ELSEVIER

Cognitive Science 27 (2003) 875–910

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Modeling knowledge-based inferences in story comprehension

Stefan L. Frank [a,b,*], Mathieu Koppen [c], Leo G.M. Noordman [a], Wietske Vonk [b,d]

[a] *Discourse Studies, Tilburg University, Tilburg, The Netherlands*
[b] *Center for Language Studies, University of Nijmegen, P.O. Box 310, 6500 AH Nijmegen, The Netherlands*
[c] *NICI, University of Nijmegen, Nijmegen, The Netherlands*
[d] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*

Received 22 October 2002; received in revised form 16 July 2003; accepted 29 July 2003

Abstract

A computational model of inference during story comprehension is presented, in which story situations are represented distributively as points in a high-dimensional "situation-state space." This state space organizes itself on the basis of a constructed microworld description. From the same description, causal/temporal world knowledge is extracted. The distributed representation of story situations is more flexible than Golden and Rumelhart's [Discourse Proc 16 (1993) 203] localist representation.

A story taking place in the microworld corresponds to a trajectory through situation-state space. During the inference process, world knowledge is applied to the story trajectory. This results in an adjusted trajectory, reflecting the inference of propositions that are likely to be the case. Although inferences do not result from a search for coherence, they do cause story coherence to increase. The results of simulations correspond to empirical data concerning inference, reading time, and depth of processing.

An extension of the model for simulating story retention shows how coherence is preserved during retention without controlling the retention process. Simulation results correspond to empirical data concerning story recall and intrusion.

Keywords: Inferencing; Story comprehension; Retention; Distributed representations; Computational modeling; Self-Organizing Maps

* Corresponding author. Tel.: +31-24-3521321; fax: +31-24-3521213.
E-mail address: stefan.frank@mpi.nl (S.L. Frank).

0364-0213/$ – see front matter © 2003 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/j.cogsci.2003.07.002

1. Introduction

A narrative text rarely states explicitly all that is the case in the story events being described. Most facts are left implicit and many can be inferred from the text. Possible inferences range from finding the correct referent of a pronoun to inferring details of the state of affairs at any moment in the story. Only few of these inferences are actually made during reading. There has been considerable debate on which inferences are made on-line (for an overview, see, e.g., Graesser, Singer, & Trabasso, 1994). In general, the inferences that are most easily made on-line are the ones that are most important to the reader's goals, require knowledge that is easily available, and contribute to the coherence of the text (McKoon & Ratcliff, 1992; Noordman, Vonk, & Kempff, 1992; Vonk & Noordman, 1990; for an overview, see Garrod & Sanford, 1994; Singer, 1994; Van den Broek, 1994).

When the reader's goal is to comprehend the story, the causes of the story's events often need to be inferred. For instance, from the short story "Bob was riding his bicycle. He hit the coffee table.", it might be inferred that Bob was riding his bicycle *indoors*, which explains the fact that he could hit a coffee table. Since being outdoors is inconsistent with hitting a coffee table, adding the inference increases the story's coherence. The inference requires the common knowledge that tables are usually found inside houses and that Bob had to be at the same place as the coffee table in order to hit it. How is this specific information selected from the large amount of knowledge about riding bicycles, coffee tables, and hitting? And how does the relevant knowledge update the text representation? Here, we present a computational model that simulates these processes.

Inference processes are only one aspect of text comprehension. Text comprehension consists of multiple processes and representations, ranging from the perception of letters and words to the comprehension of the meaning of the text and its integration with world knowledge. Theories of text comprehension generally distinguish three levels of representation that are constructed during comprehension (Kintsch & Van Dijk, 1978; Van Dijk & Kintsch, 1983) and knowledge-based inferences contribute to the highest of these. The first level is the surface representation, consisting of the text's literal wording. This gives rise to the second level, which is a network of connected propositions called the textbase. In the textbase, two propositions are connected if they share an argument. If a proposition is read for which no argument sharing proposition can be found, inferencing is necessary. Kintsch and Van Dijk do not model how these inferences come about, but they do note that

> most of the inferences that occur during comprehension probably derive from the organization of the text base into facts that are matched up with knowledge frames stored in long-term memory, thus providing information missing in the text base by a process of pattern completion. (Kintsch & Van Dijk, 1978, p. 391)

These "facts" refer to the reader's "personal interpretation of the text that is related to other information held in long-term memory" (Kintsch, 1998, p. 49). This so-called *situation model* (Kintsch, 1998; Van Dijk & Kintsch, 1983) forms the third level of text representation, which is where most knowledge-based inferences are represented.

Existing computational models of story comprehension mainly focus on the construction of a propositional representation, and rarely deal with inferences. These models are more concerned

with text propositions than with the reader's knowledge. The aim and the scope of the present model differ from most of the existing models in two respects. First, the model does not deal with a propositional representation of the story. It does not make use of textual information such as argument overlap of propositions or connectives. Stories are not represented at the textbase level. Second, and as counterpart of the first point, the model deals with story comprehension as a process that invokes knowledge. Of course, this process is based on the information in the text, but it goes beyond the text and encompasses knowledge-based inferences.

Most computational models have a different aim and scope. The Resonance model (Myers & O'Brien, 1998), for instance, simulates the fluctuating activation of story statements during reading. These activations lead to a memory representation of the story, as described by the Landscape model (Van den Broek, Risden, Fletcher, & Thurlow, 1996; Van den Broek, Young, Tzeng, & Linderholm, 1999). In the story representation constructed by these models, propositions are related either by argument overlap (Resonance) or by co-occurrence in working memory (Landscape), not by their relation in the reader's knowledge base. Therefore, stories are represented at a textbase level. Propositions that are not in the story but originate from the reader's world knowledge may be added to the story representation, but since this is done by the modeler on an *ad hoc* basis, knowledge-based inferencing is not simulated. Likewise, the model presented by Langston and Trabasso (1999; Langston, Trabasso, & Magliano, 1999) receives information about causal relations as input instead of simulating its inference. In contrast, we present a model in which stories are represented at a situational level and inferences result from the application of world knowledge to the story representation.

A major problem in modeling inferencing is the implementation of the large amount of world knowledge needed for comprehension. For the Construction-Integration model (Kintsch, 1988, 1998), this problem arises in the first of two phases that are assumed to make up text comprehension. During this so-called construction phase, the text activates a small number of propositions from the reader's world-knowledge net. By only implementing this small part of the reader's knowledge, the number of propositions in the model remains tractable. However, it is not well-defined which propositions are to be selected during the construction phase.

For instance, Schmalhofer, McDaniel, and Keefe (2002) used the Construction-Integration model to explain how bridging inferences are made. They let the model process two different texts, both starting with the sentence (1) *The director and the cameraman were preparing to shoot closeups of the actress on the edge of the roof of the 14 story building when suddenly the actress fell*. The next sentence was either (2a) *Her orphaned daughters sued the director and the studio for negligence* or (2b) *The director was talking to the cameraman and did not see what happened*. From (2a), it can be inferred that the actress died, but from (2b) it cannot. Indeed, the simulations resulted in strong activation of the proposition "the actress is dead" after processing of (2a), but not after (2b). However, for this activation to be possible, the proposition has to be part of the text representation even though it is not part of the text. Therefore, it was added during the Construction phase of sentence (1). No other proposition was added. Of course, this was because the modeler knew that one of the sentences (2a) and (2b) implies that the actress died. But what if the next sentence had turned out to be *She was released from hospital after two weeks*? In this case, the proposition "the actress is wounded" should have been added to the text representation in the Construction phase of sentence (1). And how about *The stunt coordinator was very pleased with her practice jump*? In short, any possible outcome of sentence (1) needs

to be selected during the Construction phase to make its inference possible. It is up to the modeler to choose at least the propositions that are expected to be inferred later. As a result, the process of selecting relevant world knowledge is not part of the computational model. It is exactly this problem that shall be tackled here in a computational manner.

Other well-known models solve the world-knowledge problem by reducing the size of the world in which the stories take place, allowing the model to be purely computational. For instance, in the Story Gestalt model (StJohn, 1992; StJohn & McClelland, 1992) a recurrent neural network is trained to answer questions about the stories it processes. During this training phase, the network develops distributed representations of the stories and obtains knowledge about the story world. In a later test phase, it answers questions about novel stories, using regularities in the stories on which it was trained. This process involves the making of knowledge-based inferences.

The model presented here is comparable to the Story Gestalt model in the sense that stories take place in a simplified world and are represented distributively. The major difference between the two models is that the Story Gestalt model lacks a notion of story time. The network's activation patterns represent the story events that occurred but not their temporal order. For sufficient story comprehension, however, this order is crucial.

Golden and Rumelhart (1993; Golden, Rumelhart, Strickland, & Ting, 1994) proposed a model in which the order of events described by the story text is represented explicitly. The architecture of our model is based on Golden and Rumelhart's. The main difference between the two is that propositions are represented locally in the Golden and Rumelhart model, while the current model represents them distributively in a high-dimensional situation space. For this reason, it is called the Distributed Situation Space (DSS) model.

Because of its similarity to the DSS model, the next section will explain the Golden and Rumelhart model in enough detail to understand its architecture. Following this, Section 3 describes the world knowledge that was used in our simulations. The DSS model is presented in Section 4. Also, it is shown how it can be extended to simulate story retention. Section 5 presents results of simulations and corresponding empirical data. The final section discusses implications for theories of on-line comprehension and makes suggestions for improvements to the model.

2. The Golden and Rumelhart model

Inferencing in story comprehension requires a representation of the story, knowledge about the world in which the story takes place, and a process that applies this world knowledge to the story representation. Here we describe how these three aspects are implemented in the Golden and Rumelhart model and note some limitations, which are overcome by the DSS model.

2.1. Representing a story

Golden and Rumelhart view a story as a temporal sequence of story situations. In their model, the order in which story situations occur is represented explicitly by associating to every situation a "time step" index t. The situation at time step $t - 1$ occurs before (and is a

possible cause of) the situation at t. Likewise, the situation at $t + 1$ is a possible consequence of the situation at t.

A story situation is a set of propositions that occur at one moment in the story. Let p_t denote a proposition p at story time step t. With each such p_t is associated a value $x_{p,t}$ between 0 and 1, denoting the (subjective) probability of p_t. This value represents the reader's belief that proposition p is the case at time step t. The collection of all values in the situation at story time step t is denoted by a vector $X_t = (x_{p,t}, x_{q,t}, \dots)$, with one element for each proposition. If d different propositions are needed to comprehend a story, every situation vector contains d elements. Situation vectors X_t can therefore be viewed as points in the d-dimensional unit cube $[0, 1]^d$ called the *situation-state space*. A story is a sequence of such points, or a *trajectory* in this space.

2.2. Story world knowledge

If a reader knows that proposition p can cause proposition q in the story world, then the combination of p_{t-1} (proposition p is the case at time step $t - 1$) and q_t (q is the case at the following time step t) is a plausible sequence of events. Consequently, the reader's belief in the occurrence of one of the two events can increase belief in the other. Such causal story world knowledge (or "world knowledge" for short) is implemented by assigning a value w_{pq} to each pair of propositions (p, q). A positive value of w_{pq} indicates that belief in either p_{t-1} or q_t increases belief in the other. A negative value of w_{pq} indicates the opposite: belief in p_{t-1} or q_t decreases belief in the other. If w_{pq} equals 0, no causal relation between p_{t-1} and q_t is known to exist. The values w_{pq} for all propositions p and q constitute a $d \times d$ matrix W, the world-knowledge matrix.[1]

The implementation of this general world knowledge is based on four simplifying assumptions:

1. *Single propositions*. What is modeled is how belief in a single proposition influences belief in another single proposition. The influence between beliefs in situations (i.e., combinations of propositions) only emerges as the result of these influences between pairs of propositions from the situations.
2. *Consistency over time*. Causal knowledge does not depend on the moment in the story. Although the belief in propositions fluctuates during story time, the way beliefs in propositions influence each other are "laws of nature" that remain constant.
3. *Range of influence*. Beliefs in propositions at story time step t are influenced only by beliefs regarding the neighboring time steps $t - 1$ and $t + 1$. Propositions at other time steps can only have an indirect influence if they leave an effect on the propositions at $t - 1$ or $t + 1$.
4. *Symmetry*. The influence belief in p_{t-1} has on belief in q_t is the same as the influence in the opposite direction (of q_t on p_{t-1}): Both influences are represented by the value w_{pq}. This is not in contradiction with causality being directed in time, since w_{pq} does not deal with p causing q, but with p_{t-1} and q_t causing belief in each other. For example, a stomach ache does not cause having eaten too much, but observing a stomach ache does cause us to believe that too much was eaten before.

2.3. Model processing

As input, the model receives the initial story trajectory in which $x_{p,t} = 1$ if the story text states that proposition p is the case at time step t, and $x_{p,t} = 0$ if it does not. This initial trajectory already includes *all* the story time steps, so statements do not enter the model one by one.

