013.00758

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128. doi:10.1016/j.jml.2006.07.005

Cook, A. E., & Myers, J. L. (2004). Processing discourse roles in scripted narratives: The influences of context and world knowledge. *Journal of Memory and Language*, *50*(3), 268–288. doi:10.1016/j.jml.2003.11.003

Cook, A. E., & O'Brien, E. J. (2014). Knowledge activation, integration, and validation during narrative text comprehension. *Discourse Processes*, *51*(1–2), 26–49. doi:10.1080/0163853X.2013.855107

Crocker, M. W., Knoeferle, P., & Mayberry, M. R. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, *112*(3), 189–201. doi:10.1016/j.bandl.2009.03.004

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The logic of decision and action* (pp. 81–94). Pittsburgh, PA: University of Pittsburgh Press.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. (2017). *Comprehenders rationally adapt semantic predictions to the statistics of the local environment: A Bayesian model of trial by trial N400 amplitudes*. Proceedings of the 39th annual conference of the cognitive science society, London, UK.

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, *161*, 46–59. doi:10.1016/j.cognition.2016.12.017

Delogu, F., Drenhaus, H., & Crocker, M. W. (2018). On the predictability of event boundaries in discourse: An ERP investigation. *Memory & Cognition*, *46*(2), 315–325. doi: 10.3758/s13421-017-0766-4

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. doi:10.1038/nn1504

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi:10.1016/j.cognition.2008.07.008

Elman, J., & McRae, K. (2017). A model of event knowledge. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 337–342). Austin, TX: Cognitive Science Society.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi:10.1207/s15516709cog1402_1

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. Proceedings of the 31st annual conference of the cognitive science society (pp. 1139–1144). Austin, TX.

Frank, S. L., Haselager, W. F., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, *110*(3), 358–379. doi:10.1016/j.cognition.2008.11.013

Frank, S. L., Koppen, M., Noordman, L. G., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, *27*(6), 875–910. doi:10.1207/s15516709cog2706_3

Frank, S. L., Koppen, M., Noordman, L.G.M., & Vonk, W. (2008). World knowledge in computational models of discourse comprehension. *Discourse Processes*, *45*(6), 429–463. doi:10.1080/01638530802069926

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. doi:10.1016/j.bandl.2014.10.006

Frank, S. L., & Vigliocco, G. (2011). Sentence comprehension as mental simulation: An information-theoretic perspective. *Information*, *2*(4), 672–696. doi:10.3390/info2040672

Garrod, S., & Terras, M. (2000). The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. *Journal of Memory and Language*, *42*(4), 526–544. doi:10.1006/jmla.1999.2694

Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, *26*(2–3), 67–86. doi:10.1080/01638539809545039

Gerrig, R. J., & O'Brien, E. J. (2005). The scope of memory-based processing. *Discourse Processes*, *39*(2–3), 225–242. doi:10.1080/0163853X.2005.9651681

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. doi:10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. P. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, Brain* (pp. 95–126). Cambridge, MA: MIT Press.

Golden, R. M., & Rumelhart, D. E. (1993). A parallel distributed processing model of story comprehension and recall. *Discourse Processes*, *16*(3), 203–237. doi:10.1080/01638539309544839

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441. doi:10.1126/science.1095455

Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Research*, *1146*, 210–218. doi:10.1016/j.brainres.2007.02.054

Hale, J. T. (2001). *A probabilistic Earley parser as a psycholinguistic model*. Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics on Language Technologies (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

Hale, J. T. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672. doi:10.1207/s15516709cog0000_64

Hale, J. T. (2011). What a rational parser would do. *Cognitive Science*, *35*(3), 399–443. doi:10.1111/cogs.2011.35.issue-3

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, *124*(1), 62–82. doi:10.1037/0096-3445.124.1.62

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. doi:10.1016/j.cogbrainres.2003.10.022

Isberner, M.-B., & Richter, T. (2014). Does validation during language comprehension depend on an evaluative mindset? *Discourse Processes*, *51*(1–2), 7–25. doi:10.1080/0163853X.2013.855867