Propositions stated in the text stand for given facts that cannot be denied, so their values (equal to 1) do not change. The story comprehension process comes down to updating all other values, that is, for each p_t that was not stated explicitly the probability that it is the case is computed. As derived in Appendix A, this probability depends on the other values in the trajectory and on the world-knowledge matrix W. Of course, the probabilities of all propositions at all time steps need to be estimated simultaneously. Since changing a single value will generally change the probability of many other propositions, the values are not set in a single sweep through the trajectory but are iteratively adjusted until they no longer change.

The trajectory after convergence of this process is the interpretation of the story. In this trajectory, a large value of $x_{p,t}$ indicates that proposition p is inferred to be the case at story time step t (unless it was stated in the original text).

2.4. Limitations

The model's architecture can be shown to seriously limit the world knowledge and stories that can be represented. Three main limitations are:

1. *Constraints within a time step*. One of the basic assumptions concerning the implementation of world knowledge is that values at time step t are influenced only by those at $t-1$ and $t+1$. However, it may be necessary to impose constraints on propositions *within* a time step. For instance, a story character might have reached a road junction at time step $t-1$, which will cause her to make a left or right turn at t. She cannot turn both left and right, which is a constraint within t.
2. *Disjunction*. A story statement that is a conjunction of two propositions, like "it is raining and cold," is represented by setting both $x_{\text{rain},t} = 1$ and $x_{\text{cold},t} = 1$. For disjunctions, however, this is not possible. A statement like "the butler *or* the mysterious stranger committed the crime" can only be represented as a single proposition, in which case the "or" is no longer an operator that combines two propositions.[2]
3. *Combined effect of propositions*. Occasionally, the consequences of a conjunction can be quite different from those of the individual propositions that make up the conjunction. For instance, taking *either* medicine A or B might cure a disease, but taking *both* at the same time can make things worse. Minsky and Papert (1969) showed that an architecture like Golden and Rumelhart's cannot compute the so-called exclusive-or function needed to implement such a causal relation because this requires the world-knowledge values $w_{\text{take\_A,cured}}$ and $w_{\text{take\_B,cured}}$ to be positive but their sum to be negative.[3]

As shown in Section 4.1, the DSS model solves these three problems by representing stories differently. In order to explain this alternative representation, it is convenient to first describe the world knowledge that is implemented. This is done in the following section.

3. Constructing a microworld

Understanding even the simplest story requires large amounts of knowledge about the world in which the story takes place. It is, however, impossible to implement any realistic amount of this world knowledge in a story comprehension model. The solution presented here is similar to the one proposed by StJohn (1992; StJohn & McClelland, 1992): Instead of limiting the amount of knowledge, the world itself is limited. The result is a microworld, all knowledge of which is incorporated in the model. Although the microworld allows only for rather simple and not particularly interesting stories, it is complex enough to evaluate the model's properties.

We begin by choosing a small number of basic propositions from which every microworld situation is built up. In our microworld there exist two story characters, who are named "Bob" and "Jilly." Their possible activities and states can be described using the 14 basic propositions shown in Table 1. These are not unrelated within a time step but put constraints on one another. For instance, two hard constraints are that Bob and Jilly can only play soccer when they are outside and can only play a computer game when inside (which is defined as not-outside). Other important constraints are that Bob and Jilly can only perform one activity at a time and that it is only possible for someone to win when they play soccer, hide-and-seek, or both play a computer game. It goes without saying that no proposition can be the case at the same time as its negation. There also exist soft constraints. For instance, Bob and Jilly are more likely to be at the same place and do the same thing than to be at different places and do different things.

All knowledge about constraints between (combinations of) propositions within a time step is considered non-temporal world knowledge. *Temporal* world knowledge, on the other hand, is concerned with contingencies between (combinations of) propositions at adjacent time steps. Here too, there are hard and soft constraints. Two hard constraints are that Bob and Jilly stop the game they are playing after one of them wins and that a game can only be won if it was played in the previous time step. An important soft constraint is that whoever is tired is less

Table 1
Fourteen basic microworld propositions and their intended meanings

No.	Name	Meaning
1	Sun	The sun shines.
2	Rain	It rains.
3	B outside	Bob is outside.
4	J outside	Jilly is outside.
5	Soccer	Bob and Jilly play soccer.
6	Hide-and-seek	Bob and Jilly play hide-and-seek.
7	B computer	Bob plays a computer game.
8	J computer	Jilly plays a computer game.
9	B dog	Bob plays with the dog.
10	J dog	Jilly plays with the dog.
11	B tired	Bob is tired.
12	J tired	Jilly is tired.
13	B wins	Bob wins.
14	J wins	Jilly wins.

likely to win at the next time step. Also, Bob and Jilly are more likely to stay where they are than to change place unless, of course, the weather changes.

The regularities that hold in our microworld will not be implemented in the model directly. Rather, they are used to construct a realistic sequence of situations from which the world knowledge needed by the model is extracted, as explained in the next section. Based on the temporal relations and the non-temporal constraints, a microworld description of 250 consecutive example situations was constructed. In all of these, each basic proposition is stated to be either the case or not the case. For instance, the 14th example situation states that Bob and Jilly are playing soccer outside, that the sun does not shine and it does not rain, and that nobody is tired or wins. The following, 15th example situation is identical except that Bob became tired, which is why Jilly wins in example situation number 16.

4. The Distributed Situation Space model

The three limitations of the Golden and Rumelhart model discussed in Section 2.4 can be overcome by changing the representation of propositions and situations. Every dimension of Golden and Rumelhart's situation space corresponds to exactly one proposition, so propositions are represented locally in this space. The DSS model, on the other hand, uses a distributed representation. As in the Golden and Rumelhart model, propositions in the DSS model are represented by vectors in a high-dimensional situation space. However, there is no one-to-one correspondence between propositions and dimensions of the distributed situation space.

Section 4.1 explains how propositions and story situations are represented distributively, and vice versa: how a DSS vector can be interpreted in terms of belief values of propositions. Following this, Section 4.2 discusses the distributed representation of temporal world knowledge. It explains how world knowledge affects belief values, and how this leads to measures for temporal coherence of a story and for a proposition's "fit" in a story.

The model's task is to infer which propositions are likely to be the case, given the constraints put by both the story and world knowledge. As described in Section 4.3, this process is implemented as a form of pattern completion. The DSS vectors corresponding to the successive story time steps enter the model one by one and are adjusted according to patterns of events known to occur in the world. This results in the increase of belief values of some propositions, reflecting the extent to which they are inferred. Although the inference process does not depend on story coherence, increased coherence may emerge as a result. The section concludes by showing how DSS provides a natural way of modeling story retention.

4.1. Representing a story

Several researchers have suggested distributed representations of propositions. In his Predication model, Kintsch (2000, 2001) proposes a representation in which predicate and argument vectors taken from the LSA model (Landauer & Dumais, 1997) are combined into proposition vectors in such a way that semantically related propositions have similar vectors. Likewise, in StJohn and McClelland's (1990) Sentence Gestalt model a recursive neural network is trained to develop vector representations for simple sentences. However, neither of these represen-

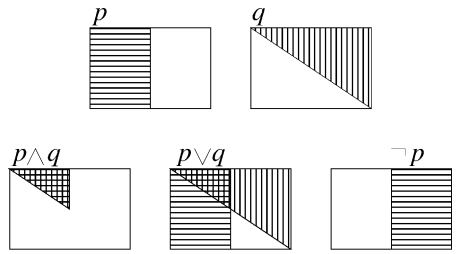

Fig. 1. Dependencies between two propositions (p and q) represented as two-dimensional areas. Proposition p corresponds to the horizontally hatched area and q corresponds to the vertically hatched area. Propositions can be combined by means of conjunction (bottom left), disjunction (bottom center), or negation (bottom right). The fraction of the map covered by a proposition or combination thereof equals its probability of occurrence in the story world.

tations for propositions is suitable for the current model since they cannot be related to the propositions' subjective probabilities. Such a relation is required to interpret the representation in terms of belief values.

In this section, we shall describe a distributed vector representation from which it is possible to directly compute the subjective probability that a proposition is the case given the story situation. Such a subjective probability is called a belief value since it indicates to what extent the proposition is believed to be the case in the situation. In this representation, propositions can be combined using the Boolean operators of negation, conjunction and disjunction, while preserving the relation between their representations and belief values.

First, in Section 4.1.1, a representation is presented in which propositions correspond to areas in two-dimensional space. From this, the representation for negations, conjunctions, and disjunctions of propositions follows naturally and the limitations of the Golden and Rumelhart model mentioned before are overcome. Next, Section 4.1.2 explains how such a representation can be extracted automatically from the description of microworld events developed in Section 3, and that this representation is equivalent to a representation as points in high-dimensional space. Finally, Section 4.1.3 shows how belief values can be computed from such a distributed representation.

4.1.1. Representing propositions and situations

Suppose there is a story world consisting of only two propositions, p and q, each of them being the case half of the time. That is, stated as probabilities, $\Pr(p) = \Pr(q) = .5$. Also suppose that in this story world p and q exclude each other to some extent, causing their conjunction to have a probability of only $\Pr(p \wedge q) = .125$ (compared to $\Pr(p \wedge q) = \Pr(p)\Pr(q) = .25$ that would result if the propositions were independent).

Fig. 1 shows how this story world can be represented by a assigning to p and q particular areas within a rectangle that confines the space of all possibilities. To each proposition is assigned an area that occupies half of the total space, reflecting the .5 *a priori* probability of both propositions. For clarity, this is shown for the two propositions separately in the top row of Fig. 1.

A story situation is a (partial) description of events at one moment in the story. The areas of p and q have been assigned in such a way that any story situation can be represented. The

two areas have an overlap occupying 1/8 of the space, reflecting that $\Pr(p \wedge q) = .125$ (Fig. 1, bottom left). This area represents the situations in which both p and q occur.

Not only the conjunction, but all Boolean operators on propositions p and q are represented faithfully by areas. For instance, $\Pr(p \vee q) = \Pr(p) + \Pr(q) - \Pr(p \wedge q) = .875$ is the size of the area occupied by at least one of p and q (Fig. 1, bottom center) and $\Pr(\neg p) = 1 - \Pr(p) = .5$ is the size of the area not occupied by p (Fig. 1, bottom right).

Note that a story situation contains more specific information if it is the conjunction of more (negations of) propositions. Consequently, in this representation the more information is available, the smaller the corresponding area becomes. Note also that this representation does not allow for a distinction between propositions and situations: They have the same status as areas of a certain extent. This inability to distinguish propositions from situations is in accordance with our claim that DSS represents stories at Van Dijk and Kintsch's (1983) situational level. Such a representation is similar to the result of experiencing the story events (Fletcher, 1994). Unlike their textual descriptions, experiences are not considered a combination of separate propositions.

The three limitations of the Golden and Rumelhart model mentioned in Section 2.4 are overcome by representing propositions and situations in this way. First, constraints between propositions within the same story time step are implemented in their representations. Second, it is now possible to represent not only conjunctions but also disjunctions. Third, since the representation of $p \wedge q$ is not the sum of the representations of p and q separately, knowledge about the causal effects of the conjunction can be qualitatively different from the combined knowledge about the individual propositions.

4.1.2. Self-Organizing Maps

For any realistic amount of propositions, it is impossible to construct by hand a map such that the projections of all propositions on this map correspond to their interdependencies. Fortunately, this can be done automatically by means of a Self-Organizing Map (SOM), also known as Kohonen Map (Kohonen, 1995). Such a map is a grid of cells that organizes itself to map propositions as described above.

For each proposition p, each cell i has a unique membership value $\mu_i(p)$ between 0 and 1, indicating the extent to which the cell belongs to that proposition's area. As explained in detail in Appendix B.1, these values are obtained by training on the 250 example situations from the microworld description developed in Section 3. During training, the membership values are adapted until they reflect the non-temporal constraints among propositions, while the temporal contingencies between consecutive situations are ignored. If a perfect mapping is not possible, the SOM makes an approximation. While these representations of propositions are important to the psychological model, the self-organizing process by which they are obtained is not considered part of the model. We do not claim that this is how actual mental representations of propositions develop.

Fig. 2 shows the resulting map for each basic proposition of our microworld. Moreover, the mappings of two combinations of propositions are given as an example. A SOM does not need to be two-dimensional, but this is convenient for visualization. Also, the exact size and form of the map (in our case, 10×15 hexagonal cells) do not have a large effect on the quality of the representations.

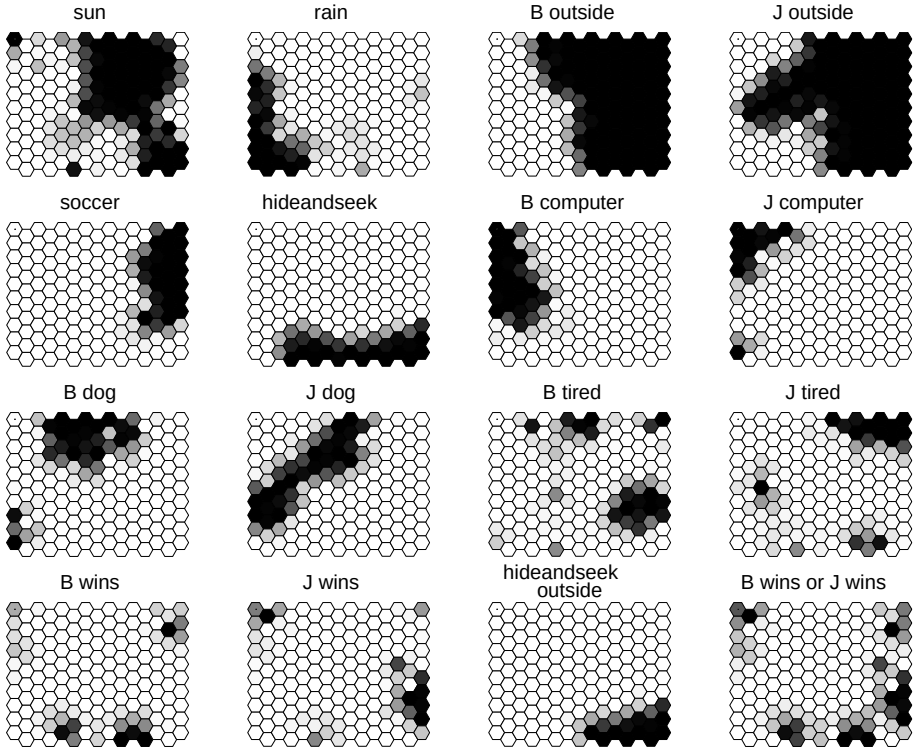

Fig. 2. Automatically constructed mappings of propositions on a Self-Organizing Map with $n = 10 \times 15$ cells. The darkness of a cell indicates its membership value for the corresponding proposition. The last two mappings of the bottom row are examples of combined propositions, representing "Hide-and-seek $\wedge$ B outside $\wedge$ J outside" (Bob and Jilly play hide-and-seek outside), and "B wins $\vee$ J wins" (Bob or Jilly wins), respectively.

The $n = 150$ cells of the SOM form a two-dimensional grid, but can also be viewed as dimensions of an n-dimensional state space $[0, 1]^n$. Any *area* on the SOM, defined by membership values $\mu_i(p)$ for all cells i, corresponds to a *point* in this distributed situation space, defined by the vector $\mu(p) = (\mu_1(p), \mu_2(p), \ldots, \mu_n(p))$. It must be kept in mind, however, that the difference between the SOM and DSS representations is purely aesthetic. The DSS vectors are used in mathematical formulas, while the SOM areas are useful for visualization purposes.

Since a proposition's area on a SOM is fuzzy instead of sharply defined, we need to resort to fuzzy set theory to define the areas corresponding to negations, conjunctions, and other complex propositions. A cell's membership values for "not p" and for "p and q" are computed as follows:[4]

$$\mu_i(\neg p) = 1 - \mu_i(p), \quad \mu_i(p \wedge q) = \mu_i(p)\mu_i(q). \tag{1}$$

It is a well-known fact that all connectives in propositional logic can be defined in terms of negation and conjunction, so any story situation can be represented as a DSS situation vector using the mappings from Fig. 2 and the rules for combining them in Eq. (1). For instance, the

membership values for the disjunction "p or q" follow from the De Morgan law: $\mu_i(p \vee q) = \mu_i(\neg(\neg p \wedge \neg q)) = \mu_i(p) + \mu_i(q) - \mu_i(p)\mu_i(q)$. Likewise, the statement "either p or q" can be represented by defining the exclusive-or operator (XOR) as: p XOR $q \equiv (p \vee q) \wedge \neg(p \wedge q)$.

A story is a sequence of situation vectors, that is, a trajectory through situation space. If $X_t \in [0, 1]^n$ is the situation vector at time step t, the trajectory of a story consisting of T situations is the T-tuple $\bar{X} = \langle X_1, X_2, \ldots, X_T \rangle$. The model takes this trajectory as input and, during the inference process, converts it to a more informative trajectory. How the resulting trajectory can be interpreted is explained next.

4.1.3. Belief values

We now know how to represent any story situation as a vector in DSS. In order to interpret the trajectory that results from the inference process it will also be necessary to take the opposite route: given some vector, reconstruct the situation. This is not generally possible, since only few points in DSS correspond exactly to some combination of propositions. We can, however, compute the belief value of any proposition given a DSS vector.

Let $X = (x_1, x_2, \ldots, x_n)$ be a situation vector (or, equivalently, a SOM area), with n the number of situation-space dimensions. As an "abuse of notation," the symbol X will also be used to refer to the situation represented by the vector X. As a result of training the SOM, the subjective unconditional probability that situation X occurs in the microworld equals the fraction of the map that it covers. This value, denoted $\tau(X)$, is the belief value of situation X and equals

$$\tau(X) = \frac{1}{n}\sum_i x_i. \tag{2}$$

Now suppose we want to compute the subjective probability of some proposition p given that situation X is the case (in fact, p itself can be a combination of propositions). This is the belief value of p in situation X, denoted $\tau(p|X)$. From the fact that $\Pr(p|X) = \Pr(p \wedge X)/\Pr(X)$ and Eqs. (1) and (2), it follows that[5]

$$\tau(p|X) = \frac{\sum_i \mu_i(p) x_i}{\sum_i x_i}. \tag{3}$$

4.2. Temporal story world knowledge

Apart from the important difference in story representation, the DSS model and the Golden and Rumelhart model have identical architectures. Be reminded that world-knowledge implementation in the Golden and Rumelhart model was based on four simplifying assumptions (see Section 2.2). Of these, only the assumption that world knowledge is implemented as influences between propositions does not apply to the DSS model. Instead, it is assumed that world knowledge concerns influences between "dimensions" or "SOM cells."

The model can mathematically be considered a Markov random field (MRF) (see Appendix A). From this theory and the four assumptions, it follows that temporal world knowledge can be implemented using one $n \times n$ matrix of parameters W. As explained in detail in Appendix B.2, this matrix is based on the temporal contingencies between consecutive situations of the mi-

croworld description after converting them into the distributed representation developed in Section 4.1.

The world-knowledge matrix can be used to find a situation that is likely to occur at time step t if the situations at $t-1$ and $t+1$ are known. This is done by calculating, for each SOM cell i at time step t, the expected value of $x_{i,t}$, given the situation vectors X_{t-1} and X_{t+1}. As proven in Appendix A, this expected value equals

$$E_{i,t} = \sigma(X_{t-1}W_{\cdot i} + W_{i\cdot}X'_{t+1}). \tag{4}$$

Here, σ is the sigmoidal function defined as $\sigma(z) = (1 - \mathrm{e}^{-z})^{-1} - (1/z)$, with $\sigma(0) = 1/2$. $W_{\cdot i}$ is the ith column of W and $W_{i\cdot}$ is its ith row. The column vector X'_{t+1} is the transpose of the row vector X_{t+1}. In case there is no previous or next situation, so $t = 1$ or $t = T$, it is defined that $X_0 = \vec{0}$ or $X_{T+1} = \vec{0}$, respectively. The *expected situation vector* at time step t is formed by the collection of expected values: $E_t = (E_{1,t}, \ldots, E_{n,t})$.

Eq. (4) is crucial for the inference process described in Section 4.3. Also, it is used to compute the belief value of proposition p at time step t, given the situations at the neighboring time steps $t-1$ and $t+1$. Taking Eq. (3) for computing belief values, but using the expected vector E_t instead of the actual situation vector X_t, results in the expression

$$\tau(p_t|X_{t-1}, X_{t+1}) = \frac{\sum_i \mu_i(p) E_{i,t}}{\sum_i E_{i,t}} \tag{5}$$

for the subjective probability that p is the case at time step t, given what is known about $t-1$ and $t+1$. This belief value is important for computing two measures that give information about the temporal relatedness of situations: *proposition fit* and *story coherence*, defined below.

Magliano, Zwaan, and Graesser (1999) present data showing that the extent to which a sentence fits in a story context, is rated higher if the sentence is more causally connected to the other story statements. Likewise, we define proposition fit as the proposition's strength of relation with neighboring situations. Suppose proposition p can be expected to occur at time step t given the previous and next situations, for instance because X_{t-1} is a possible cause of p_t, and X_{t+1} is a possible consequence. In that case, the belief value of p_t given X_{t-1} and X_{t+1} will be larger than the unconditional belief value $\tau(p)$. The difference between the two is the proposition fit of p_t:

$$\text{prop. fit}(p_t) = \tau(p_t|X_{t-1}, X_{t+1}) - \tau(p). \tag{6}$$

Story coherence is a measure for the extent to which a sequence of situations is in concordance with temporal world knowledge. If situations X_{t-1} and X_{t+1} increase the amount of belief in the intermediate situation X_t, then the trajectory $\langle X_{t-1}, X_t, X_{t+1}\rangle$ is temporally coherent. The coherence of a complete story trajectory is the increase in belief value by neighboring situations, averaged over all time steps:

$$\text{coh}(\bar{X}) = \frac{1}{T}\sum_t (\tau(X_t|X_{t-1}, X_{t+1}) - \tau(X_t)). \tag{7}$$

4.3. Model processing

4.3.1. Inference

The statements of a story, together with world knowledge, put constraints on the propositions that can be the case in the story. According to the DSS model, inference is reflected in the propositions' belief values changing to satisfy these constraints. This means that propositions that are likely to be the case in the story have their belief values increased, while the belief values of unlikely propositions are decreased, which generally results in increased story coherence. It is important to note that story coherence and belief values do not control or even influence the inference process, but only reflect its outcome.

Before running the model on a story, the story's situations are converted into vectors in situation space, as explained in Section 4.1. Next, the model starts processing on the first two situation vectors X_1 and X_2 simultaneously because no temporal inferences are possible with only a single situation. Contrary to the Golden and Rumelhart model, all the following story situations enter the model one by one, and are processed as they come in. When the inference process is completed for the story so far, the next situation (if any) enters the model and the process resumes. Processing of the older situations resumes as well, so existing inferences about earlier time steps can be withdrawn and new inferences can be made.

During the inference process, the model uses temporal world knowledge to convert the sequence of situation vectors (i.e., the trajectory) corresponding to the story read so far, into one that contains the information present in the story as well as new information inferred from it. This means that the process needs to solve two problems. First, the facts given by the story should be preserved. Second, the story trajectory should be adapted to temporal world knowledge.

4.3.1.1. Preventing inconsistency. Preventing text-given propositions from being denied is straightforward in Golden and Rumelhart's localist situation space: These propositions are never allowed to have their belief values changed. In the DSS model, there is no direct connection between situation vectors and propositions. Still, it is not difficult to prevent conclusions that are inconsistent with the original story. Everything outside a story situation's SOM area belongs to the negation of the situation and may therefore not be inferred during the inference process. A story situation plus extra information is always a subarea of the original situation's area. If $x^0_{i,t}$ is the value of SOM cell i at time step t of the original story, then after any amount of processing time, the current value $x_{i,t}$ may not be larger than $x^0_{i,t}$.

4.3.1.2. Applying temporal knowledge. Knowledge about the temporal patterns occurring in the microworld is encoded in matrix W. During the inference process, the trajectory is brought into closer correspondence with this matrix. This temporal pattern matching is accomplished by adjusting all individual values $x_{i,t}$ towards levels that are more likely considering the current trajectory and the values in W.

Eq. (4) gives the expected value $E_{i,t}$ of SOM cell i at time step t, given the rest of the trajectory and world knowledge. However, it follows from MRF theory that the *most likely* value for any SOM cell i at time step t is either $x_{i,t} = 0$ or $x_{i,t} = 1$ (see Appendix A). If the

expected value $E_{i,t}$ is larger than .5, the most likely value is $x_{i,t} = 1$, so $x_{i,t}$ has to increase (taking into account that its maximum value is $x^0_{i,t}$). When $E_{i,t} < .5$, the most likely value is $x_{i,t} = 0$, so $x_{i,t}$ should decrease (taking into account that it cannot become negative). This is done for all values in the trajectory $\bar{X}$ in parallel.