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194. doi:10.1207/s15516709cog2002_1

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. doi:10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*(2), 173–202. doi:10.1207/s15516709cog2502_1

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127. doi:10.1016/j.cognition.2004.03.002

Knoeferle, P., Habets, B., Crocker, M. W., & Münte, T. F. (2008). Visual scenes trigger immediate syntactic reanalysis: Evidence from ERPs during situated spoken comprehension. *Cerebral Cortex*, *18*(4), 789–795. doi:10.1093/cercor/bhm121

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. doi:10.1016/j.brainres.2006.12.063

Kuperberg, G. R., Paczynski, M., & Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, *23*(5), 1230–1246. doi:10.1162/jocn.2010.21452

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. doi:10.1016/S1364-6613(00)01560-6

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum & J. Requin (Eds.), *Preparatory states and processes* (pp. 217–237). Hillsdale, NJ: Lawrence Erlbaum.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi:10.1037/0033-295X.104.2.211

Langston, M., & Trabasso, T. (1999). Modeling causal integration and availability of information during comprehension of narrative texts. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 29–69). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data. *Brain and Language*, *120*(3), 271–281. doi:10.1016/j.bandl.2011.09.001

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. doi:10.1038/nrn2532

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi:10.1016/j.cognition.2007.05.006

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312. doi:10.1006/jmla.1997.2543

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. doi:10.1016/j.jml.2012.01.001

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 92–102.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*(2–3), 131–157. doi:10.1080/01638539809545042

Nieuwland, M. S., & van Berkum, J.J.A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111. doi:10.1162/jocn.2006.18.7.1098

O'Brien, E. J., & Albrecht, J. E. (1992). Comprehension strategies in the development of a mental model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 777–784.

O'Brien, E. J., & Cook, A. E. (2016). Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Processes*, *53*(5–6), 326–338. doi:10.1080/0163853X.2015.1123341

Otten, M., & van Berkum, J.J.A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. doi:10.1080/01638530802356463

Parsons, T. (1990). *Events in the semantics of English*. Cambridge, MA: MIT Press.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 285–408. doi:10.1037/0033-295X.95.3.385

Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, *52*(5–6), 337–355. doi:10.1080/0163853X.2015.1025665

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). *Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing*. Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 (pp. 324–333). Stroudsburg, PA: Association for Computational Linguistics.

Rohde, D.L.T. (2002). *A connectionist model of sentence comprehension and production* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. Laberge & S. Samuels (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 211–236). New York, NY: Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. doi:10.1038/323533a0

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance* (vol. 6). Hillsdale, NJ: Erlbaum.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language*, *54*(4), 574–591. doi:10.1016/j.jml.2005.11.003

Singer, M. (2013). Validation in reading comprehension. *Current Directions in Psychological Science*, *22*(5), 361–366. doi:10.1177/0963721413495236

Singer, M., & Doering, J. C. (2014). Exploring individual differences in language validation. *Discourse Processes*, *51*(1–2), 167–188. doi:10.1080/0163853X.2013.855534

Smith, N. J., & Levy, R. (2008). *Optimal processing times in reading: A formal model and empirical investigation*. Proceedings of the 30th annual conference of the Cognitive Science Society (pp. 595–600). Austin, TX: Cognitive Science Society.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between Cloze, corpus, and subjective probabilities in language processing. Proceedings of the 33rd annual conference of the Cognitive Science Society (pp. 1637–1642). Austin, TX: Cognitive Science Society.

St. John, M. F. (1992). The Story Gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*(2), 271–306. doi:10.1207/s15516709cog1602_5

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*(1–2), 217–257. doi:10.1016/0004-3702(90)90008-N

St. John, M. F., & McClelland, J. L. (1992). Parallel constraint satisfaction as a comprehension mechanism. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist Approaches to Natural Language Processing*, (pp. 97–136). Hillsdale, NJ: Lawrence Erlbaum.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. doi:10.1126/science.7777863

van Berkum, J. J. A., Brown, C. M., 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.

van Berkum, J.J.A. (2012). The electrophysiology of discourse and conversation. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 589–614). Cambridge, UK: Cambridge University Press.