Since the expected values at time step t depend on the current values of $X_{t\pm1}$, and the expected values at $t \pm 1$ depend on X_t, the xs cannot be set to 0 or x^0 directly. Instead, a first-order differential equation states how the *change* in value of $x_{i,t}$ over processing time, denoted $\dot{x}_{i,t}$, depends on the current trajectory:

$$\dot{x}_{i,t} = \begin{cases} \left(E_{i,t} - \frac{1}{2}\right)(x^0_{i,t} - x_{i,t}) & \text{if } E_{i,t} > \frac{1}{2} \\ \left(E_{i,t} - \frac{1}{2}\right) x_{i,t} & \text{if } E_{i,t} \leq \frac{1}{2}. \end{cases} \tag{8}$$

The factor $(E_{i,t} - 1/2)$ makes sure that $x_{i,t}$ always changes towards a more likely value: It increases as long as $E_{i,t} > .5$ and decreases when $E_{i,t} < .5$. If $x_{i,t}$ is increasing, its rate of change is multiplied by its distance to the maximum value $x^0_{i,t}$, which prevents $x_{i,t}$ from becoming larger than this maximum. If $x_{i,t}$ is decreasing, its rate of change is multiplied by its distance to 0, which prevents $x_{i,t}$ from becoming negative.

Given an initial trajectory, Eq. (8) can be solved approximately, giving the development of the trajectory over continuous time expressed in arbitrary "model processing time" units. The original story trajectory $\bar{X}^0$ serves as the initial value for this evaluation. The equation is solved by the function ODE45 in MATLAB 6.1, using a method developed by Dormand and Prince (1980).

4.3.1.3. Depth of processing. When a situation has been sufficiently processed, the next situation is allowed to enter the model. The criterium for sufficient processing is controlled by a positive depth-of-processing parameter θ. Eq. (8) is evaluated until the trajectory's total rate of change is less than the threshold value $1/\theta$:

$$\sum_{i,t} |\dot{x}_{i,t}| < \frac{1}{\theta}, \tag{9}$$

where t ranges from 1 to the number of story situations in the model at that moment. Large values of θ correspond to deep processing, since story situations are added when inferencing on the previous situations is mostly completed. As θ decreases, the criterium for convergence becomes less stringent and the process halts even if much can still be inferred, corresponding to shallower processing. In all simulations presented here, the value of θ was set to 0.3 unless stated otherwise.

4.3.1.4. Amount of inference. At any moment during the inference process, the trajectory can be interpreted by computing the belief value of any proposition-at-time-step p_t. If the process results in an increase of belief in p_t, this means that p_t is positively inferred. Likewise, if its belief value decreases, p_t is negatively inferred, meaning that it is inferred not to be the case.

Formally, the amount of inference is defined as the increase of the proposition's belief value relative to its value in the original story:

$$\inf(p_t) = \tau(p_t|X_t) - \tau(p_t|X_t^0). \tag{10}$$

For validation against empirical findings, we also require a measure for the total amount of inference that takes place during processing. This is not simply Eq. (10) summed over all basic propositions because complex propositions should be taken into account as well. The total amount of inference is therefore determined by directly comparing the initial story trajectory to the result of its processing.

The unconditional belief value of a story situation X_t^0 is computed using Eq. (2). When new facts about the situation at time step t are inferred, it is replaced by a more informative situation X_t. Since this new situation is more specific, it is less likely to occur and has a lower unconditional belief value. The total amount of inference on the situation at time step t equals its decrease in unconditional belief value: $\tau(X_t^0) - \tau(X_t)$. This can be interpreted as the increase in the amount of knowledge there is about the situation. The total amount of inference on a trajectory is the sum of the amounts of inference on its individual situations:

$$\text{total inf}(\bar{X}) = \sum_t (\tau(X_t^0) - \tau(X_t)), \tag{11}$$

where t ranges from 1 to the number of story situations in the model at that moment. Note that the total amount of inference is largest when $\tau(X_t) = 0$, which is only the case if X_t equals the nil vector. If this happens, the model has inferred that the situation was inconsistent with the rest of the story and should not be believed. A reader making such an inference may well discard it and accept the story at face value, awaiting further information and resulting in no inference made. The model does not include a process that evaluates its inferences. Therefore, when faced with a sequence of situations that is inconsistent according to world knowledge, it can be inferred that one of the situations is impossible.

4.3.2. Retention

The DSS model can easily be adapted for modeling retention of story propositions. Over time, the memory trace of a story becomes weaker, meaning that the amount of information in it decreases. In DSS, a situation that covers a large part of the SOM contains less information than a situation that covers only a small part. Therefore, reducing the amount of information in a trajectory corresponds to an increase in the SOM cell values $x_{i,t}$. The rate of increase depends on world knowledge and on the rest of the trajectory. If there is much evidence that some $x_{i,t}$ has a small value (i.e., $E_{i,t}$ is small), this value will drift up to 1 more slowly. The weakening of a story's memory trace over retention time is given by the differential equation

$$\dot{x}_{i,t} = E_{i,t}(1 - x_{i,t}). \tag{12}$$

The value of $x_{i,t}$ increases at a rate equal to $E_{i,t}$, multiplied by its distance to 1 in order to prevent $x_{i,t}$ from exceeding this maximum value. The story trajectory resulting from the inference process serves as the initial value for the evaluation of Eq. (12). The time over which the equation is evaluated corresponds to the amount of time that elapsed since the story was read, expressed in arbitrary "model retention time" units.

The retention process, like the inference process, is not affected by belief values but results in adjusted belief values. As retention time grows, the belief values of the story's propositions regress to their unconditional levels. The extent to which a proposition is still retained can therefore be defined as the difference between its current and its unconditional belief value:

$$\mathrm{ret}(p_t) = \tau(p_t|X_t) - \tau(p). \tag{13}$$

5. Results

5.1. World-knowledge implementation

As explained in detail in Appendix B, the world knowledge used by the model was extracted in two steps from the microworld description developed in Section 3. First, the microworld description was used to train a Self-Organizing Map, the result of which is shown in Fig. 2. Next, these mappings were used to convert the microworld description into a distributed representation, from which the world-knowledge matrix W was computed.

How can we ascertain that world knowledge was implemented successfully? Be reminded that the belief values, based on this world-knowledge implementation, can be interpreted as the subjective probabilities of propositions. The probabilities also follow directly from the microworld description. By comparing these "actual" probabilities (Pr) to the belief values (τ) as computed by Eqs. (2), (3) and (5), it can be established whether the model's world knowledge reflects the regularities that hold in the microworld.

First, the subjective and actual probabilities of all conjunctions $p \wedge q$ of (negations of) basic propositions were compared. The resulting scatter plot is shown in the left panel of Fig. 3. Second, we tested whether the non-temporal dependencies among propositions are captured by their vector representations. If a (negation of a) basic proposition p is given, the probability that a positive basic proposition s is the case at the same moment in the microworld description equals $\Pr(s|p)$. Of all situations in which p is the case, this is the proportion that include s. The overall proportion of situations in which proposition s occurs is its *a priori*

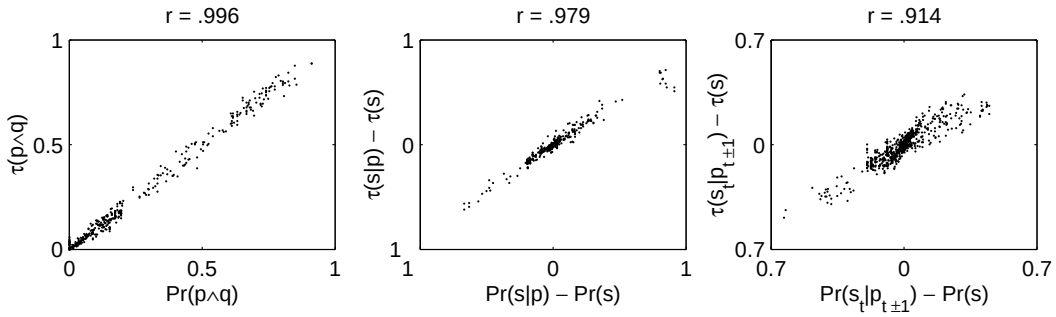

Fig. 3. Scatter plots of actual probabilities (Pr) versus subjective probabilities (τ), and coefficients of correlation (r). Propositions denoted by p or by q are basic propositions or negations thereof. Propositions denoted by s are positive basic propositions.

Table 2
Three stories used to model specific inferences

Story	t	Situation	Possible text
1	1	¬Rain ∧ ¬Sun	It doesn't rain and the sun doesn't shine.
	2	(Soccer ∨ Hide-and-seek ∨ (B computer ∧ J computer)) ∧ (B wins ∨ J wins)	Bob and Jilly are playing a game and one of them wins.
2	1	Sun	The sun is shining.
	2	Hide-and-seek	Bob and Jilly are playing hide-and-seek.
	3	¬(B outside) ∧ ¬(J outside)	They are inside.
3	1	Sun ∧ Soccer	The sun shines and Bob and Jilly play soccer.
	2	B tired ∧ ¬(J tired)	Bob is tired, but Jilly isn't.
	3	B wins ∨ J wins	Next, one of them wins.
	4	B tired ∧ J tired	Now they are both tired.
	5	Rain	It starts raining.
	6	B inside ∧ J inside ∧ Hide-and-seek	Bob and Jilly go and play hide-and-seek inside.
	7	J tired ∧ ¬(B tired)	Only Jilly is tired.
	8	B wins ∨ J wins	Someone wins.
	9	B computer ∧ J dog	Later, Bob is playing a computer game, and Jilly is playing with the dog.

For each situation, it is shown how it is constructed from basic propositions and a possible text describing this situation is given.

probability $\Pr(s)$, so the probability of s changes by an amount $\Pr(s|p) - \Pr(s)$ under influence of p. The center panel of Fig. 3 shows the scatter plot of these actual probability differences versus their corresponding subjective probability differences. Third, we tested whether the world-knowledge matrix W correctly captures temporal dependencies among propositions. In the microworld, the amount of influence that a proposition p at time step $t \pm 1$ has on a proposition s at t is $\Pr(s_t|p_{t\pm1}) - \Pr(s)$, the change in probability of s_t. All positive basic propositions s_t and (negations of) basic propositions $p_{t\pm1}$ were used for the scatter plot of actual versus subjective probability differences in the right panel of Fig. 3.

In all three cases, the correlation between actual and subjective probabilities was very high: .996, .979, and .914, respectively. Also, the three scatter plots show that there are no outliers. In short, the vector representation of propositions and the world-knowledge matrix W did capture the regularities that occurred in the microworld description.

5.2. Inference

5.2.1. Specific inferences

In order to test the model's ability to make specific inferences, three simple sequences of situations ("stories") were constructed. These stories, shown in Table 2, varied in length from two to nine situations. Each story was meant to evoke one or more specific inferences.

- *Story 1: realizing the exclusive-or relation*. From the fact that Bob or Jilly wins, it can be inferred that they must have been at the same place in the previous time step. This inference requires the exclusive-or relation: If *either* Bob *or* Jilly is outside at t, winning

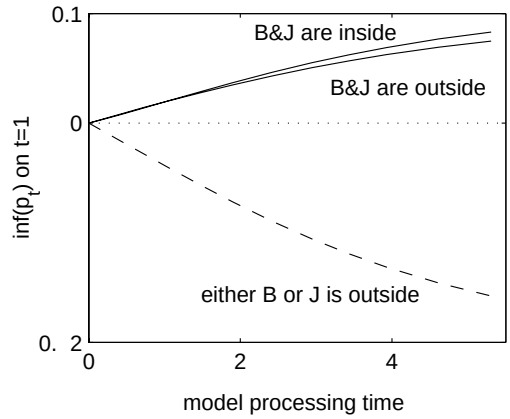

Fig. 4. Amount of inference (Eq. (10)) of "Bob and Jilly are outside" (B outside $\wedge$ J outside), of "Bob and Jilly are inside" ($\neg$(B outside) $\wedge$ $-$(J outside)), and of "either Bob or Jilly is outside" (B outside XOR J outside) during processing of Story 1.

cannot occur at $t+1$; if *both* are (not) outside at t, winning is possible at $t+1$. Story 1 tests whether this knowledge was successfully implemented in matrix W. The situation at $t = 1$ gives no indication where Bob and Jilly might be. Following this, someone wins, which means that they both must have been either outside or not outside (which is equivalent to being inside) at $t = 1$. The model is able to correctly infer this, as can be seen from Fig. 4. This result shows that the DSS model can handle the exclusive-or relation required to make this inference.

The amounts of inference of "Bob and Jilly are outside" and of "Bob and Jilly are inside" seem fairly low. There are two reasons for this. First, these two situations exclude each other and can therefore never be both strongly inferred. Second, Bob and Jilly are *a priori* more likely to be at the same place than to be at different places. As a result, the belief values for "Bob and Jilly are (not) outside" are high to begin with and cannot increase much more.

- *Story 2: retracting an inference*. After reading the first two sentences of Story 2, one might infer that Bob and Jilly play hide-and-seek *outside*. This inference is based on the information that the sun shines and on the knowledge that this usually causes them to be outside. However, the third sentence tells us that they are in fact inside at $t = 3$. This does not necessarily mean that they were already inside at $t = 2$, but it does make that more likely. Therefore, the inference that Bob and Jilly are outside at $t = 2$ should be retracted. As Fig. 5 shows, this is indeed what the model does. At first, the belief value of "Bob and Jilly are outside" at $t = 2$ increases. After 5.38 units of model processing time have passed, the process stabilizes enough (i.e., the total trajectory change is less than $1/\theta$, as in Eq. (9)) to allow the third situation to be added to the story trajectory. At that moment, the belief value decreases almost to its original level: it is no longer inferred that Bob and Jilly are outside during story time step $t = 2$.
- *Story 3: inferring who wins at what*. Whoever is tired, is less likely to win. In Story 3, it is Bob who is tired at first, so the one who wins at $t = 3$ is probably not him, but Jilly.

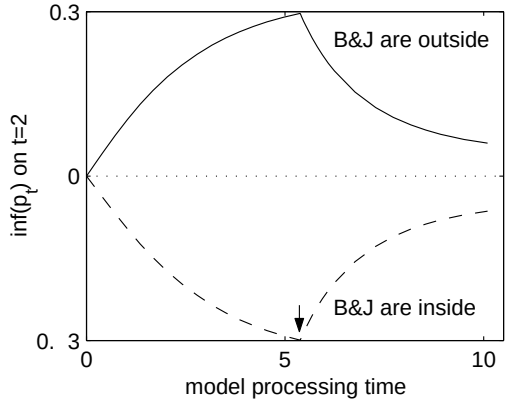

Fig. 5. Amount of inference (Eq. (10)) of "Bob and Jilly are outside" and of "Bob and Jilly are inside" at $t = 2$, during processing of Story 2. The third situation enters the model after 5.38 units of processing time, as indicated by the arrow.

The left graph in Fig. 6 shows that the model infers exactly this. The right graph shows that Bob wins later in the story ($t = 8$), when Jilly is tired.

Also, the model infers what Bob and Jilly are playing when one of them wins. Note that the game being played is mentioned two time steps before it is stated that someone wins. Still, it is inferred that the game being won is soccer at $t = 3$ and hide-and-seek at $t = 8$. Since situations are only directly influenced by the previous and next time steps, this information must have travelled through the intermediate time steps $t = 2$ and $t = 7$, respectively, showing that indirect influence from more distant situations is indeed possible.

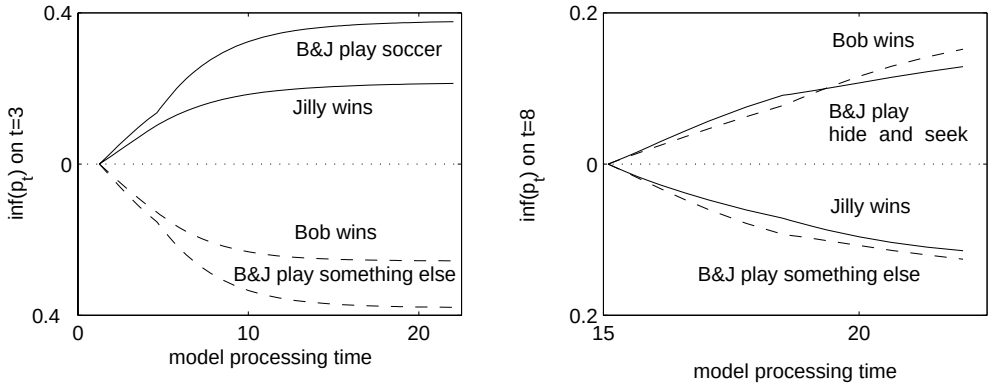

Fig. 6. Amounts of inference (Eq. (10)) during processing of Story 3. Left: inference at $t = 3$ of "Bob wins," of "Jilly wins," of "Bob and Jilly play soccer," and of "Bob and Jilly play something else" (Hide-and-seek $\vee$ (B computer $\wedge$ J computer)). The third situation is added to the story trajectory at 1.19 units of processing time after the inference process began with the first two situations. Right: inference at $t = 8$ of "Bob wins," of "Jilly wins," of "Bob and Jilly play hide-and-seek" and of "Bob and Jilly play something else" (Soccer $\vee$ (B computer $\wedge$ J computer)). The eighth situation enters the model at 15.10 units of processing time.

5.2.2. Inferences in general

The previous section shows that the model's inferences correspond to our intuitions: propositions that are implied by the story are inferred. In order to test this more systematically, 100 random stories were constructed and used as input to the inference model. The stories varied in length from three to seven situations. There were 20 random stories of each length, so the total number of story situations equaled $20 \times (3 + 4 + 5 + 6 + 7) = 500$. Each such situation consisted of exactly one basic proposition or its negation.

In general, propositions that are inferred on the basis of temporal world knowledge should be

- implied by the story;
- possible given the story situation. If it is stated that Bob is outside, it cannot be inferred that he is inside at the same moment in story time;
- not already given by the story situation. If Bob and Jilly play soccer, then they must be outside. This inference does not require information from other story situations, and is therefore not an inference in the sense of Eq. (10).

For each situation of each random story, the proposition fit (Eq. (6)) and amount of inference (Eq. (10)) of all basic propositions were obtained (except for the proposition that constituted the situation). The correlation between amount of inference and fit of propositions was .66 (based on 500 situations $\times$ (14 $-$ 1) propositions $=$ 6,500 observations), indicating that the model does indeed infer propositions that are implied. Moreover, propositions with positive fit were inferred to be the case (positive inference) and propositions with negative fit were inferred to be not the case (negative inference).

Whether a proposition p_t is possible given the original story situation X_t^0, can be seen from its initial belief value $\tau(p_t|X_t^0)$. If this value is close to 0, p_t is unlikely to be the case at that moment in the story and should not be inferred even if it is a likely proposition given the rest of the story. Likewise, if the initial belief value is close to 1, p_t is already likely given story situation X_t^0 and it should not be inferred to be the case at t from situations at other story time steps.

Indeed, this is what the model predicts. All 500 situations $\times$ 14 basic propositions $=$ 7,000 observations were divided into two groups. The "non-inferable" group contained cases with initial belief values so close to 0 or1 (less than .001 or more than .999) that inference was not expected to occur. The "inferable" group contained the others. The average absolute proposition fit was .11 among the non-inferables and .08 among the inferables, indicating that the latter would be inferred less if only proposition fit would matter. However, the opposite was the case: The average absolute amount of inference was .07 among the inferables but only 2.4×10^{-5} among the non-inferables.

5.2.3. Inference and coherence

As noted in the introduction, the inferences that readers most easily make on-line are inferences that contribute to the coherence of the story, which in the model is defined in Eq. (7). The coherences of the 100 random stories ranged from $-$.23 to .25, with an average of .001. Since coherence is a measure for the match between a story and temporal world knowledge, and the inference process adapts the trajectory to world-knowledge matrix W, the story coherences of the trajectories increased through this process. The result was a larger coherence value for

Table 3
Stories with different relatedness levels

t	Relatedness level	Situation	Possible text
1		Soccer	Bob and Jilly play soccer.
2	1	B tired $\wedge$ $\neg$(J tired)	Only Bob is tired.
	2	B tired	Bob is tired.
	3	B tired XOR J tired	One of them is tired.
	4	J tired	Jilly is tired.
	5	J tired $\wedge$ $\neg$(B tired)	Only Jilly is tired.
3		B wins	Bob wins.

Relatedness level is varied by using one of the five situations at $t = 2$.

all 100 stories (the average was .28), showing that the inferences contributed to the stories' coherence. However, this is not a built-in consequence of the model's equations: Transient decreases of coherence during processing were observed for 17 stories, taking 4.5% of their processing time.

5.2.4. Relatedness, inference and reading time

A story sentence is read faster when it is more related to the preceding sentence. Myers, Shinjo, and Duffy (1987), and also Golding, Millis, Hauselt, and Sego (1995), showed this by having subjects read stories consisting of just two events. The relatedness between those events varied: The second story event was either unrelated to the first event or was predictable to a certain degree. They found that reading the second sentence took more time when it was less related to the first sentence. Murray (1995, 1997) also had subjects read two-sentence stories but included stories in which the events were adversatively related, meaning that the first story event made the second event less likely to occur. He found that the second sentence took more time to read when it was adversatively related to the first sentence than when it was unrelated. Using more realistic texts, Sanders and Noordman (2000) showed that a sentence is read faster when it is embedded in a text that causally implies it, than when it is not causally related to the rest of the text.

To test whether the model predicts the same relation between relatedness and reading time, five stories with different levels of relatedness were constructed. Each of the stories, shown in Table 3, consisted of three situations, the first of which was "soccer" and the last was "Bob wins." Relatedness was varied among stories by modifying the second situation. Since Bob is more likely to win when Jilly is tired, stating that Jilly is tired and Bob is not, should result in the highest relatedness to the last situation. If, on the other hand, Bob is tired and Jilly is not, relatedness is lowest. Intermediate levels of relatedness are obtained in a similar way.

The time needed by the model to process the last situation and the amount of inference that took place during this process, are plotted in Fig. 7. These results clearly show that a higher level of relatedness leads to shorter processing time and less inference, which is consistent with the generally accepted idea that the on-line construction of an inference takes time. For instance, Vonk and Noordman (1990) had subjects read texts that contained an inference evoking sentence. When the information to be inferred was explicitly stated in the text before

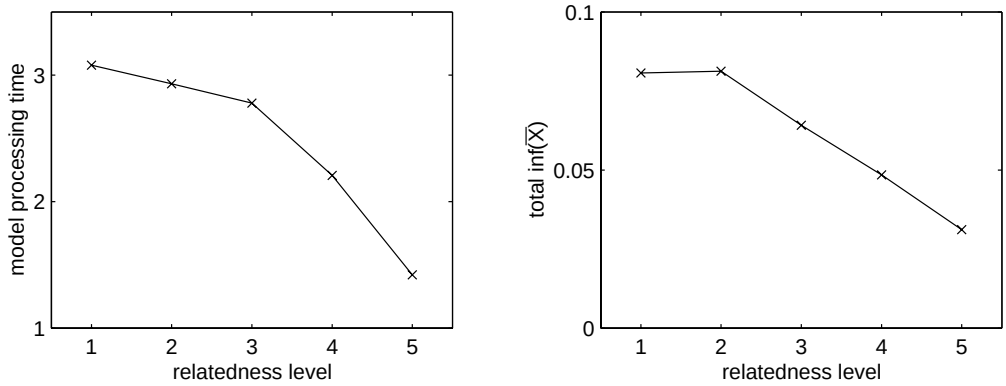

Fig. 7. Amount of time needed to process the situation "Bob wins" (left) and total amount of inference (Eq. (11)) that took place during this process (right), as a function of the situation's relatedness to the previous situation.

the inferring sentence, reading times on the inferring sentence were shorter than when the information was not stated but had to be inferred.

Stories describing less related events evoke more inferences, which slows down reading. To test whether this relation holds in general, the model was run on all stories consisting of just two situations, with each situation consisting of exactly one (negation of a) basic proposition. Since there are 14 positive basic propositions, the number of stories was $(2 \times 14)^2 = 784$. Story coherence was taken as a measure of relatedness. These ranged from $-.34$ to .38, so the relatedness of the two situations ranged from adversative to predictable.

Fig. 8 directly compares the model's results to those of Golding et al. (1995) and Myers et al. (1987). Since they did not use stories with adversatively related sentences, only the model's results for the 394 stories with non-negative coherence are plotted. The effect of story

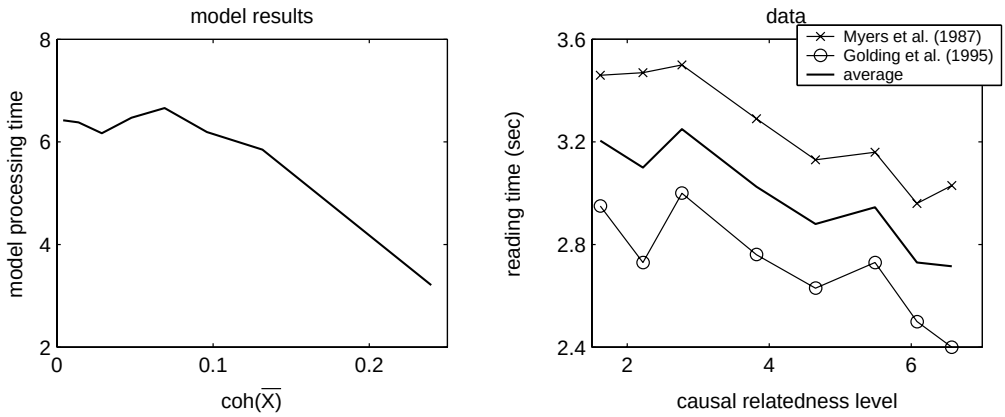

Fig. 8. Left: amount of processing time needed to process two-sentence stories, as a function of story coherence (Eq. (7)). Each of the eight points in the graph is the average of processing times and coherences for 49 or 50 stories. Right: reading time on the second sentence of two-sentence stories, as a function of sentence relatedness (Golding et al., 1995; Myers et al., 1987).