van Berkum, J.J.A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671. doi:10.1162/089892999563724

van Berkum, J.J.A. (2009). The "neuropragmatics" of simple utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Basingstoke, UK: Palgrave Macmillan.

van Berkum, J.J.A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research*, *17*(3), 701–718. doi:10.1016/S0926-6410(03)00196-4

van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 165–187). Hillsdale, NJ: Lawrence Erlbaum Associates.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. doi:10.1037/0033-2909.123.2.162

Appendix A

Observation sampling

Given n atomic propositions, there are 2^n distinct microworld observations. To construct an $m \times n$ situation-state space, we want to sample m observations from these 2^n possibilities, such that each observation satisfies the hard world knowledge constraints, and such that the entire set of observations approximately reflects the probabilistic structure of the world. To this end, we employ a nondeterministic, incremental inference-driven sampling algorithm (cf. Frank & Vigliocco, 2011) that employs three-valued logic (1: True, 0: False, 0.5: Undecided). The algorithm starts from a fully unspecified observation (all n atomic proposition states set to 0.5), and then repeats the following procedure until the state of all atomic propositions is decided (i.e., is either 1 or 0):

(1) Pick a random, undecided proposition p_x.
(2) Set proposition p_x to be either the case (1) or not (0), depending on its probability given the observation constructed thus far.
(3) Draw all inferences that must follow from deciding the state of proposition p_x in the current observation:
  (a) Randomly pick the next, undecided proposition p_y;
  (b) Construct two alternative observations: observation o_1 in which proposition p_y is set to be the case (1), and observation o_2 in which it is not (0);
  (c) Check for both observations if they violate any hard world knowledge constraints, and act depending on the outcome:
    - Both o_1 and o_2 are consistent with world knowledge: The state of proposition y is uncertain, and cannot be inferred (the state of p_y remains 0.5);
    - Only o_1 is consistent with world knowledge: Infer p_y to be the case (the state of p_y is set to 1);
    - Only o_2 is consistent with world knowledge: Infer p_y not to be the case (the state of p_y is set to 0);
    - Both o_1 and o_2 violate world knowledge: The state of affairs before trying to infer p_y conflicts with world knowledge (start all over from a fully unspecified observation).
  (d) Repeat from step (a) until all undecided propositions have been tried to infer.
(4) Repeat from step 1 until there are no more undecided propositions.

Appendix B

Dimension selection

For a situation-state space to accurately reflect the probabilistic structure of a microworld, a sufficiently large sample of m observations is required. This yields an $m \times n$ situation-state space and situation vectors of m components. Typically, these vectors are rather large for use in neural network models. Ideally, we would therefore like to reduce the $m \times n$ situation-state space to a more practical $k \times n$ space (where $k < m$), while preserving the information encoded in the $m \times n$ space. To this end, we employ a nondeterministic dimension selection algorithm (rather than a dimension reduction algorithm; cf. Frank et al., 2009):

(1) Compute a vector $\vec{c}_m$ containing the comprehension score $comprehension(a, b)$ for each combination of the n atomic propositions from the $m \times n$ space.
(2) Randomly select k rows from the original $m \times n$ situation-state space.
(3) Reject the $k \times n$ space if any of its columns contains zeros only; start over from step 2.
(4) Compute a vector $\vec{c}_k$ containing the comprehension score $comprehension(a, b)$ for each combination of the n atomic propositions from the $k \times n$ space.
(5) Reject the $k \times$ n space if any perfect comprehension scores (-1 or $+1$) in $\vec{c}_m$ and $\vec{c}_k$ do not match; start over from step 2.
(6) Compute the correlation coefficient $\rho(\vec{c}_k, \vec{c}_m)$; if this is the highest ρ thus far, the current $k \times n$ space is the current best approximation of the $m \times n$ space.
(7) Repeat from step 2 for x epochs, and find the $k \times n$ space with the highest ρ.

For the present model, we sampled 150 observations from a 15000 × 45 DSS, with $\rho(\vec{c}_{150}, \vec{c}_{15000}) = .90$.