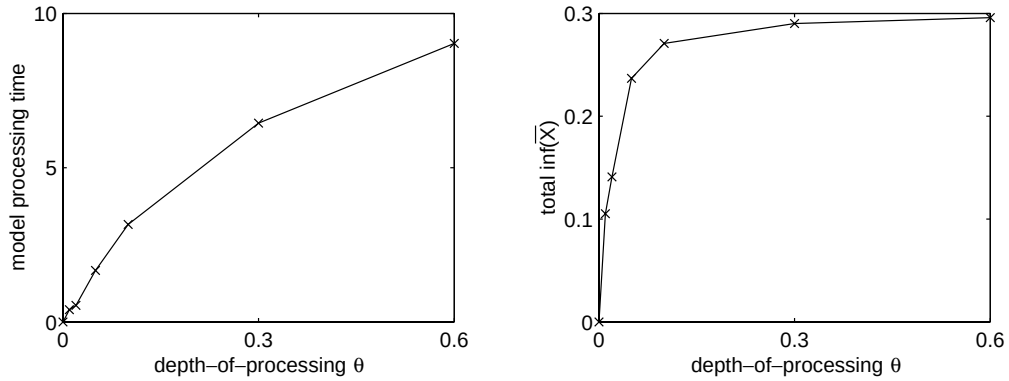

Fig. 9. Effect of depth-of-processing parameter θ on average processing time per situation (left) and total amount of inference (Eq. (11)) during processing of each situation (right).

coherence on processing time as found by the model is quite similar to the effect of relatedness on reading time as found by Myers et al. and Golding et al.

Over all 784 stories tested, the correlation between story coherence and amount of inference was $-.42$. A sequence of story situations that violates temporal world knowledge will evoke more inferences than a story that is in accordance with world knowledge. This results in an increase in processing time. Accordingly, there was a negative correlation ($r = -.36$) between story coherence and model processing time. The model correctly predicts that stories with adversatively related events are processed slower than stories that describe unrelated events, and that stories describing positively related events are processed quickest. The strong relation between amount of inference and processing time was also reflected in the high positive correlation ($r = .93$) between the two.

5.2.5. Inference and depth of processing

Noordman et al. (1992) varied subjects' reading goal by either instructing them to check for inconsistencies in a text, or by not giving such an instruction. They found that the consistency-checking instruction led to more inferences and longer reading times. Stewart, Pickering, and Sanford (2000) used another method to manipulate the reading process. Their subjects read single sentences and had to answer a related question after every sentence. In one condition, all of these questions could be answered without making any inference from the sentences, while in the other condition inferences needed to be made from every sentence. It was found that reading slowed down when inferencing was required compared to when it was not.

Supposedly, instructing readers to check for inconsistencies or having them answer inference-requiring questions leads to deeper processing of the texts. In the model, depth of processing is controlled by parameter θ. Fig. 9 shows the effect of varying θ on average processing time per situation and total amount of inference during processing of each situation, for the 100 random stories constructed in Section 5.2.2. In accordance with empirical data, deeper processing resulted in longer processing times[6] and more inference.

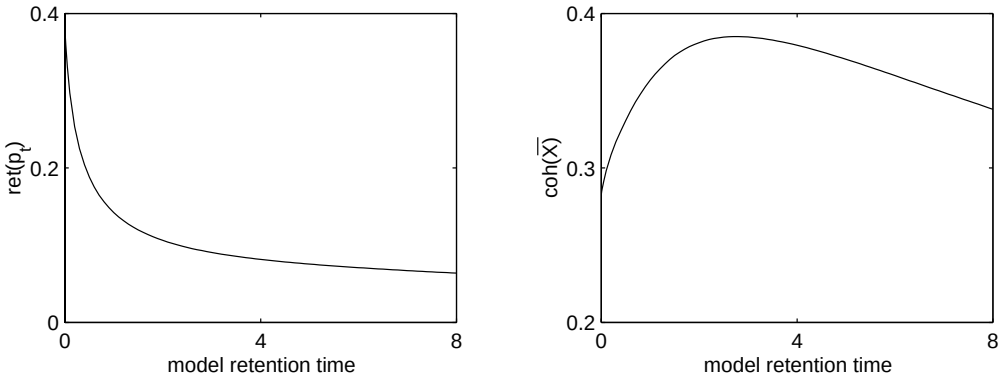

Fig. 10. Left: average amount of retention (Eq. (13)) of story propositions of the 100 random stories, as a function of retention time. Right: average coherence (Eq. (7)) of retained story trajectories as a function of retention time.

5.3. Retention

The inference results showed that inferences contributed to the stories' coherences. The retention process behaves similarly: Although retention of story propositions decreases as retention time grows (Fig. 10, left) the average coherence of the retained trajectories shows an increase before it starts to decrease after approximately three units of retention time (Fig. 10, right). As retention time reaches infinity, all SOM cell values approach 1, which by Eqs. (5) and (7) means that the coherence equals 0.

There are two explanations for the increase in coherence during retention. First, propositions may be forgotten selectively. Indeed, it is well known that some story propositions are recalled more easily than others. In a cued recall task, Myers et al. (1987) found that, in general, a sentence was more likely to be recalled if it was more related to the one that was given as a recall cue. However, the highest levels of relatedness resulted in a small decrease in recall. This last effect was not found by Varnhagen, Morrison, and Everall (1994). They had children read a number of stories and asked them to recall as much of the stories as possible, without giving any sentences as cue. Story propositions with many causal connections in the story were recalled more often than propositions with fewer connections. No decrease in free recall for the highest levels of connectivity was found. The same relation between number of causal connections and recall probability was found by Trabasso and Van den Broek (1985) and by Fletcher and Bloom (1988). In short, propositions that form the "causal backbone" of the story are remembered best. Moreover, Goldman and Varnhagen (1986) found that this effect is stronger in a delayed free recall task than in immediate recall. Not surprisingly, they also found that fewer story propositions are recalled in delayed recall than in immediate recall, as did many other researchers (e.g., Duffy, Shinjo, & Myers, 1990; Trabasso & Van den Broek, 1985).

Another reason for the increasing coherence might be the occurrence of intrusions. Readers occasionally recall propositions that were never part of the text, but are part of their knowledge. Bower, Black, and Turner (1979) as well as Smith and Graesser (1981) found that propositions that form part of a story script and are therefore highly predictable in the story, are falsely recalled more often than less predictable propositions. Luftig (1982) too found higher intru-

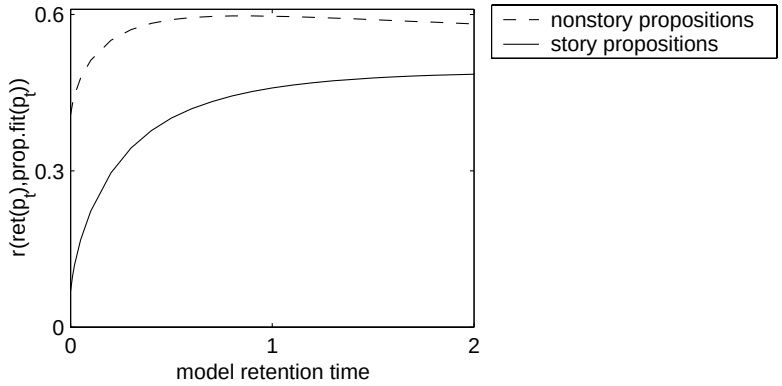

Fig. 11. Correlation between retention (Eq. (13)) and fit (Eq. (6)) of propositions as a function of model retention time, for story propositions and non-story propositions (intrusions).

sion rates of propositions that, according to world knowledge, follow from the text than of propositions that do not. Moreover, this effect was stronger in a delayed recall task than in immediate recall.

If the model accounts for these empirical data, there should be a positive correlation between proposition fit and retention both for story propositions and for non-story propositions (intrusions). Moreover, these correlations should increase over retention time. Fig. 11 shows that the model does predict all of these effects. After two units of retention time, the correlation between fit and retention is .49 (based on 500 observations) for story propositions and .58 (based on 6,500 observations) for non-story propositions.

6. Discussion

6.1. Evaluation of the model

6.1.1. Inference

The DSS model takes as input a temporal sequence of story situations and uses world knowledge about temporal contingencies to infer which propositions that were not stated in the story are likely to be the case. The results show that this was successful: The inferred propositions were temporally implied by the story statements, were not impossible or already given in the story, and contributed to the story's coherence. The model does not make any distinction between reasoning forwards and backwards in story time, nor between inferring a single-situation event (like winning) and events than span multiple story time steps (like rain). All of these are handled by the same inference process. More importantly, the inference model was validated against several experimental findings. Processing of less coherent stories took more time because these stories evoked more inferences than did more coherent stories. Also, increasing depth of processing led to more inference and slower reading. That the model accounts for these experimental results is not trivial. The model was only designed to perform inferencing by adjusting incomplete descriptions of story events to world knowledge, and not

to simulate particular experimental data. Therefore, accounting for these data is an emergent property of the model.

The two major theories of on-line inference are the minimalist theory (McKoon & Ratcliff, 1992), which claims that readers do not commonly create elaborate situation models during reading, and the constructionist theory (Graesser et al., 1994), which says that readers do form such situation models. Clearly, the DSS model leans more towards the latter account since all of its inferences are based on situation models. However, the model also differs from the constructionist theory in one important respect. Graesser et al. claim that readers actively try to accomplish coherence of a text, according to the so-called search-after-meaning principle. In other words, inferencing is driven by a need for coherence. The model offers a reverse interpretation: Increased coherence results from inferences, which emerge from matching the events described in the story to patterns of events known to occur in the world. There is no search for story coherence. Rather, incoming information automatically adjusts the story trajectory, generally resulting in increased coherence.

Technically, making the model coherence-driven is not hard to do. A standard gradient-ascent algorithm can be applied to search for a local maximum of story coherence (Eq. (7)) starting with the original story trajectory. However, such a coherence-driven implementation is theoretically excluded in our approach. The definition of story coherence is based on belief values, which depend on the situation vectors but cannot influence them. Therefore, the inference process can never be controlled by the story's coherence. Nevertheless, if the increase of coherence is wrongly interpreted as the driving force of the process instead of its consequence, this leads to the illusion of an active search for coherence. The switch from localist to distributed representations makes clear how belief values and coherence form an abstraction, based on a story representation, and can therefore not change the story representation. This shows that using distributed representations is not only useful in practice, allowing for more flexibility in representing situations and world knowledge than localist representations like Golden and Rumelhart's, but is also of theoretical importance.

6.1.2. Retention

The retention model showed that story coherence can increase over retention time although story propositions are forgotten. Like the inference process, the retention process does not look for coherence nor are propositions in any way selected to be retained or forgotten. Preservation of coherence simply follows as an emergent property from the differential equation that defines the retention process. This equation does not know about coherence or even propositions, and cannot make use of such higher level concepts. Nevertheless, the model correctly predicted empirical recall data. Propositions that were more related to the story were retained better than unrelated propositions and intrusion rates were higher for predictable propositions than for unpredictable ones. Both of these effects increased as retention time grew.

6.2. Limitations and possible improvements

6.2.1. Temporal world knowledge

Using the MRF framework guarantees a mathematically sound model. However, certain architectural assumptions were made in order to simplify the MRF analysis, and it is unclear to

what extent these limit the model's abilities. In particular, the symmetry assumption claimed that the temporal knowledge matrix W can be used to reason forwards in story time and the transposed matrix W' to reason backwards. However, there is no reason to assume that the real world shows the same symmetry.

It is important to note that from the symmetry assumption it does not follow that the belief values are symmetrical. In general, $\tau(p_t|q_{t\pm1}) \neq \tau(q_{t\pm1}|p_t)$. To give an example: If Bob or Jilly wins at t, it is certain that they did not play with the dog at $t-1$. This is reflected in the high belief value $\tau(\neg(\mathrm{B} \vee \mathrm{J\,dog})_{t-1}|\mathrm{B} \vee \mathrm{J\,wins}_t) = .90$. On the other hand, given that Bob and Jilly do not play with the dog at $t-1$, it is not at all certain that one of them will win at t. Indeed, the corresponding belief value is $\tau(\mathrm{B} \vee \mathrm{J\,wins}_t|\neg(\mathrm{B} \vee \mathrm{J\,dog})_{t-1}) = .24$.

The high correlation between microworld probability differences ($\Pr(p_t|q_{t\pm1}) - \Pr(p)$) and belief value differences ($\tau(p_t|q_{t\pm1}) - \tau(p)$) shows that at least in our microworld, the symmetry assumption does not seriously limit the quality of the knowledge matrix. In other (micro)worlds, this might be different. Fortunately, the MRF approach is not necessary for the model's functioning and can easily be replaced by only changing the definition of the $E_{i,t}$-function (Eq. (4)). This function gives the expected value for $x_{i,t}$, given the previous and/or next situations and world knowledge. It is the model's central function, since it states how world knowledge is implemented and applied to a story representation. If a better implementation of world knowledge is found, or a better way to apply it to the story trajectory, only the $E_{i,t}$-function needs to be changed accordingly. All of the four assumptions on which world-knowledge implementation is based can be discarded if a better $E_{i,t}$-function is to be found without them.

6.2.2. Stories versus texts

As far as the model is concerned, a story is no more than a temporal sequence of situations. This makes story comprehension no different from understanding events going on in the real world. The reader of a text, however, can make use of information that is not available to an observer of real world events. In particular, causal connectives like "because" and "although" can influence the processing of a text (Millis & Just, 1994) and its recall (Millis, Graesser, & Haberlandt, 1993), but are of course not available in the real world.

The same is true for temporal connectives. For instance, in Story 2 of Section 5.2.1, Bob and Jilly are inside at $t = 3$ from which it is inferred that they were also inside at $t = 2$, when they were playing hide-and-seek. It is not possible to tell the model that a new episode had started at $t = 3$ by adding a connective like "next" or "then." Bestgen and Vonk (1995) showed that temporal markers like "then" reduce the availability of information in the previous sentence, so such a connective could signal that the situations at $t = 2$ and 3 do not need to influence each other.

The current model does not make use of textual information because it represents stories at the situational level of text representation. Situations (or "facts," as Kintsch and Van Dijk call them) are related by the effect that they have on each other's probabilities: "relations between facts in some possible world [. . .] are typically of a conditional nature, where the conditional relation may range from possibility, compatibility, or enablement via probability to various kinds of necessity" (Kintsch & Van Dijk, 1978, p. 390). At the textbase level, propositions are not related by probability, but "connection relations between propositions in a coherent text base are typically expressed by connectives such as 'and,' 'but,' 'because,' 'although,' 'yet,' 'then,' 'next,' and so on" (p. 390).

This raises the question whether such a textbase level can be added to the DSS model. Textual information carried by, for instance, connectives is present at this level and can influence the inferencing process that takes place at the situational level. Currently, we are investigating ways to extend the DSS model with such a textbase level.

Notes

1. The world-knowledge matrix is denoted by B in Golden and Rumelhart's notation.
2. In the field of propositional logic, the De Morgan law states how disjunction can be rewritten in terms of negation and conjunction: $p \vee q \equiv \neg(\neg p \wedge \neg q)$. However, this does not help here, since the conjunction $\neg p \wedge \neg q$ must be represented as a single proposition in order to be negated.
3. This follows directly from the expression for the probability of a proposition (Eq. (A.5) in Appendix A). If it is known that proposition q causes p and that r causes p, both w_{qp} and w_{rp} are positive. If only q_{t-1} is the case, the probability of p_t equals the logistic function of $a_p + w_{qp}$. If both q_{t-1} and r_{t-1} occur, the probability of p_t is the logistic function of $a_p + w_{qp} + w_{rp}$, which must be larger since $w_{qp} > 0$, $w_{rp} > 0$ and the logistic function is monotonically increasing. Ergo, it is impossible to represent the knowledge that both q and r cause p, but $q \wedge r$ causes $\neg p$.
4. This is not the only way to model negation and conjunction in fuzzy logic. In particular, $\mu_i(p \wedge q) = \min\{\mu_i(p), \mu_i(q)\}$ is often used. However, using the product to model conjunction yields Eq. (3), which has the useful property that $\tau(p|X) = 1 - \tau(\neg p|X)$.
5. Note that the belief value of p given an $X = p$ can be somewhat less than 1, reflecting the uncertainty inherent in fuzzy logic systems. In practice, however, the subjective probabilities correspond very closely to the actual probabilities in the microworld, as is shown in the results of Section 5.1.
6. From Eq. (9), which determines when processing of a situation is completed, it might seem as if processing time can never decrease with increasing θ. However, this only is true for the first two story situations. Deeper processing can lead to shorter processing time for the situation at $t + 1$ if this situation is highly compatible with an inference that was made at story time step t. With shallower processing this "$t + 1$-compatible" inference might not be made, leading to longer processing time for $t + 1$. In fact, when comparing $\theta = 0.3$ to $\theta = 0.6$, this effect occurs in three of the 100 random stories.

Acknowledgments

We would like to thank three anonymous reviewers and Gary Dell for their helpful comments. The research presented here was supported by Grant 575-21-007 of the Netherlands Organization for Scientific Research (NWO) awarded to L.G.M. Noordman, W. Vonk, and the late E.E. Roskam.

Appendix A. Markov random fields

The mathematics of both the Golden and Rumelhart and the DSS model are based on Markov random field theory. A simplified introduction to this theory, applied to the two models, is presented here. For a more thorough explanation, see for instance Golden (1996, Chap. 6.3) or Cressie (1991, Chap. 6.4).

A.1. Model architecture

Suppose we have m random variables $z_1, \ldots, z_m$, all real valued on the interval [0, 1]. If the probability of the value of z_i is dependent on the value of z_j, then z_i is said to be connected to z_j. Note that, if z_i is connected to z_j, then z_j is also connected to z_i. Since a value depends on itself, all variables are connected to themselves. Such a system is called a Markov random field if every combination of values of $z_1, \ldots, z_m$ has a positive probability density. Since probability densities can be arbitrarily close to 0, this is not a serious restriction. For the Golden and Rumelhart model, the random variables are the values of propositions-at-time-steps. For the DSS model, they are the values of SOM-cells-at-time-steps. From here on, we shall use the term "cell" to refer to both propositions and SOM cells, and a cell-at-a-time-step will be referred to as a "node."

Any particular configuration of values $Z = (z_1, \ldots, z_m)$ is an instantiation of a field and has associated with it a probability density $P(Z)$. Since this refers to a complete instantiation, it is a *global* probability density. It is not easy to compute, but we can compare the probability densities of two instantiations.

The Hammersley–Clifford theorem (1971, unpublished) as described in Besag (1974) states how a valid probability distribution over a Markov random field can be constructed. First, we need to define the notion of a clique: A clique *is a set of variables that are all connected to each other*. Since variables are connected to themselves, every single variable forms a clique. Now let Z_1 and Z_2 be two instantiations of a Markov random field. The Hammersley–Clifford theorem states that $P(Z)$ forms a valid probability density function if and only if

$$\frac{P(Z_1)}{P(Z_2)} = \mathrm{e}^{Q(Z_1)-Q(Z_2)} \tag{A.1}$$

with Q a function of the form:

$$\begin{aligned} Q(z_1, \ldots, z_m) = &\sum_{i=1}^{m} z_i G_i(z_i) + \sum_{i=1}^{m}\sum_{j>i}^{m} z_i z_j G_{ij}(z_i, z_j) + \sum_{i=1}^{m}\sum_{j>i}^{m}\sum_{k>j}^{m} z_i z_j z_k G_{ijk}(z_i, z_j, z_k) \\ &+ \cdots + z_1 z_2 \ldots z_m G_{1,2,\ldots,m}(z_1, z_2, \ldots, z_m). \end{aligned} \tag{A.2}$$

Here, the G's are functions such that $G_{ij\ldots}(z_i, z_j, \ldots) = 0$ if variables $z_i, z_j, \ldots$ do not form a clique. For use in the two models, Eq. (A.2) simplifies greatly. Two simplifications follow from the models' architecture, shown graphically in Fig. 12. First, it can easily be seen that there exist no fully connected groups (cliques) of more than two nodes. This means that every G function with more than two arguments equals 0 and these terms disappear from Eq. (A.2). Only the first two terms are left over.

Second, the strength of dependencies between nodes is the same for all story time steps. This means that not all connected pairs of nodes need to be stated in Eq. (A.2) separately

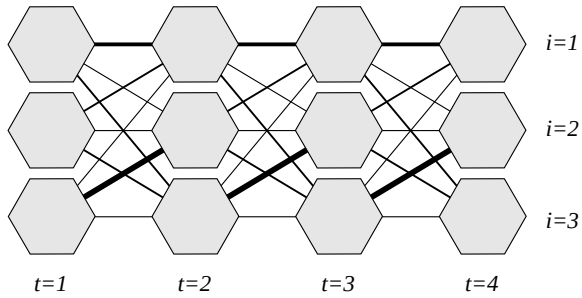

Fig. 12. Architecture of the Golden and Rumelhart and the DSS model, in a story world consisting of $n = 3$ cells ($i = 1, 2, 3$) and $T = 4$ time steps ($t = 1, \ldots, 4$), making a total of $m = 12$ nodes. Every row corresponds to a cell. Every column corresponds to a time step. Two nodes are connected only if they are from neighboring time steps. The thickness of a connection indicates the strength of the dependency.

because they can be summed over all time steps. The variables z_i and z_j are therefore replaced by variables $x_{i,t-1}$ and $x_{j,t}$ that have additional time step indices. Instead of summing over m variables, we now sum over n cells and T time steps.

Finally, it is assumed that a node's contribution to Q increases linearly with its value. This is accomplished by turning the G functions into constants. The first term of Eq. (A.2) gives rise to n of these: $G_1, \ldots, G_n$, which will be denoted by the vector $A = (a_1, \ldots, a_n)$. The second line of Eq. (A.2) gives rise to $n \times n$ constants $G_{11}, \ldots, G_{nn}$, which form a matrix that will be denoted $W = (w_{ij})_{i,j=1,\ldots,n}$. The resulting, simplified Q function is

$$Q(\bar{X}) = \sum_{t=1}^{T}\sum_{i=1}^{n} x_{i,t}a_i + \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{i,t-1}x_{j,t}w_{ij} = \sum_{t=1}^{T}(X_tA + X_{t-1}WX_t). \tag{A.3}$$

The xs refer to a trajectory $\bar{X} = \langle X_1, X_2, \ldots, X_T \rangle$ consisting of T time steps. For the equation to be valid at $t = 1$, all values at the non-existent time step $t = 0$ are defined to be 0.

In the Golden and Rumelhart model, the unconditional probability of a proposition i is a function of a_i only. In the DSS model this value has become obsolete, since knowledge about the unconditional probabilities is incorporated in the propositions' vector representations. Therefore, in the DSS model A is set to the zero vector and the corresponding term drops from Eq. (A.3).

A.2. Expected value

Both in the Golden and Rumelhart and the DSS model the expected value of a node is computed using the local probability distribution $P_{i,t}$. This is the probability distribution of node (i, t) given the values of all other nodes. Eq. (A.1) gives the ratio of two global probabilities, but the ratio of the corresponding local probabilities is easily shown to be the same.

Let $\bar{X}^*_{i,t}$ denote the collection of all values of the trajectory except that of node (i, t). The ratio of the local probabilities $P_{i,t}(x_1)$ and $P_{i,t}(x_2)$ equals

$$\frac{P_{i,t}(x_1)}{P_{i,t}(x_2)} = \frac{P(x_1|\bar{X}^*_{i,t})}{P(x_2|\bar{X}^*_{i,t})} = \frac{P(x_1, \bar{X}^*_{i,t})/P(\bar{X}^*_{i,t})}{P(x_2, \bar{X}^*_{i,t})/P(\bar{X}^*_{i,t})} = \frac{P(x_1, \bar{X}^*_{i,t})}{P(x_2, \bar{X}^*_{i,t})} = \mathrm{e}^{Q(x_1,\bar{X}^*_{i,t}) - Q(x_2,\bar{X}^*_{i,t})}.$$

From Eq. (A.3) it follows that

$$Q(x_1, \bar{X}^*_{i,t}) - Q(x_2, \bar{X}^*_{i,t}) = (x_1 - x_2)(a_i + X_{t-1}W_{\cdot i} + W_{i\cdot}X'_{t+1}).$$

Here, $W_{\cdot i}$ and $W_{i\cdot}$ are the ith column and the ith row of W, respectively. For this equation to be valid at every time step, the vectors X_0 and X_{T+1} are defined to consist of 0s only. We shall use the shorthand notation

$$\Delta Q_{i,t} = a_i + X_{t-1}W_{\cdot i} + W_{i\cdot}X'_{t+1},$$

with, as noted above, $a_i = 0$ in the DSS model. Note that $\Delta Q_{i,t}$ is a function of $\bar{X}^*_{i,t}$, but does not depend on the value x_1 or x_2 of node (i, t). The ratio of local probabilities can now more simply be written as

$$\frac{P_{i,t}(x_1)}{P_{i,t}(x_2)} = \mathrm{e}^{(x_1-x_2)\Delta Q_{i,t}}. \tag{A.4}$$

Both for the Golden and Rumelhart model and the DSS model the local probability distribution can be derived from (A.4). First, for the Golden and Rumelhart model there are theoretically only two possible values, 0 and 1, for a node. Taking $x_1 = 0$ and $x_2 = 1$ in Eq. (A.4), and using $P_{i,t}(0) + P_{i,t}(1) = 1$, leads to

$$P_{i,t}(1) = \frac{1}{1 + \mathrm{e}^{-\Delta Q_{i,t}}}, \tag{A.5}$$

that is, the logistic function of $\Delta Q_{i,t}$. With 0 and 1 as the possible values, this probability that $x_{i,t} = 1$ equals the expected value of the local probability distribution of node (i, t).

For the DSS model, the situation is more complicated, since nodes can now theoretically have any value between 0 and 1. Consequently, the local probability $P_{i,t}$ is to be replaced by a probability density. Applying Eq. (A.4) to this density, with $x_2 = 0$ and $x_1 = x$, $x \in [0, 1]$, leads to the following equation for the density $P_{i,t}$:

$$P_{i,t}(x) = P_{i,t}(0)\,\mathrm{e}^{x\Delta Q_{i,t}}.$$

Being a probability density, $P_{i,t}$ has to integrate to unity over the interval [0, 1]. Thus, for $\Delta Q_{i,t} \neq 0$,

$$\begin{aligned}\int_0^1 P_{i,t}(x)\,\mathrm{d}x &= P_{i,t}(0)\int_0^1 \mathrm{e}^{x\Delta Q_{i,t}}\,\mathrm{d}x = P_{i,t}(0)[(\Delta Q_{i,t})^{-1}\mathrm{e}^{x\Delta Q_{i,t}}]_0^1 \\ &= P_{i,t}(0)(\Delta Q_{i,t})^{-1}(\mathrm{e}^{\Delta Q_{i,t}} - 1) = 1,\end{aligned}$$

showing that $P_{i,t}(0) = \Delta Q_{i,t}(\mathrm{e}^{\Delta Q_{i,t}} - 1)^{-1}$ and so

$$P_{i,t}(x) = \frac{\Delta Q_{i,t}\mathrm{e}^{x\Delta Q_{i,t}}}{\mathrm{e}^{\Delta Q_{i,t}} - 1}.$$

If $\Delta Q_{i,t}$ approaches zero, this density approaches the uniform density $P_{i,t}(x) = 1$ on the interval [0, 1]. This density also results directly from applying the above argument to the case $\Delta Q_{i,t} = 0$.

Inspection of the expression for $P_{i,t}$ makes clear that the maximum probability density is always obtained for one of the extreme values: for $x_{i,t} = 0$ if $\Delta Q_{i,t} < 0$, and for $x_{i,t} = 1$ if $\Delta Q_{i,t} > 0$. This is why the inference model described in Section 4.3 lets each $x_{i,t}$ approach either 0 or its maximum value.

For $\Delta Q_{i,t} \neq 0$, the expected value of $x_{i,t}$ is obtained through integration by parts:

$$\begin{aligned} E_{i,t}(\Delta Q_{i,t}) &= \int_0^1 x P_{i,t}(x)\,\mathrm{d}x = \frac{\Delta Q_{i,t}}{\mathrm{e}^{\Delta Q_{i,t}} - 1} \int_0^1 x\,\mathrm{e}^{x\Delta Q_{i,t}}\,\mathrm{d}x \\ &= \frac{\Delta Q_{i,t}}{\mathrm{e}^{\Delta Q_{i,t}} - 1} \left(\left[(\Delta Q_{i,t})^{-1} x\,\mathrm{e}^{x\Delta Q_{i,t}} \right]_0^1 - \int_0^1 (\Delta Q_{i,t})^{-1} \mathrm{e}^{x\Delta Q_{i,t}}\,\mathrm{d}x \right) \\ &= \frac{\Delta Q_{i,t}}{\mathrm{e}^{\Delta Q_{i,t}} - 1} \left((\Delta Q_{i,t})^{-1} \mathrm{e}^{\Delta Q_{i,t}} - \left[(\Delta Q_{i,t})^{-2} \mathrm{e}^{x\Delta Q_{i,t}} \right]_0^1 \right) \\ &= \frac{1}{1 - \mathrm{e}^{-\Delta Q_{i,t}}} - \frac{1}{\Delta Q_{i,t}}. \end{aligned}$$

According to this expression, $E_{i,t}(0) = 1/2$ in the limit for $\Delta Q_{i,t}$ going to zero, corresponding to the expected value of the uniform density valid for $\Delta Q_{i,t} = 0$.

Appendix B. Implementation of world knowledge

The world knowledge that the model uses is extracted from the microworld description constructed in Section 3, consisting of 250 example situations, in each of which every basic proposition is either known to be the case or known to be not the case. Since there are 14 basic propositions (see Table 1) an example situation can be represented by a vector consisting of 14 binary elements, one for each proposition. An element has a value of 1 if the corresponding proposition is the case, or 0 if it is not. For instance, the situation in which the sun shines (the first proposition) and Bob is outside (the third proposition), and no other basic proposition is the case, corresponds to the vector $S = (1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$.

Implementing world knowledge is a two-stage process. First, a Self-Organizing Map is trained on the individual example situation vectors S. Contingencies between situations are ignored in this stage. Next, the resulting mappings are used to convert each vector S into its distributed representation. The contingencies between situations at adjacent time steps are used to compute the temporal world-knowledge matrix W from the distributed example situation vectors.

B.1. Vector representations of propositions

The 250 example situations serve as input to a two-dimensional SOM consisting of $10 \times 15 = 150$ hexagonal cells, as in Fig. 2. Between every two cells i and j a distance $d(i, j)$ is defined. This equals the minimum number of steps needed to get from i to j, if every step takes you from a cell to its immediate neighbor. The distance between two neighboring cells is 1, and the largest distance on a 10×15 map with hexagonal cells is 16. The *neighborhood* of cell i is defined as the set of cells j that lie within a certain distance of N from i, so j is in the neighborhood of i iff $d(i, j) \leq N$. Note that i is in its own neighborhood.

With each cell i is associated a vector μ_i of weights between 0 and 1. These weight vectors consist of one element for each basic proposition, so $\mu_i = (\mu_i(\text{Sun}), \mu_i(\text{Rain}), \ldots, \mu_i(\text{J wins}))$. Training the SOM comes down to setting these vectors so that the structure of the input vectors is mapped onto the two dimensions of the SOM.

Before training begins, a learning rate parameter $\alpha = .9$ and a neighborhood size parameter $N = 16$ are set. Next, the SOM is trained by repetitively presenting it with all example vectors. After presentation of each vector S:

1. The euclidean distances between each weight vector and S are computed.
2. Let i be the cell whose weight vector μ_i is closest to S. All cells in the neighborhood of i now have their weight vectors moved towards S. If j is one of these cells, its weight vector changes by an amount $\alpha(S - \mu_j)$.
3. The value of N is reduced by 4×10^{-4} and α is reduced by 3.52×10^{-5}.

These steps are repeated until all example vectors have been presented to the SOM 100 times. By then, $N = 6$ and $\alpha = .02$. Next, the training process continues but without changing α. After presenting all example vectors 60 more times, $N = 0$ and training is completed.

The vector representation of a proposition is obtained by taking from each cell's weight vector the element corresponding to the proposition. For instance, the first value of weight vector μ_i is $\mu_i(\text{Sun})$. This is the extent to which cell i belongs to the representation of "the sun shines." The full vector representation of "the sun shines" is $\mu(\text{Sun}) = (\mu_1(\text{Sun}), \mu_2(\text{Sun}), \ldots, \mu_{150}(\text{Sun}))$.

B.2. Temporal world knowledge

The temporal world-knowledge matrix W is also based on the microworld description but its values are not obtained by a training procedure. Instead, W is computed directly from the example situations.

Let $S_1, S_2, \ldots, S_{250}$ be the sequence of example situations developed in Section 3. The kth example can be represented by a vector $\mu(S_k)$ in distributed situation space by applying the rules in Eq. (1) to the vector representations of propositions. Before W can be computed, these vectors need to be normalized:

$$\nu_i(S_k) = \frac{\mu_i(S_k)}{\bar{\mu}(S_k)} - 1.$$

Each vector $\mu(S_k)$ is divided by the average value of its elements, $\bar{\mu}(S_k)$. As a result, all vectors have the same average value of 1. Next, from all vectors, 1 is subtracted, making the averages of each vector ν equal to 0. Each entry in W is computed from the normalized vectors ν:

$$w_{ij} = \frac{1}{K-1} \sum_{k=1}^{K-1} \nu_i(S_k)\nu_j(S_{k+1})$$

where $K = 250$, the number of training situations. If it often happens that two SOM cells i and j both have a high value or both have a low value in consecutive example situations, then w_{ij} will become positive. If i and j often have dissimilar values in consecutive examples, w_{ij} will become negative. In this way, w_{ij} reflects the temporal contingencies between cells i and j.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, *36*, 192–236.

Bestgen, Y., & Vonk, W. (1995). The role of temporal segmentation markers in discourse processing. *Discourse Processes*, *19*, 385–406.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*, 177–220.

Cressie, N. A. C. (1991). *Statistics for spatial data*. New York: Wiley.

Dormand, J. R., & Prince, P. J. (1980). A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, *6*, 19–26.

Duffy, S. A., Shinjo, M., & Myers, J. L. (1990). The effect of encoding task on memory for sentence pairs varying in causal relatedness. *Journal of Memory and Language*, *29*, 27–42.

Fletcher, C. R. (1994). Levels of representation in memory for discourse. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 589–607). San Diego, CA: Academic Press.

Fletcher, C. R., & Bloom, C. P. (1988). Causal reasoning in the comprehension of simple narrative texts. *Journal of Memory and Language*, *27*, 235–244.

Garrod, S. C., & Sanford, A. J. (1994). Resolving sentences in a discourse context: How discourse representation affects language understanding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 675–698). San Diego, CA: Academic Press.

Golden, R. M. (1996). *Mathematical methods for neural network analysis and design*. Cambridge, MA: MIT Press.

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

Golden, R. M., Rumelhart, D. E., Strickland, J., & Ting, A. (1994). Markov random fields for text comprehension. In D. S. Levine & M. Aparicio (Eds.), *Neural networks for knowledge representation and inference* (pp. 283–309). Hillsdale, NJ: Erlbaum.

Golding, J. M., Millis, K. M., Hauselt, J., & Sego, S. A. (1995). The effect of connectives and causal relatedness on text comprehension. In R. F. Lorch & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 127–143). Hillsdale, NJ: Erlbaum.

Goldman, S. R., & Varnhagen, C. K. (1986). Memory for embedded and sequential story structures. *Journal of Memory and Language*, *25*, 401–418.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163–182.

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

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*, 257–266.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin: Springer.

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*, 211–240.

Langston, M. C., & 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: Erlbaum.

Langston, M. C., Trabasso, T., & Magliano, J. P. (1999). A connectionist model of narrative comprehension. In A. Ram & K. Moorman (Eds.), *Understanding language understanding: Computational models of reading* (pp. 181–226). Cambridge, MA: MIT Press.

Luftig, R. L. (1982). Effects of paraphrase and schema on intrusions, normalizations, and recall of thematic prose. *Journal of Psycholinguistic Research*, *11*, 369–380.

Magliano, J. P., Zwaan, R. A., & Graesser, A. (1999). The role of situational continuity in narrative understanding. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 219–245). Mahwah: Erlbaum.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*, 440–466.

Millis, K. K., Graesser, A. C., & Haberlandt, K. (1993). The impact of connectives on the memory for expository texts. *Applied Cognitive Psychology*, *7*, 317–339.

Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*, *33*, 128–147.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Murray, J. D. (1995). Logical connectives and local coherence. In F. Lorch & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 107–125). Hillsdale, NJ: Erlbaum.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, *25*, 227–236.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*, 131–157.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, *26*, 453–465.

Noordman, L. G. M., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language*, *31*, 573–590.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, *29*, 37–60.

Schmalhofer, F., McDaniel, M. A., & Keefe, D. (2002). A unified model for predictive and bridging inferences. *Discourse Processes*, *33*, 105–132.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479–515). San Diego, CA: Academic Press.

Smith, D. A., & Graesser, A. C. (1981). Memory for actions in scripted activities as a function of typicality, retention interval, and retrieval task. *Memory & Cognition*, *9*, 550–559.

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language*, *42*, 423–443.

StJohn, M. F. (1992). The Story Gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*, 271–306.

StJohn, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257.

StJohn, 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). Hove, UK: Erlbaum.

Trabasso, T., & Van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*, 612–630.

Van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539–588). San Diego, CA: Academic Press.

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). Mahwah, NJ: Erlbaum.

Van den Broek, P., Young, M., Tzeng, Y., Linderholm, T., 1999. The landscape model of reading: Inferences and the online construction of a memory representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71–98). Mahwah, NJ: Erlbaum.

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

Varnhagen, C. K., Morrison, F. J., & Everall, R. (1994). Age and schooling effects in story recall and story production. *Developmental Psychology*, *30*, 969–979.

Vonk, W., & Noordman, L. G. M. (1990). On the control of inferences in text understanding. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 447–464). Hillsdale, NJ: Erlbaum.